Editora
Ibict

Tecnologias utilizadas
em pesquisas acadêmicas em Ciências Sociais Aplicadas

Brasília - 2023

Organizadores

Milton Shintaku
Diego José Macêdo
Luciano Heitor Gallegos Marin

TECNOLOGIAS UTILIZADAS EM PESQUISAS ACADÊMICAS EM CIÊNCIAS SOCIAIS APLICADAS

PRESIDÊNCIA DA REPÚBLICA

Luiz Inácio Lula da Silva
Presidente da República

Geraldo José Rodrigues Alckmin Filho
Vice-Presidente da República

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

Luciana Santos
Ministra da Ciência, Tecnologia e Inovação

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA

Tiago Emmanuel Nunes Braga
Diretor

Carlos Andre Amaral de Freitas
Coordenador de Administração - COADM

Ricardo Medeiros Pimenta
Coordenador de Ensino e Pesquisa em Informação para a Ciência e Tecnologia - COEPI

Henrique Denes Hilgenberg Fernandes
Coordenador de Planejamento, Acompanhamento e Avaliação - COPAV

Cecília Leite Oliveira
Coordenadora-Geral de Informação Tecnológica e Informação para a Sociedade - CGIT

Washington Luís Ribeiro de Carvalho Segundo
Coordenador-Geral de Informação Científica e Técnica - CGIC

Alexandre Faria de Oliveira
Coordenador-Geral de Tecnologias de Informação e Informática - CGTI

Milton Shintaku
Coordenador de Tecnologias para Informação - COTEC

Organizadores
Milton Shintaku
Diego José Macêdo
Luciano Heitor Gallegos Marin

TECNOLOGIAS UTILIZADAS EM PESQUISAS ACADÊMICAS EM CIÊNCIAS SOCIAIS APLICADAS

Ibict
Brasília
2023

Organizadores do livro
Milton Shintaku
Diego José Macêdo
Luciano Heitor Gallegos Marin

Autores
Alex Fabianne de Paulo
Alex Sebastião Constâncio
Amanda Damasceno de Souza
Caio Saraiva Coneglian
Denise Fukumi Tsunoda
Diego José Macêdo
Eduardo Ribeiro Felipe
Fábio Castro Gouveia
Fernanda Farinelli
Guilherme Noronha
Henrique Leal Tavares
Ingrid Torres Schiessl
Lucas Rodrigues Costa
Luciano Heitor Gallegos Marin
Milton Shintaku
Rebeca dos Santos de Moura
Tiago Rodrigo Marçal Murakami

Revisão gramatical e ortográfica
Flavia Furlan Granato
Rafael Souza

Design gráfico, diagramação e capa
Rafael Fernandez Gomes

Normalização
Alda Melânia César
Fernanda Maciel Rufino
Ingrid Torres Schiessl
Maison Roberto
Marcela Albuquerque
Raíssa Menêses

Dados Internacionais de Catalogação-na-Publicação (CIP)

T255u Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas [recurso eletrônico] / Organizadores: Milton Shintaku, Diego José Macêdo, Luciano Heitor Gallegos Marin. Brasília: Ibict, 2023.
1 recurso online [328 p.] : il.

Modo de acesso: WWW
Publicação digital (e-book) no formato PDF. [16,1 MB]
ISBN 978-65-89167-93-8
DOI 10.22477/9786589167938

1.Tecnologias informacionais. 2. Pesquisa Acadêmica. 3. Ciências Sociais Aplicadas. I. Instituto Brasileiro de Informação em Ciência e Tecnologia. II. Shintaku, Milton (org.). III. Macêdo, Diego José (org.). IV. Marin, Luciano Heitor Gallegos (org.).

CDU 303.064

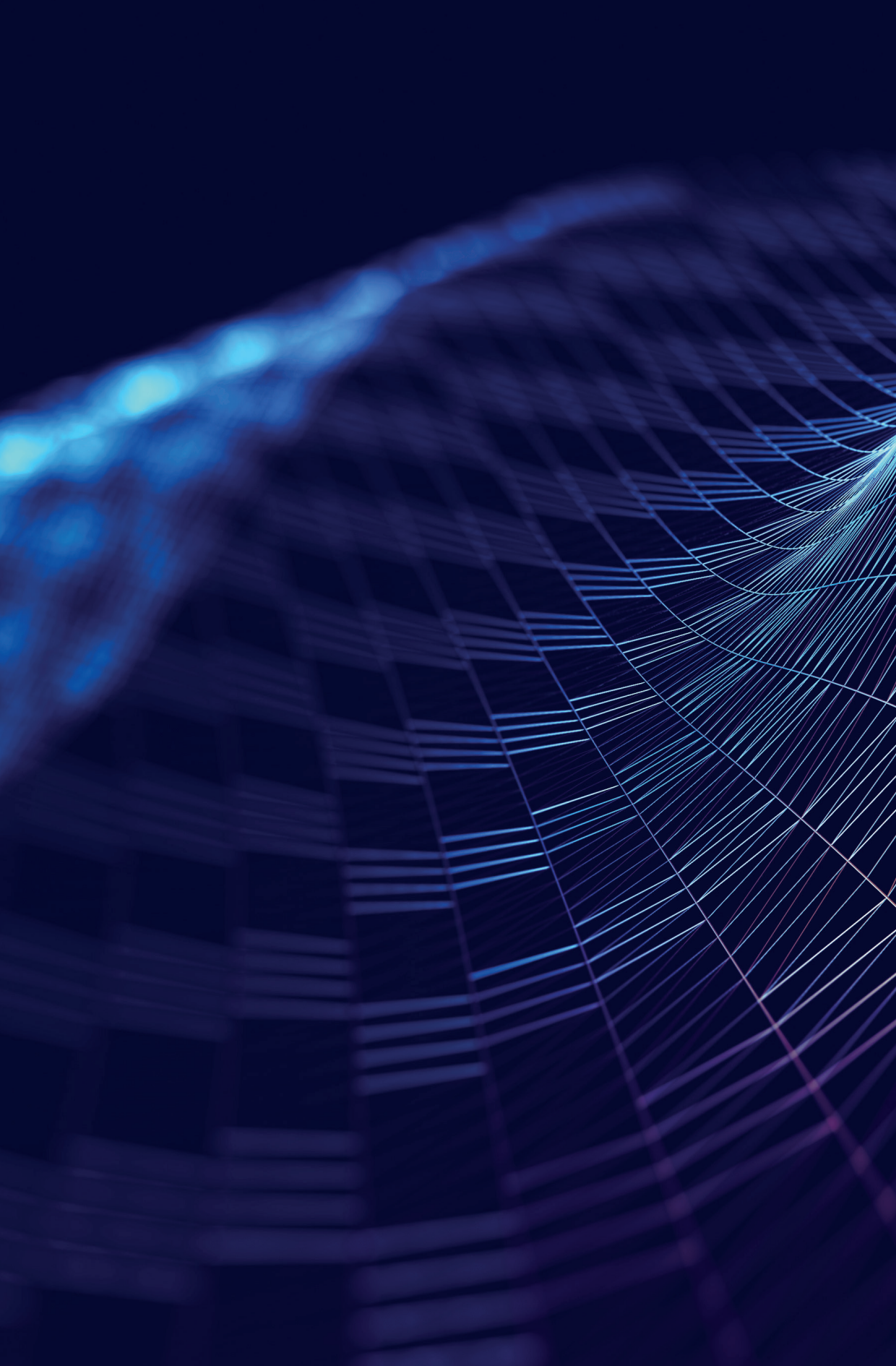
Ficha catalográfica elaborada por: Alda M. César - CRB 1/3253

+Como referenciar este livro:

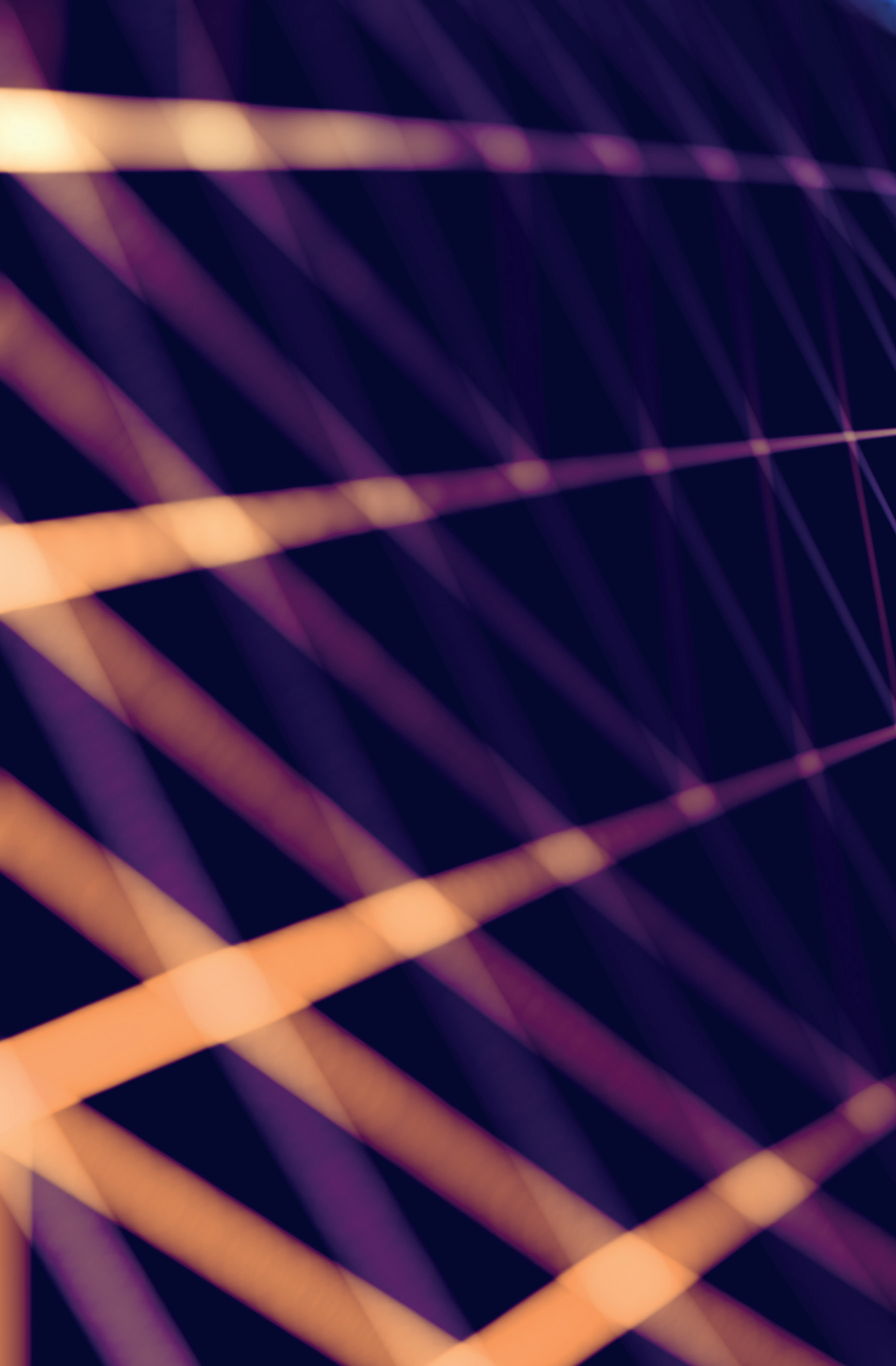
SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas**. Brasília, DF: Ibict, 2023. 332 p. ISBN 978-65-89167-94-5. DOI: 10.22477/9786589167938.

Ibict - Brasília
Setor de Autarquias Sul (SAUS),
Quadra 5, Lote 6, Bloco H - 5º Andar
CEP 70.070-912, Brasília, DF

Ibict - Rio de Janeiro
Rua Lauro Müller, 455 - Botafogo
CEP 22.290-160
Rio de Janeiro, RJ



Sumário



APRESENTAÇÃO

A presente obra é o resultado de inquietações envolvendo projetos de pesquisas sobre Ecossistema de Informação, desenvolvida no âmbito da Coordenação de Tecnologia da Informação (Cotec), do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict). Esses projetos possuem enfoque em estudo de soluções tecnológicas que atendam à gestão da informação em ambientes complexos, tais como aqueles que necessitam do uso de ferramentas tecnológicas e de sistemas de informação para que possam funcionar de forma equilibrada, integrada e colaborativa. Essas soluções são aplicadas a materiais, arquivos e produções humanas, que são elementos inerentes às ciências sociais aplicadas, nos quais as opções de soluções com base em ferramentas tecnológicas e de suas aplicações visam a atender a grande diversidade de cenários práticos possíveis.

As ferramentas tecnológicas são elementos importantes nas pesquisas envolvendo as ciências sociais aplicadas, seja na (1) formação do ecossistema ou (2) como ferramenta para pesquisa e trabalho. No caso (1), que é mais comum, existe a presença de tecnologias constituintes de sistemas de informação, seja pela metodologia utilizada, pelos equipamentos e, principalmente, pelos softwares e sistemas informatizados que fazem uso delas. Entretanto, é no contexto do uso dessas tecnologias, como ferramentas utilizadas em pesquisas e apresentadas ao longo desta obra, inseridas e agregadas a softwares de apoio, no desenvolvimento, na implementação, no monitoramento, na avaliação e no aperfeiçoamento de sistemas de informação, que salientam a importância de suas aplicações.

As pesquisas das ciências sociais aplicadas, na maioria das vezes, ocorrem com o apoio de sistemas de informação, atuando com base em um ou mais fenômenos sociais em cenários reais estabelecidos. Elas são úteis em estudos acadêmicos e profissionais, sendo o último ainda pouco comum

no Brasil, embora ocorram naturalmente em instituições de forma tímida e não coordenada. Muitos desses estudos dependem e estão presentes, principalmente, em sistemas de informação exitosos nessas instituições.

O conhecimento de ferramentas tecnológicas existentes em sistemas de informação que, por sua vez, compõem ecossistemas de tecnologias aplicáveis às pesquisas em ciências sociais aplicadas, possuem caráter prático, do estado-da-arte, e de inovação. A inovação, como procedimento de melhoria nos processos de produção, de serviços, de desenvolvimento e de pesquisas, é resultado de estudos de aplicações de tecnologias com base científica. Assim, o uso de ferramentas tecnológicas em pesquisas transcende o âmbito dos trabalhos acadêmicos e as aplicações em organizações, estendendo a sua atuação para o público em geral. Nos países desenvolvidos essa prática é comum e a própria população participa ativamente de iniciativas de aplicações com base em ferramentas tecnológicas com perfil social aplicado.

Assim, esta obra visa à apresentação de ferramentas tecnológicas, no âmbito das ciências sociais aplicadas, voltada principalmente a pesquisas científicas. Não se trata de uma produção extensiva e completa sobre o tema, mas oferece à comunidade um extrato das principais opções de coleta, processamento e análise de fenômenos sociais aplicados, que pode, por sua vez, contribuir com os estudos e trabalhos de autores e interessados para a facilitação de pesquisas em ciências sociais aplicadas no Brasil e no mundo.

INTRODUÇÃO

Desde os primórdios da humanidade, a curiosidade científica e a tecnológica estão presentes. As ideias de aproveitamento dos ensinamentos passados, experimentação e repasse do novo conhecimento são a base da ciência, e estão presentes na história da humanidade. Assim, para formalização da ciência, foi preciso esquematizar os processos, criar formas de registro dos conhecimentos e estabelecer metodologias de experimentação que permitissem a repetição, a fim de se obter a estrutura aproximada da comunicação científica, como se estabeleceu. Da mesma forma, a tecnologia também está presente no cotidiano desde sempre, incentivando a criação de ferramentas que facilitem a execução de processos e tarefas antes mesmo de o ser humano lascar pedras para criar objetos cortantes, e onde possivelmente tenha utilizado galhos e gravetos como ferramenta para diversos fins. Além disso, a ciência e a tecnologia fazem parte da vida das pessoas, estando presentes nas mais diversas atividades, até mesmo sem serem reconhecidas.

A ciência inicia-se com enfoque no estudo dos fenômenos naturais advindos da filosofia física, posteriormente denominada de ciências naturais. Complementarmente, surgiu uma forma de filosofia que trata dos fenômenos da alma, da qual surgiu, mais tarde, a ciência motivada por aspectos das humanidades e das ciências sociais. A ciência é uma iniciativa humana que busca adquirir conhecimento sobre o mundo natural e tecnológico por meio de um processo sistemático de investigação, observação, experimentação e análise crítica, envolvendo métodos rigorosos e disciplinados para entender os fenômenos, formular hipóteses, realizar experimentos, coletar dados, analisar resultados e tirar conclusões baseadas em evidências empíricas. A ciência pode ser, também, aplicada à busca de conhecimentos sobre o aspecto "social" do ser humano.

O termo "social" está relacionado à interação entre indivíduos ou grupos de pessoas em sociedade, e refere-se à dimensão coletiva da vida humana, abrangendo os relacionamentos, normas, valores, instituições e estruturas que moldam as interações e o comportamento das pessoas dentro de uma comunidade ou sociedade. Os principais conceitos envolvendo o termo social são: interações sociais (refere-se às maneiras como as pessoas se comunicam, se relacionam e se influenciam mutuamente), normas sociais (são regras e expectativas compartilhadas pela sociedade sobre o comportamento apropriado e inadequado em diferentes contextos), cultura (engloba seus valores, crenças, costumes, tradições e práticas compartilhadas de um grupo social), instituições (são estruturas organizacionais que desempenham funções específicas na sociedade, como família, educação, religião, governo e economia) e problemas sociais (são problemas ou desafios que afetam grupos de pessoas ou a sociedade como um todo). Nesse sentido, tornam-se possíveis pesquisas sociais com base científica, como ocorre com as ciências sociais.

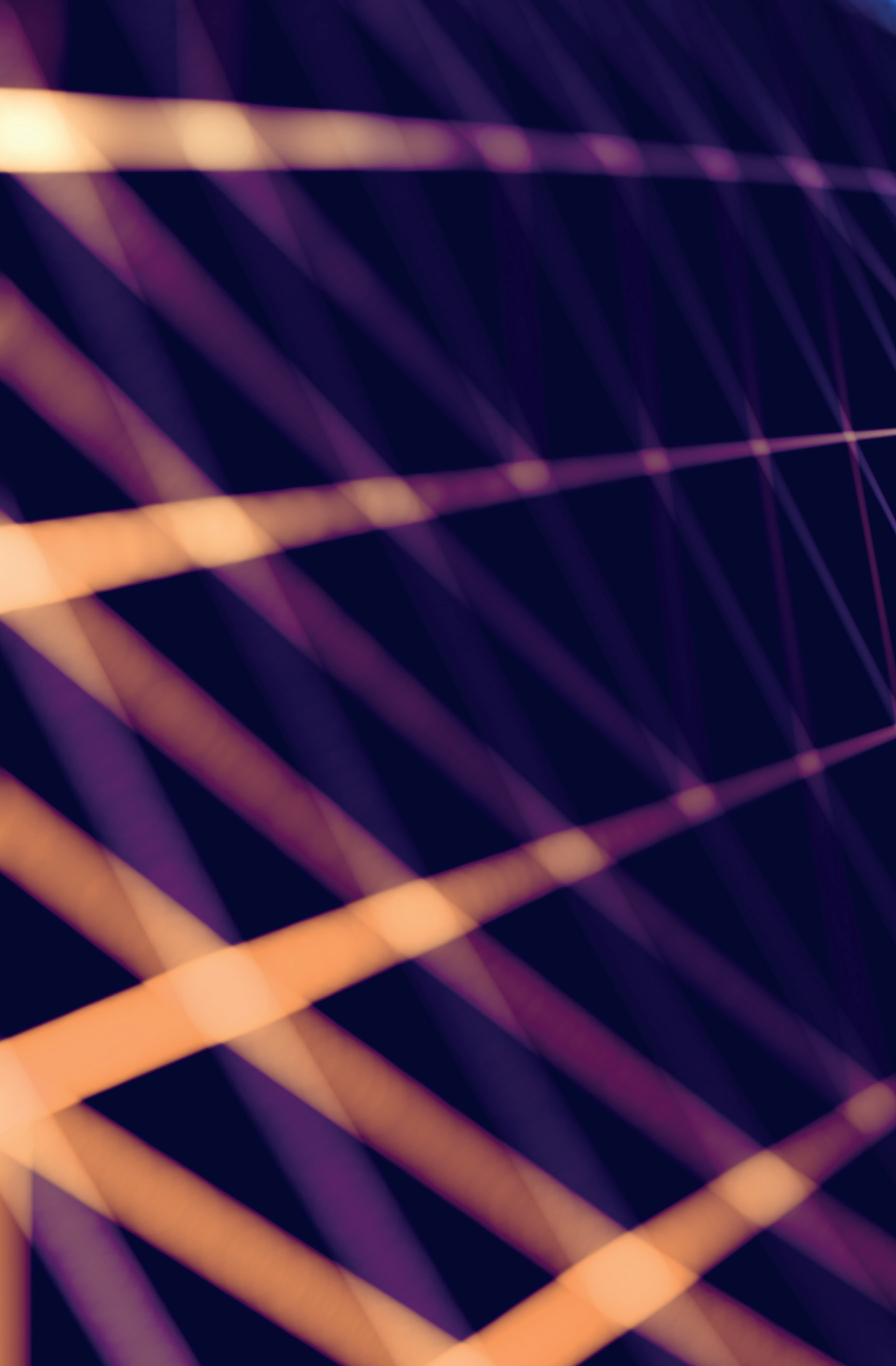
As ciências sociais são um conjunto de conhecimentos humanos organizados em disciplinas acadêmicas dedicadas ao estudo sistemático e científico dos aspectos sociais, culturais, políticos e econômicos da sociedade humana. As principais disciplinas das ciências sociais são: sociologia (estuda a estrutura, organização e dinâmica das sociedades humanas, bem como os padrões de interação social, estratificação social, mudança social e influências culturais), antropologia (investigação das culturas humanas, incluindo aspectos como costumes, tradições, crenças e relações sociais em diferentes grupos étnicos e culturais) e psicologia social (foca no estudo do comportamento humano em contextos sociais e como fatores sociais afetam o pensamento e o comportamento das pessoas). Tais disciplinas podem ser utilizadas de forma transversal, atuando conjuntamente para fornecer novas ideias sobre os fenômenos sociais, permitindo uma compreensão mais profunda da sociedade, seus desafios e oportunidades. Além disso, as ciências sociais desempenham um papel fundamental na formulação de políticas públicas, no desenvolvimento de estratégias de intervenção social e na promoção do entendimento intercultural e da justiça social.

As ciências sociais, em sua perspectiva aplicada, são um campo interdisciplinar de pesquisas e estudos com enfoque na compreensão, no processamento e na análise das complexas interações sociais, culturais, econômicas e políticas que moldam a sociedade contemporânea. Comumente conhecido por

Ciências Sociais Aplicadas, este vasto campo engloba conhecimentos humanos advindos de áreas como Ciências Econômicas, Ciência Política, Administração de Empresas, Psicologia Social, Gestão da Informação – todas voltadas para a investigação e a solução de problemas e fenômenos sociais. Elas desempenham um papel fundamental na compreensão e na resolução dos desafios sociais e políticos que enfrentamos em nosso mundo em constante transformação, contribuindo para a promoção do desenvolvimento sustentável, da justiça social e do progresso humano. Normalmente, as pesquisas em ciências sociais computacionais são apoiadas por ferramentas tecnológicas, e a presente obra trata das principais opções para coleta, processamento e análise de fenômenos sociais aplicados.

Boa leitura,

Os organizadores.



1. NARRATIVA COMPUTACIONAL COM JUPYTER NOTEBOOK COMO APOIO À PESQUISA NAS CIÊNCIAS SOCIAIS APLICADAS

Milton Shintaku
Rebeca dos Santos de Moura
Lucas Rodrigues Costa

1.1 INTRODUÇÃO

O termo "narrativa" tem sido muito empregado no meio político, sobretudo no início da segunda década do século XXI, para denominar a elaboração de histórias sobre determinado evento. Nesses casos, o termo torna-se o mais apropriado para tal fim. Entretanto, em sua etimologia, narrativa vem do latim *narrare*, que significa tornar algo conhecido, que, por sua vez, tem origem em *gnarus*, isto é, o que tem conhecimento. *Gnarus* deu origem a várias palavras do português, como ignorante, aquele que não tem conhecimento etc.

Assim, a narrativa, em sua forma mais pura, tem relação com a transmissão do conhecimento. Na linguística, por exemplo, Paiva (2008) descreve a narrativa como algo contado e recontado a partir de algo real ou fictício, uma série de eventos, relato de acontecimentos, entre outras possibilidades. Para essa área do conhecimento, o estudo das narrativas é importante por possibilitar o entendimento da história e da literatura escrita.

Na visão educativa de Kearney (2012), entende-se que narrativas são instrumentos de compreensão da condição humana, com uma complexidade filosófica que pode ser discutida em cinco visões diferentes, a saber: enredo, recriação, alívio, sabedoria e ética. Nesse caminho, o autor encerra a discussão com a questão ética da narrativa, na qual o narrador é o sujeito (agente) que sofre das interferências da própria história, de forma que ela nunca será neutra.

Squire (2014), por sua vez, discute a definição de narrativas, primeiro com foco na concentração, movimentos, sucessões de signos narrativos, descrição e construção de sentidos. Com isso, constata que a narrativa adota questões sociais, culturais e históricas, atribuindo um contexto particular. Por isso, ela pode assumir diversos formatos e tipos, com raízes semióticas.

Como metodologia de pesquisa, os estudos narrativos têm sido amplamente utilizados. Por isso, Clandinin e Connelly (2011) defendem a pesquisa com narrativas, na medida em que representam experiências pessoais e sociais, com continuidade que possibilita verificar presente, passado e futuro. Nesse sentido, apresenta um cenário complexo, em que a experiência estudada pode ser fruto de outras, o que exige aprofundamento de pesquisa.

Nota-se, no entanto, que narrativas são instrumentos relacionados às ciências humanas e sociais, muito utilizadas como metodologia de pesquisa. Como consequência, ao termo "narrativa" podem ser atribuídos qualificadores, de forma a apresentar *facetas* refinadas, como em textos históricos, literários, visuais, digitais etc. Com isso, possibilita-se a visão das narrativas de acordo com certo viés.

Dentre as novas terminologias destacadas neste capítulo, ressalta-se a narrativa computacional, para a qual se apresenta uma nova tendência nas ciências da computação com o uso das narrativas. Com isso, aliam-se os pontos discutidos nas ciências sociais e humanas, com as ciências técnicas e a computação. Salienta-se a computação como disciplina transversal e instrumento de estudos. Em um cenário em que as tecnologias digitais estão presentes em todas as áreas do conhecimento, as narrativas computacionais podem ser úteis nas ciências sociais aplicadas, servindo como ferramenta para apoio à pesquisa.

Pela construção do termo, narrativa computacional compreende a atividade narrativa restrita a questões voltadas para a ciência da computação. Assim, utilizam-se as técnicas da narrativa para descrever algoritmos, uma vez que, mesmo sendo uma representação da lógica, torna-se algo pessoal, como uma história contada pelo programador e influenciada por experiências pessoais, formação, atuação profissional, entre outras.

Com o desenvolvimento das Tecnologias da Informação e Comunicação (TIC), a oferta de algoritmos computacionais tem sido cada vez mais utilizada em todas as áreas do conhecimento. Assim, torna-se essencial que se compreendam os algoritmos, uma vez que eles tornam a lógica acessível a todos. Logo, deve-se narrar a lógica do algoritmo como se contasse uma história, apresentando o conhecimento contido no programa de computador.

1.2 SOBRE AS NARRATIVAS COMPUTACIONAIS

Desde o início da computação, um dos pontos principais é a construção de programas que materializam um algoritmo, ou seja, uma lógica que representa uma sequência organizada de elementos, com complexidade cada vez maior, voltada à resolução de uma atividade. Entretanto, os programas tendem a ser a representação da forma de pensar do programador, a lógica pessoal amparada por todas as experiências do profissional. Nesse sentido, durante muito tempo havia a máxima na computação, em que somente o autor de um programa conseguia fazer sua manutenção.

Em parte, como solução na computação organizacional, foi criada a linguagem de programação *Common Business Oriented Language* (*Cobol*), que atuava quase como um inglês estruturado. Criado pelo Departamento de Defesa Norte-Americano para ser utilizada em computadores de grande porte (*mainframe*), possui uma estrutura rígida e burocrática, mas que possibilita ser construída, executada e mantida por equipe de colaboradores, devido às características apresentadas pela linguagem.

Entretanto, o *Cobol* apresenta certas limitações, mesmo com todos os avanços apresentados pela sua evolução temporal, desde a sua criação até as versões mais modernas, como a lançada em 2002, pela *International Standardization Organization* (*ISO*). Com estrutura e funcionalidades voltadas para programação, principalmente em lotes, o *Cobol* apresenta certas restrições quanto ao uso em pesquisa. Assim, sua função se volta mais à informatização de atividades administrativas e contábeis.

Mesmo com todas as limitações, o *Cobol* apresentava um embrião do que seria considerado como narrativas computacionais, no qual a lógica

do programa era explicada. Nesse caso, o próprio código e estrutura do programa ofertava explicações que facilitavam a compreensão de seu funcionamento. Outro ponto de destaque era a existência dos comentários no programa, muito utilizados para apresentar sua lógica.

Numa visão mais atual, narrativa computacional tem relação com a Inteligência Artificial (IA), tanto que Riedl (2016) defende o uso dessas narrativas com um viés da construção de formas práticas de criar contextos socioculturais voltados para máquinas. Da mesma forma, possibilita que outros programadores possam entender a lógica utilizada nos programas.

Ontañón e Zhu (2011) descrevem as narrativas computacionais como estórias criadas para serem contadas para computadores, a fim de poderem processá-las. Com isso, adiciona novos elementos no longo processo entre humanos e máquinas, desde a criação de padrões que mesclavam questões sintáticas com semânticas, a fim de que humanos e máquinas pudessem compreender os objetos digitais voltados para o processamento.

1.2.1 PESQUISAS COM A NARRATIVA COMPUTACIONAL

Vall-Vargas, Zhu e Ontañón (2017) categorizam os estudos da Narrativa Computacional em duas áreas: geração e análise. Ambos os casos são estudos de desenvolvimento de algoritmos para que computadores gerem ou analisem narrativas, a fim de tratar melhor os dados ou informações, numa tentativa de aproximar o processamento das pessoas.

No que tange à análise e extração de informação de narrativas, há alguns desafios, tendo em vista a amplitude de possibilidades que as narrativas oferecem, tratando de extração de informação de textos, entendimento de contexto etc. Nesse sentido, muitos casos requerem algoritmos especializados em processamento de linguagem natural a fim de estabelecer modelos que atendam a determinados tipos de narrativas, levando-se em conta a imensa variedade de tipos textuais.

No que diz respeito a dados, no entanto, em alguns pontos torna-se mais simples. Assim, inicialmente as bases de dados eram quase pessoais, atendendo apenas a uma aplicação. Posteriormente, com os administradores de bases de dados, surgiram os dicionários de dados, metadados

que definiam os elementos de dados. A proposta atual é que as narrativas computacionais ofereçam o contexto dos dados, possibilitando seu compartilhamento não apenas em sentido institucional, mas de forma global, principalmente na chamada ciência de dados.

Para tanto, o uso de narrativas computacionais com dados torna-se mais eficaz com o apoio dos chamados cadernos de notas computacionais, ou, no original em inglês, *computational notebooks* (Rule; Tabard; Holland, 2018). Segundo esses autores, os *notebooks* são apropriados para o compartilhamento de narrativas, as quais são compostas não apenas pelas informações, mas pelos dados, códigos, entre outros elementos, atendendo a humanos e máquinas.

Nesse sentido, entendem-se narrativas como a forma pela qual um pesquisador registra seus estudos utilizando um notebook, depositando informações, formas de processamento, resultados em gráficos, tabelas etc. A narrativa computacional se converte, então, na capacidade de publicação organizada de vários elementos da pesquisa, dos dados e da forma como eles se processaram, a fim de possibilitar sua validação, reúso, compartilhamento de código, entre outros fatores.

1.2.2 JUPYTER NOTEBOOK E A NARRATIVA COMPUTACIONAL

O *Jupyter Notebook* é uma ferramenta poderosa e versátil para programação interativa e análise de dados, que permite a criação de documentos dinâmicos, os quais combinam texto explicativo, código executável, visualizações e outras mídias, como imagens e vídeos, tudo em um único ambiente integrado (Jupyter Notebook, 2023a, *on-line*). O *Jupyter Notebook* suporta uma ampla variedade de linguagens de programação, incluindo *Python*, *R*, *Julia*, *Scala* e muitas outras, por meio de *kernels* específicos para cada linguagem (Jupyter Notebook, 2023b, *on-line*). Os *kernels* são processos computacionais que fornecem suporte para a execução de código em uma linguagem de programação específica do ambiente *Jupyter Notebook*. Dessa forma, é possível usar o mesmo ambiente para trabalhar com diferentes linguagens e tecnologias.

Além de oferecer a execução de códigos de forma interativa, facilitando a realização de testes e experimentos rapidamente, com o *Jupyter Notebook*

ainda é possível criar relatórios e documentos ricos em conteúdo, que podem ser compartilhados com outras pessoas. É possível, também, criar recursos avançados para visualização de dados, permitindo a criação de gráficos e visualizações complexas e interativas, facilitando a análise, a comunicação e a manipulação dos documentos por meio dos resultados obtidos.

Uma das principais vantagens do *Jupyter Notebook* é sua flexibilidade e adaptabilidade, uma vez que pode ser usado tanto localmente, em uma máquina individual, quanto em servidores remotos, como na nuvem, permitindo a colaboração em projetos de equipe. Com isso, podem ser criados conjuntos de desenvolvedores em colaboração, ou mesmo grupos de pesquisadores atuando com conjunto de dados, compartilhando códigos e algoritmos.

No entanto, é importante ressaltar que o *Jupyter Notebook* depende da linguagem de programação *Python* para funcionar corretamente. Isso significa que é necessário ter um ambiente *Python* configurado e instalado em seu sistema antes de começar a trabalhar com os notebooks. A infraestrutura necessária para o bom funcionamento da plataforma requer certo conhecimento de *Python*.

Uma das vantagens de se utilizar o *Python* em conjunto com o *Jupyter Notebook* é a possibilidade de aproveitar funções nativas da linguagem, bem como instalar quantos pacotes adicionais forem necessários para atender às demandas de cada projeto. Dessa forma, é possível contar com recursos avançados de processamento de dados e análise estatística.

Dentre os pacotes e bibliotecas mais populares e amplamente utilizados para tratamento e análise de dados, podem ser citados:

> **Pandas** é uma biblioteca que oferece estruturas de dados e ferramentas de análise de alto nível, permitindo a manipulação de grandes conjuntos de dados de forma rápida e eficiente (Panda, 2023).

> **Matplotlib** é uma biblioteca para criação de gráficos e visualizações de dados em *Python*, que permite a criação de diversos tipos de gráficos, incluindo linhas, barras, dispersão e histogramas (Matplotlib, 2023).

> **Numpy** é uma biblioteca que permite a realização de cálculos numéricos complexos e operações matemáticas avançadas, como álgebra linear e transformadas de *Fourier* (NumPy, 2023).

Há outras bibliotecas e pacotes com a oferta de outras facilidades voltadas ao tratamento de dados, muitos dos quais com certas especificidades. Assim, essas bibliotecas apresentam o básico necessário para atuação em dados de pesquisas em ciência social aplicada. Muitas delas, aliás, estão voltadas a operações mais simples, sem grandes quantidades de cálculos estatísticos, muitas das quais dizem respeito à geração de gráficos que serão discutidos nos textos.

1.2.2.1 EXEMPLO DE USO DE JUPYTER NOTE E NARRATIVAS EM MANIPULAÇÃO DE DADOS

A coleta de dados, em alguns casos, pode ser efetuada em fontes que oferecem, entre outras vantagens, a exportação em formatos abertos, como o *Comma Separated Value* (*CSV*). Toma-se como exemplo a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), que agrega teses e dissertações defendidas em 133 instituições brasileiras, com mais de 800 mil documentos em texto integral. Assim, torna-se uma fonte indispensável para pesquisas.

Numa busca simples pelo termo *Python*, feita no dia 24 de abril de 2023, foram encontrados 487 documentos, como mostra a Figura 1. A BDTD exporta os dados, identificando os campos em inglês com a codificação *Unicode Transformation Format 8* (*UTF-8*). Logo, pode ser que sejam necessários ajustes para leitores padrão *ISO*.

Figura 1 - Dados extraídos da BDTD no formato *CSV* abertos em leitor de planilha.

	id	title	authors	subjectsPOR	institutions	types	accesslevel	publicationDates	urls	formats	languages
2	UFG_ea73bfca9	AritmÃ©tica com Py	primary[Cavalcar	LÃ³gica|Programaã	UFG	masterThesis	openAccess	2018	http://r	masterThes	por
3	UERJ_43e00404	RetroanÃ¡lise com o	primary[Santiago	Engenharia civil|En	UERJ	masterThesis	openAccess	2019	http://r	masterThes	por
4	UTFPR-12_b60b	O ensino de matemá	primary[Pesente	MatemÃ¡tica - Estu	UTFPR	masterThesis	openAccess	2019	http://r	masterThes	por
5	UNICAMP-30_9	Desenvolvimento de	primary[GonÃ§a	ProgramaÃ§Ã£o em	UNICAMP	masterThesis	openAccess	2007	https://r	masterThes	por
6	UTFPR-1_6649c	O ensino de matemá	primary[Pesente	MatemÃ¡tica - Estu	UTFPR	masterThesis	openAccess	2019	http://r	masterThes	por
7	URGS_794af535	A Python library for	primary[Sarate, I	FÃ-sica teÃ³rica|Te	UFRGS	masterThesis	openAccess	2022	http://l	masterThes	eng
8	USP_66ebecd3	AutomaÃ§Ã£o de pr	primary[Paulo CÃ©zar Peixoto de Se		USP	masterThesis	openAccess	2019	https://r	masterThes	por
9	UFMT_5e139fb	Estudo de Ã³rbitas p	primary[Almeida	Leis de Kepler|Grav	UFMT	masterThesis	openAccess	2016	http://r	masterThes	por
10	UFPB-2_f927e2	Monty Python e por	primary[Ramos,	Audiovisual|Cultur	UFPB	masterThesis	openAccess	2017	https://r	masterThes	por
11	UFPE_e0e94a6	IntegraÃ§Ã£o pytho	primary[REIS, Eli	Engenharia Civil|El	UFPE	masterThesis	openAccess	2018	https://r	masterThes	por
12	UNICAMP-30_6	ImplementaÃ§Ã£o c	primary[GonÃ§a	CompressÃ£o de im	UNICAMP	masterThesis	openAccess	2003	https://r	masterThes	por
13	UNICAMP-30_7	Ambiente de suport	primary[Silva, Al	Processamento de i	UNICAMP	masterThesis	openAccess	2003	https://r	masterThes	por
14	FGV_3ba41619	VisualizaÇÃ£o de c	primary[Oliveira	ColeÃ§Ãµes Digitai	FGV	masterThesis	openAccess	2021	https://r	masterThes	por
15	SCAR_c4740dc2	Efeitos da alimentaÇ	primary[Ciprianc	Fisiologia|FunÃ§Ãe	UFSCAR	masterThesis	openAccess	2013	https://r	masterThes	por

Fonte: Captura de tela (2023).

Usando a *biblioteca Pandas* do *Python*, é possível ler e manipular dados em diversos formatos, incluindo arquivos *CSV* e *Excel*. Usando o arquivo exportado da BDTD, o exemplo a seguir mostra como ler um arquivo *CSV* e criar um *dataframe*, isto é, uma estrutura de dados semelhante a uma tabela em um banco de dados. Com ela, os dados podem ser manipulados de forma mais fácil.

Quadro 1 - Comandos em *Python* para ler arquivos *CSV* e criar um *dataframe*.

```
# Importação das bibliotecas necessárias
import pandas as pd

# Lê o arquivo CSV
df = pd.read_csv('search_results.csv', sep=';')

# Mostrar as colunas do dataframe
print(df.columns)

# Selecionar algumas colunas do dataframe
df_selected = df[['title', 'authors', 'institutions',
'types', 'publicationDates', 'languages']]

# Mostrar as primeiras linhas de conteúdo
print(df_selected .head())
```

Fonte: Elaborado pelos autores (2023).

Nesse exemplo, o comando *pd.read_csv("search_results.csv", sep=";")* comporta um arquivo *CSV* chamado *search_results.csv* e o armazena em um *DataFrame* do *Pandas* chamado *df*. O parâmetro *sep=";"* especifica que o separador utilizado no arquivo *CSV* é o ponto e vírgula. Tal comando permite ler e estruturar os dados do arquivo *CSV* em uma estrutura tabular, facilitando a manipulação e análise dos dados por meio das funcionalidades do *Pandas*. O comando *df.columns* apresenta as colunas existentes no *dataframe*, com o seguinte resultado:

Quadro 2 - Código *Python* de apresentação das colunas existentes no *dataframe*.

```
['id', 'title', 'authors', 'subjectsPOR', 'institutions',
'types', 'accesslevel', 'publicationDates', 'urls',
'formats', 'languages']
```

Fonte: Elaborado pelos autores (2023).

A partir das colunas do *dataframe*, podemos realizar a seleção vertical ou *slice*, por meio do comando *df[['title', 'authors', 'institutions', 'types', 'publicationDates', 'languages']]*, que seleciona as colunas designadas. Em seguida, as primeiras linhas de conteúdo do *dataframe* são apresentadas com o comando *print(df_selected .head())*, cujo resultado é o seguinte:

Quadro 3 - Código *Python* de apresentação dos resultados obtidos pelo comando chamado *df*.

```
  authors                      institutions  types
pubDates languages
0 Cavalcante, Rogério da Silva UFG      masterThesis 2018
      por
1 Santiago, Carlos Alexandre  UERJ     masterThesis  2019
      por
2 Pesente, Guilherme Moraes  UTFPR          masterThesis
2019         por
3 Gonçalves Neto, Jahyr     UNICAMP  masterThesis  2007
      por
4 Pesente, Guilherme Moraes  UTFPR          masterThesis
2019         por
```

Fonte: Elaborado pelos autores (2023).

Para extrair apenas a lista de autores do conjunto de dados, podemos usar uma simples seleção de colunas *print(df['authors'])*. Entretanto, os dados dessa coluna apresentam informações de *primary* (autor primário) e o *link* do *Lattes* do autor entre parênteses, como vemos no quadro a seguir:

Quadro 4 - Código *Python* de apresentação dos resultados obtidos pelo comando *print (df['authors'])*.

```
0                primary[Cavalcante, Rogério da Silva(NA)]
1      primary[Santiago, Carlos Alexandre de Almeida(...
2      primary[Pesente, Guilherme Moraes(http://latte...
3                primary[Gonçalves Neto, Jahyr, 1980-(NA)]
4      primary[Pesente, Guilherme Moraes(http://latte...
                              ...
482                           primary[Hertzog, Lucas(NA)]
483    primary[Botelho, Gilberto Garcia(http://lattes...
484    primary[Marinho, Jos?? Lino do Nascimento(http...
485    primary[Xavier, Pedro Armentano Mudado(http://...
486    primary[Ferreira, Leandro Martins(http://latte...
Name: authors, Length: 487, dtype: object
```

Fonte: Elaborado pelos autores (2023).

Podemos utilizar os métodos *str.replace()* e *str.strip()* do *Pandas* alinhado ao uso de expressões regulares para tratar o dado e remover as informações desnecessárias, como mostrado a seguir:

Quadro 5 - Comandos em *Python str.replace( )* e *str.strip( )* para uso de expressões regulares.

```
# Remover os prefixos e sufixos da coluna 'authors'
df['autores'] = df['authors'].str.replace(r'^primary\
[|\]$', '', regex=True)

# Remover os conteúdos entre parênteses da coluna
'autores'
df['autores'] = df['autores'].str.replace(r'\([^()]*\)',
'', regex=True).str.strip()
```

Fonte: Elaborado pelos autores (2023).

Assim, com o resultado do processamento, tem-se a lista de autores após tratamento:

Quadro 6 - Código *Python* de apresentação dos resultados obtidos pelo comando *str.replace ( ) e str.strip ( ).*

```
0                Cavalcante, Rogério da Silva
1      Santiago, Carlos Alexandre de Almeida
2                   Pesente, Guilherme Moraes
3                Gonçalves Neto, Jahyr, 1980-
4                   Pesente, Guilherme Moraes
                         ...
482                             Hertzog, Lucas
483                  Botelho, Gilberto Garcia
484         Marinho, Jos?? Lino do Nascimento
485            Xavier, Pedro Armentano Mudado
486                  Ferreira, Leandro Martins
Name: autores, Length: 487, dtype: object
```

Fonte: Elaborado pelos autores (2023).

Uma função interessante do *Pandas* é a *value_counts()*, que calcula valores únicos em determinada coluna do *dataframe*. Com essa função, podemos extrair informações importantes dos tipos de arquivos do conjunto de dados, conforme o exemplo a seguir:

Quadro 7 - Comandos em *Python value_counts ( )* para calcular valores únicos.

```
# Contar valores únicos em uma coluna:
contagem_tipos = df['types'].value_counts()

# Imprime o dataframe resultante
print(contagem_tipos)
```

Fonte: Elaborado pelos autores (2023).

O resultado desse comando é o seguinte:

Quadro 8 - Código *Python* de apresentação dos resultados obtidos pelo comando *value_counts ( ).*

```
masterThesis      382
doctoralThesis    105
```

Fonte: Elaborado pelos autores (2023).

Esse exemplo pode ser aprimorado para mostrar o resultado da porcentagem em relação ao total usando o parâmetro *normalize=True* no método *value_counts()* e multiplicando por 100 para converter em porcentagem, como mostrado a seguir:

Quadro 9 - Comandos em *Python normalize=True* para visualização em porcentagem.

```
# Contar valores únicos em uma coluna, com normalização
contagem_tipos_porcentagem = df[‘types’].value_
counts(normalize=True) * 100

# Imprime o dataframe resultante
print(contagem_tipos_porcentagem)
```

Fonte: Elaborado pelos autores (2023).

O resultado desse comando é o seguinte:

Quadro 10 - Código *Python* de apresentação dos resultados obtidos pelo comando *normalize=True*.

```
masterThesis      78.439425
doctoralThesis    21.560575
```

Fonte: Elaborado pelos autores (2023).

O *Matplotlib* é uma biblioteca para visualização de dados em *Python*. Ela permite criar gráficos de linhas, barras, dispersão, histogramas, entre outros tipos de visualizações. O exemplo a seguir une o método *value_counts()*e o *gráfico de pizza* da *biblioteca Matplotlib*. Os comandos básicos para o gráfico são o *df['institutions'].value_counts()*, que conta a quantidade de itens de cada instituição, e o *plt.pie(...)*, que gera o *gráfico de pizza*. Outros comandos são necessários para estilizar as *labels* e a legenda, além de adicionar um título ao gráfico *(plt.title("Distribuição das Instituições"))*.

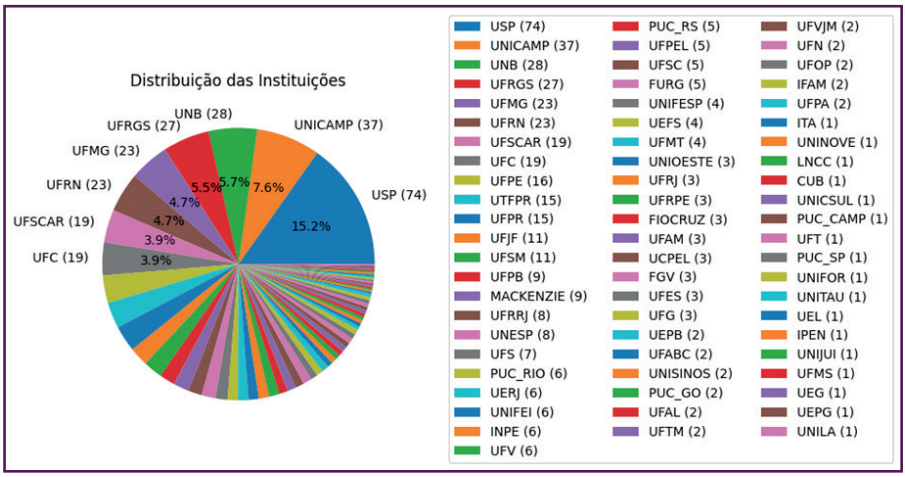
Quadro 11 - Comandos em *Python* utilizando a *biblioteca Matplotlib* para visualizar *gráficos pizza.*

```
# Importação das bibliotecas necessárias
import pandas as pd
import matplotlib.pyplot as plt

# Contagem das instituições
contagem_instituicoes = df['institutions'].value_counts()

# Filtra os itens com porcentagem acima do limite mínimo
itens_labels = contagem_instituicoes[contagem_instituicoes
>= 19]

# Criação do gráfico de pizza
labels = [f'{label} ({contagem_instituicoes[label]})'
if label in itens_labels else '' for label in contagem_
instituicoes.index]
plt.pie(contagem_instituicoes, labels=labels,
autopct=lambda pct: f'{pct:.1f}%' if pct > 3.5 else '')

# Adiciona um título ao gráfico
plt.title("Distribuição das Instituições")

# Função de formatação condicional para os rótulos da
legenda
def format_legend(label):
    return f'{label} ({contagem_instituicoes[label]})'

# Criação da legenda com duas colunas e formatação
condicional nos rótulos
legend_labels = [format_legend(label) for label in
contagem_instituicoes.index]
plt.legend(labels=legend_labels, loc='best', ncol=3, bbox_
to_anchor=(1.1, 1.25))

# Exibe o gráfico
plt.show()
```

Fonte: Elaborado pelos autores (2023).

Por fim, o gráfico é exibido na tela utilizando o comando *plt.show()*.

Figura 2 - Apresentação do gráfico de pizza obtido pelo comando *Python*.



Fonte: Elaborado pelos autores (2023).

No próximo exemplo, usamos o gráfico de barra da biblioteca Matplotlib para mostrar a quantidade de itens publicados por ano. Primeiro, calculamos os itens por ano utilizando o método *value_counts()* e ordenamos esse objeto com o método *sort_index()* para garantir que os anos estejam em ordem crescente. Por fim, criamos o gráfico de barras, usando a função *plt.bar()*, passando os anos filtrados como valores do eixo x e as quantidades de itens correspondentes como valores do eixo y. Adicionamos rótulos aos eixos x e y, e um título ao gráfico.

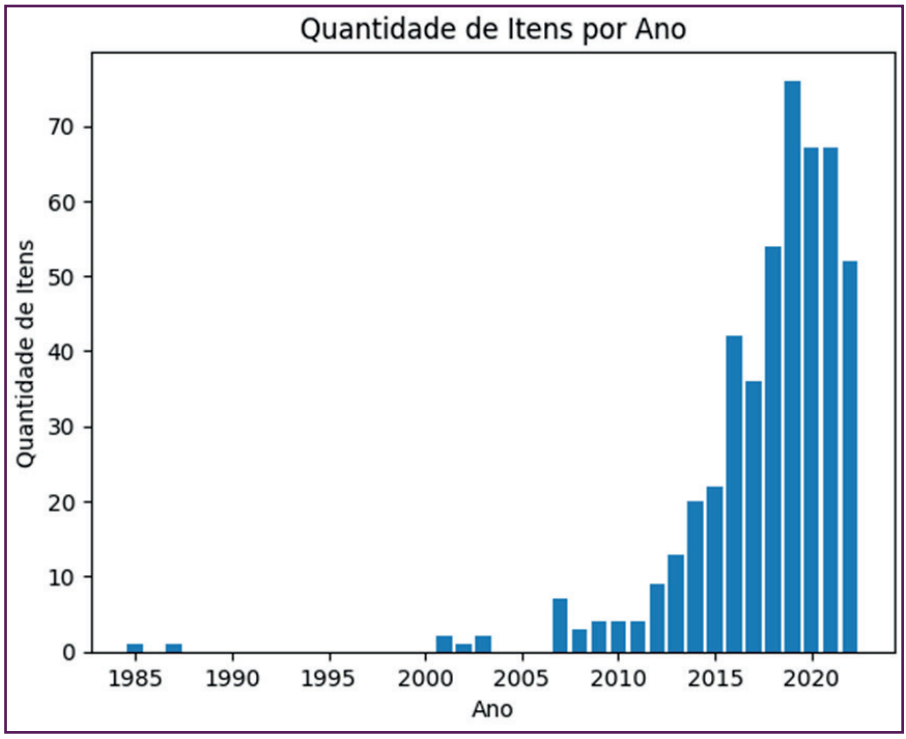
Quadro 12 - Comandos em *Python* utilizando a *biblioteca Matplotlib* para visualizar gráficos em barras.

```
# Importação das bibliotecas necessárias
import pandas as pd
import matplotlib.pyplot as plt

# Contagem dos itens por ano
contagem_por_ano = df['publicationDates'].value_counts().
sort_index()

# Criar o gráfico de barras
plt.bar(contagem_por_ano.index, contagem_por_ano.values)

# Adicionar rótulos aos eixos x e y, e um título ao gráfico
plt.xlabel('Ano')
plt.ylabel('Quantidade de Itens')
plt.title('Quantidade de Itens por Ano')

# Exibir o gráfico
plt.show()
```

Fonte: Elaborado pelos autores (2023).

Por fim, o gráfico é exibido na tela utilizando o comando *plt.show()*.

Figura 3 - Apresentação do gráfico em barras obtido pelo comando *Python*.



Fonte: Elaborado pelos autores (2023).

O *NumPy* é uma biblioteca para computação numérica em *Python*. Ela é amplamente utilizada em ciência de dados, *machine learning* e outras áreas, nas quais é necessário lidar com cálculos matemáticos complexos. No exemplo a seguir, a *biblioteca NumPy* é utilizada para criar um *array* com valores aleatórios, por meio do método *np.random.rand(5)*. Em seguida, é calculada a média dos valores desse *array* utilizando o método *np.mean(a)*.

Quadro 13 - Comandos em *Python* utilizando a biblioteca *NumPy* para cálculos matemáticos.

```
# Importação das bibliotecas necessárias
import numpy as np

# Cria um array NumPy com valores aleatórios
a = np.random.rand(5)

# Calcula a média dos valores do array
media = np.mean(a)

# Imprime a média
print(media)
```

Fonte: Elaborado pelos autores (2023).

Por fim, a média é impressa na tela através do comando *print(media)*.

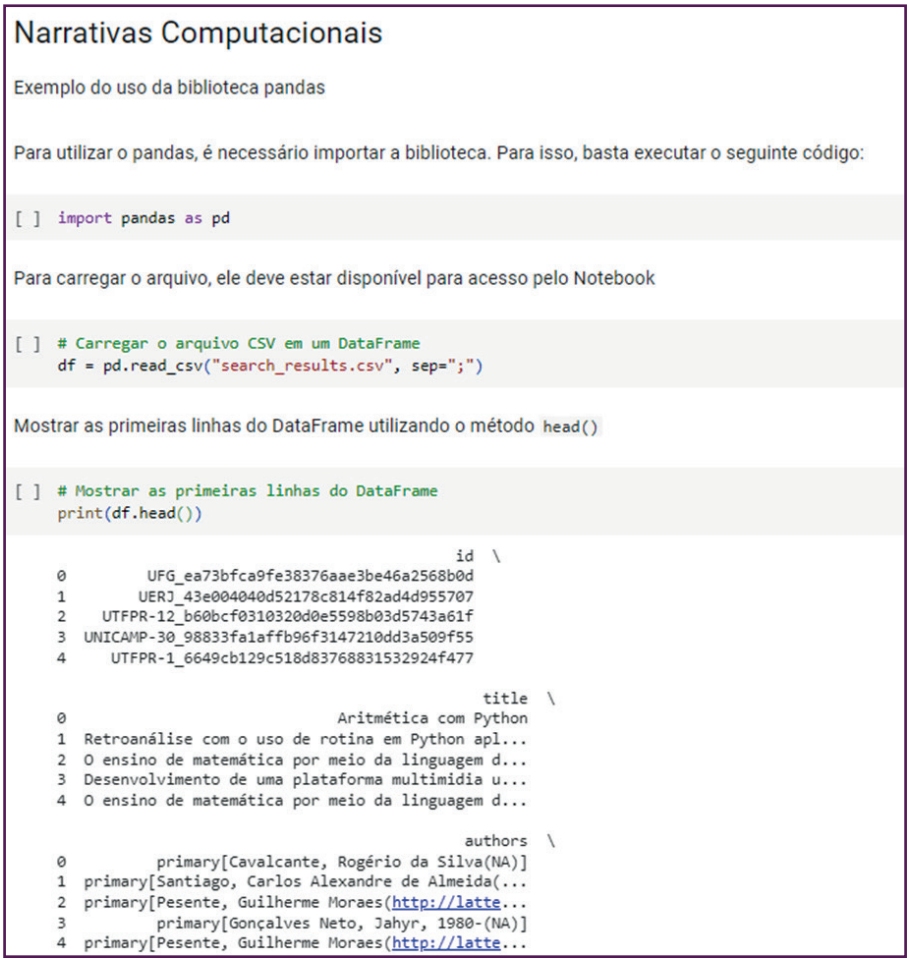
Quadro 14 - Apresentação dos resultados obtidos pelo comando *Python*.

```
0.33580568767017327
```

Fonte: Elaborado pelos autores (2023).

A visualização padrão do *Jupyter Notebook* é a seguinte:

Figura 4 - Visualização dos comandos *Python* na plataforma *Jupyter Notebook*



Narrativas Computacionais

Exemplo do uso da biblioteca pandas

Para utilizar o pandas, é necessário importar a biblioteca. Para isso, basta executar o seguinte código:

```
[ ] import pandas as pd
```

Para carregar o arquivo, ele deve estar disponível para acesso pelo Notebook

```
[ ] # Carregar o arquivo CSV em um DataFrame
    df = pd.read_csv("search_results.csv", sep=";")
```

Mostrar as primeiras linhas do DataFrame utilizando o método `head()`

```
[ ] # Mostrar as primeiras linhas do DataFrame
    print(df.head())
```

```
                                          id  \
0        UFG_ea73bfca9fe38376aae3be46a2568b0d
1       UERJ_43e004040d52178c814f82ad4d955707
2   UTFPR-12_b60bcf0310320d0e5598b03d5743a61f
3  UNICAMP-30_98833fa1affb96f3147210dd3a509f55
4    UTFPR-1_6649cb129c518d83768831532924f477

                                       title  \
0                          Aritmética com Python
1  Retroanálise com o uso de rotina em Python apl...
2  O ensino de matemática por meio da linguagem d...
3  Desenvolvimento de uma plataforma multimidia u...
4  O ensino de matemática por meio da linguagem d...

                                       authors  \
0         primary[Cavalcante, Rogério da Silva(NA)]
1  primary[Santiago, Carlos Alexandre de Almeida(...
2  primary[Pesente, Guilherme Moraes(http://latte...
3         primary[Gonçalves Neto, Jahyr, 1980-(NA)]
4  primary[Pesente, Guilherme Moraes(http://latte...
```

Fonte: Captura de tela (2023).

Esses exemplos são apenas o básico do que é possível fazer com as bibliotecas. O *Pandas*, *Numpy* e *Matplotlib* são amplamente utilizados na análise e visualização de dados, além de possuírem muitas funções e recursos para ajudar a tornar a análise mais eficiente e produtiva, especialmente quando acoplados ao *Jupyter Notebook*.

1.3 CONSIDERAÇÕES FINAIS

O processamento de dados em pesquisa requer, inevitavelmente, ferramentas informatizadas, principalmente pela quantidade. Mesmo nas ciências humanas e sociais, os estudos quantitativos ou mistos não estão totalmente descartados. Em muitos casos, os estudos qualitativos e quantitativos unidos oferecem a melhor das perspectivas, isto é, a precisão do quantitativo com a profundidade qualitativa. Assim, a presença dessas ferramentas se torna essencial para a execução do estudo.

Além disso, adotar narrativas computacionais na metodologia ajuda na apresentação da lógica utilizada no desenvolvimento dos programas de processamento, a fim de possibilitar seu reúso. Nas ciências, a reprodutividade da metodologia faz parte da base fundamental da criação do conhecimento, validando o estudo por meio de análise da reprodutibilidade, no sentido de serem alcançados resultados confiáveis.

REFERÊNCIA

CLANDININ, D. Jean; CONNELLY, F. Michael. **Pesquisa narrativa:** experiência e história em pesquisa qualitativa. Tradução do Grupo de Pesquisa Narrativa e Educação de Professores ILEEI/UFU. Uberlândia, MG: EDUFU, 2011. 250 p.

JUPYTER NOTEBOOK. Documentation. **The Jupyter Notebook**. 2023a. Disponível em: https://jupyter-notebook.readthedocs.io/en/latest/notebook.html. Acesso em: 20 set. 2023.

JUPYTER NOTEBOOK. **Project Jupyter**. 2023b. Disponível em: https://jupyter.org. Acesso em: 20 set. 2023.

KEARNEY, Richard. Narrativa. **Educação & Realidade**, Porto Alegre, v. 37, n. 2, 2012. Disponível em: https://seer.ufrgs.br/educacaoerealidade/article/view/30354. Acesso em: 19 set. 2023.

MARIANI, Fábio; MATTOS, Magda. Pesquisa narrativa: experiência e história em pesquisa qualitativa. **Revista de Educação Pública**, Cuiabá, v.

21, n. 47, p. 663-667, 10 jul. 2012. DOI 10.29286/rep.v21i47.1766. Disponível em: https://periodicoscientificos.ufmt.br/ojs/index.php/educacaopublica/article/view/1766. Acesso em: 19 set. 2023.

MATPLOTLIB. **Matplotlib**: Visualization with Python. 2023. Disponível em: https://matplotlib.org/. Acesso em: 20 set. 2023.

NUMPY. **NumPy**: the fundamental package for scientific computing with Python. 2023. Disponível em: https://numpy.org/. Acesso em: 20 set. 2023.

ONTAÑÓN, Santiago; ZHU, Jichen. On the role of domain knowledge in analogy-based story generation. *In*: TWENTY-SECOND INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, Barcelona, 16-22 July 2011. **Proceedings** [...]. Barcelona: AAAI Press, 2011. p. 1717-1722. Disponível em: https://dl.acm.org/doi/10.5555/2283516.2283681. Acesso em: 19 set. 2023.

PAIVA, Vera Lúcia Menezes de Oliveira e. A pesquisa narrativa: uma introdução. **Revista Brasileira de Linguística Aplicada**, Belo Horizonte, v. 8, n. 2, 2008. DOI 10.1590/18984-63982008000200001. Disponível em: https://www.redalyc.org/articulo.oa?id=339829603001. Acesso em: 19 mar. 2023.

PANDAS. **Pandas**: version 2.1.1. 2023. Disponível em: https://pandas.pydata.org/. Acesso em: 20 set. 2023.

RIEDL, Mark O. **Computational Narrative Intelligence**: A Human-Centered Goal for Artificial Intelligence. 21 Feb. 2016. Disponível em: http://arxiv.org/abs/1602.06484. Acesso em: 19 set. 2023.

RULE, Adam; TABARD, Aurélien; HOLLAN, James D. Exploration and Explanation in Computational Notebooks. *In*: CHI '18: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 19 Apr. 2018. **Proceedings** [...]. Montreal: ACM, 2018. p. 1-12. DOI 10.1145/3173574.3173606. Disponível em: https://dl.acm.org/doi/10.1145/3173574.3173606. Acesso em: 20 set. 2023.

SQUIRE, Corinne. O que é narrativa? **Civitas:** Revista de Ciências Sociais, Porto Alegre, v. 14, n. 2, p. 272-284, jun. 2014. DOI 10.15448/1984-7289.2014.2.17148. Disponível em: https://revistaseletronicas.pucrs.br/ojs/index.php/civitas/article/view/17148. Acesso em: 19 set. 2023.

VALLS-VARGAS, Josep; ZHU, Jichen; ONTAÑÓN, Santiago. From computational narrative analysis to generation: a preliminary review. *In*: FDG'17: INTERNATIONAL CONFERENCE ON THE FOUNDATIONS OF DIGITAL GAMES 2017, Hyannis, 2017. **Proceedings** [...]. Hyannis, MA: ACM, 2017. p. 1-4. DOI 10.1145/3102071.3106362. Disponível em: https://dl.acm.org/doi/10.1145/3102071.3106362. Acesso em: 20 set. 2023.

DADOS DOS AUTORES:

Milton Shintaku

Milton Shintaku é Doutor em Ciência da Informação pela Universidade de Brasília. Coordenador de Tecnologia para Informação (Cotec) do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

https://orcid.org/0000-0002-6476-4953
shintaku@ibict.br

Rebeca dos Santos de Moura

Rebeca dos Santos de Moura é Mestre em Engenharia de Sistemas Eletrônicos e de Automação e Bacharel em Engenharia da Computação pela Universidade de Brasília (UnB). Desenvolvedora e assistente de pesquisa no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

https://orcid.org/0000-0002-7685-8826
becahp@gmail.com

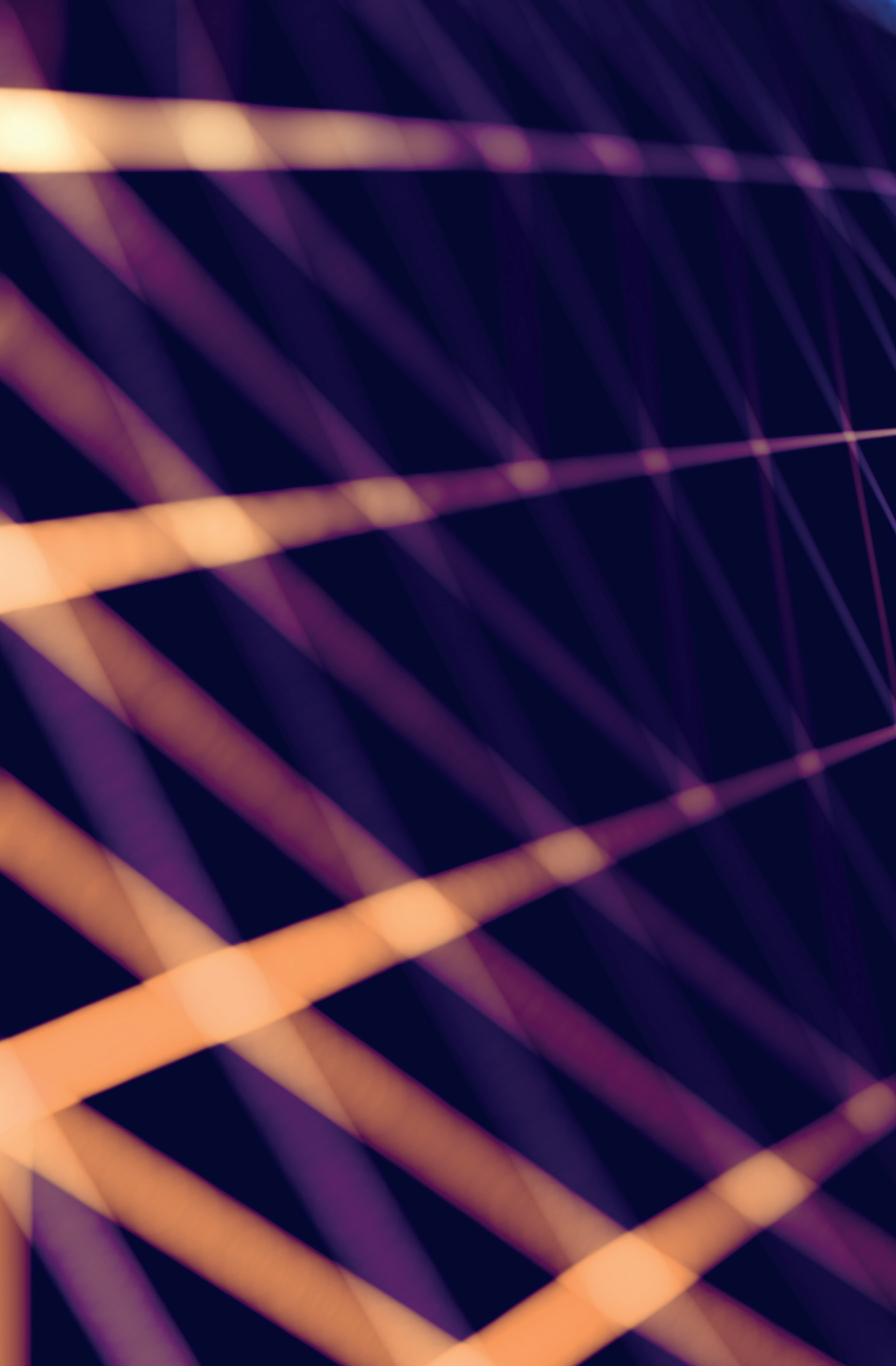
Lucas Rodrigues Costa

Lucas Rodrigues Costa é Doutor em Ciência da Computação pela Universidade de Brasília (UnB), professor substituto da UnB, desenvolvedor e assistente de pesquisa no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

https://orcid.org/0000-0002-0973-4866
lucasrodrigues@ibict.br

Como referenciar o capítulo 1:

SHINTAKU, Milton; MOURA, Rebeca dos Santos de; COSTA, Lucas Rodrigues. Narrativa computacional com Jupyter Notebook como apoio à pesquisa nas ciências sociais aplicadas. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 1. p. 15-37. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap1.



2. PROCESSAMENTO DE LINGUAGEM NATURAL PARA ANÁLISE DE SENTIMENTO UTILIZANDO PYTHON

Guilherme Noronha

2.1 INTRODUÇÃO

Um dos desafios da pesquisa social aplicada é o processamento e gerenciamento de abundância de dados. A era do *Big Data* exige que o(a) pesquisador(a) esteja capacitado(a) para manusear tecnologias que capturam, limpam e extraiam a informação de dados disponibilizados nas mais variadas fontes. Somente em 2022, usuários, na *internet,* movimentaram 97 *zettabytes* de dados, o equivalente a 10^{21} *bytes*. A cada minuto, usuários no *Twitter* compartilharam mais de 347 mil *tweets*, 231 milhões de e-mails foram enviados e 16 milhões de mensagens de texto foram trocadas (Domo, 2022).

Como lidar com esses dados, principalmente quando eles estão na forma desestruturada de textos? A área responsável por lidar com esse problema é o Processamento de Linguagem Natural, ou PLN. Ela é responsável por estudar técnicas que convertam a linguagem natural em formas capazes de serem processadas por computadores, permitindo a análise automatizada de dados. O PLN conta com ajuda de aprendizado de máquina que, segundo Burkov (2019), são métodos de inteligência artificial capazes de fazer tarefas sem serem explicitamente programados para tal.

A área moderna do PLN data os anos do pós-Segunda Guerra Mundial com iniciativas de Noam Chomsky (1972) por meio da chamada gramática gerativa e a tradução por máquinas criada pela Universidade de Georgetown e IBM (Hutchins, 2004), a citar algumas. Trata-se de um campo extenso de pesquisa com aplicações distintas, por exemplo, tradução de idiomas, sintetizador de voz, sumarizador de textos, reconhecimento de entidades, extração de informação e análise de sentimentos, tema deste capítulo.

Análise de sentimentos é uma técnica que permite extrair informação subjetiva de textos. Nela, o(a) pesquisador(a) é, a partir de um conjunto de documentos, capaz de identificar, extrair e quantificar *metadados* importantes para análise (os sentimentos presentes). Embora a análise seja comumente associada aos sentimentos negativo, positivo e neutro, essa técnica pode se aplicar a diferentes classificações que o pesquisador deseja medir. Só na língua portuguesa, por exemplo, existem mais de setecentos sentimentos que podem ser usados para classificação (Ramos; Freitas, 2019).

Este capítulo pretende auxiliar o(a) pesquisador(a) da área de ciência social aplicada a dominar a metodologia básica do processamento de linguagem natural para fazer análise de sentimentos. Por meio de um caso de uso utilizando o *Python*, explica-se o passo a passo desde o tratamento de dados até a construção de um analisador de sentimentos que seja capaz de medir com alto grau de acurácia sentimentos que podem ser aplicados a diferentes tipos de pesquisa.

Este capítulo foi dividido da seguinte maneira: a seção 2 explica as tecnologias usadas para o desenvolvimento do caso de uso; a seção 3 apresenta um caso de uso detalhando cada passo desde a análise exploratória, a transformação, modelagem e análise de resultados. A seção 4 traz algumas pesquisas feitas nos últimos cinco anos que exemplificam a análise de sentimentos como método. Por fim, a seção 5 traz as considerações finais.

2.2 TECNOLOGIAS USADAS

Para o exemplo mostrado neste capítulo foram usados o *Python* e algumas bibliotecas para PLN e aprendizado de máquina como *spaCy*[1], o *scikit-learn*[2] e o *Pandas*[3]. O código-fonte utilizado pode ser encontrado no *Github*[4].

1 Disponível em: https://github.com/explosion/spaCy .

2 Disponível em: https://scikit-learn.org/.

3 Disponível em: https://pandas.pydata.org/.

4 Disponível em: https://github.com/guilhermenoronha/sentiment_analysis_chapter.

O *spaCy* é um *software* de código aberto usado para processamento avançado de linguagem natural. Seu código-fonte é mantido e desenvolvido pela comunidade. Na versão 3.0, o *spaCy* oferece suporte para mais de 70 idiomas, incluindo o português, técnicas que acompanham a evolução do estado da arte em PLN, componentes de análise prontos para uso, suporte para aplicação de aprendizado de máquina, entre outras funcionalidades. Neste capítulo o *spaCy* será usado como ferramenta para limpeza e tratamento de texto.

O *scikit-learn* é uma biblioteca que oferece soluções de aprendizado de máquina para *Python*. Ela fornece diferentes pacotes para a resolução de diferentes problemas de aprendizado, como classificação, regressão e *clusterização* de dados. A análise de sentimentos é, em sua natureza, um problema de classificação e o *scikit-learn* possui meios para facilitar essa implementação. Assim como o *spaCy*, ele também é de código aberto, além de ser amplamente adotado tanto academicamente quanto industrialmente. Neste capítulo o *scikit-learn* foi usado para treinar o modelo de aprendizado de máquina que fará a análise de sentimentos automática.

Por fim, a biblioteca *Pandas* é a responsável pela análise e manipulação de dados. Ela é a principal biblioteca do *Python* para manipulação de dados no formato tabular. Além de armazenar dados no formato de linhas e colunas, ela possui uma série de funções embutidas úteis na análise de dados. *Pandas* foi usado neste capítulo para armazenar a base e fazer as análises iniciais.

Instruções de como instalar e utilizar o código-fonte podem ser encontradas no *Github* do próprio capítulo.

2.3 ANÁLISE DE SENTIMENTO NA PRÁTICA

É preciso fazer uma ressalva antes de passar para a parte prática. Como dito anteriormente, a análise de sentimentos é um problema de classificação de dados. Ou seja, dado um texto, quer-se saber qual classe ele possui. No exemplo deste capítulo, veremos se o sentimento é positivo ou negativo.

Para que isso seja possível, é necessário treinar um *algoritmo* a fim de que ele seja capaz de identificar, automaticamente, as classes de sentimentos. Esse treinamento requer a presença de uma base de dados com textos pré-selecionados e classificados. Nos casos em que o(a) pesquisador(a) não disponha dessa base, ele deverá montá-la, preferencialmente usando textos semelhantes ao que deseja classificar automaticamente. O(A) pesquisador(a) deverá fazer o trabalho de coleta de dados e posterior classificação. A coleta pode ser feita por meio de bibliotecas para raspagem de dados como o *scrapy*[5] ou por meio de *APIs* (*Application Programming Interface*) especializadas. As *APIs* são aplicações que fornecem uma *interface* de comunicação a um determinado serviço, como *Twitter*, *Instagram*, Governo Federal etc. Elas podem ser encontradas nos sites dos fornecedores como o *tweepy*[6] para o *Twitter*, *Graph API*[7] para o *Instagram* e as *APIs* Governamentais[8] para o Governo Federal.

Após a coleta, o(a) pesquisador(a) deve anotar manualmente os textos com o sentimento que ele deseja classificar. A anotação de textos pode ser feita por meio de *softwares* especializados como o *Prodigy*[9], *Label Studio*[10] etc. Essa coleta inicial é feita somente com o intuito de treinar um *algoritmo* para que ele consiga classificar novos textos manualmente. Quanto maior a quantidade de textos coletados, melhor "treinado" será o *algoritmo*. Embora a coleta de dados seja uma etapa importante, ela não é contemplada neste capítulo cujo foco está apenas no processamento.

Neste capítulo usou-se uma base de dados de comentários de filmes feitos no *IMDB* (Internet Movie Database)[11], site especializado em crítica de obras audiovisuais. Essa base contém duas colunas de interesse: a primeira, com a crítica em língua portuguesa de um usuário sobre um determinado filme

5 Disponível em: https://scrapy.org/.

6 Disponível em: https://www.tweepy.org/.

7 Disponível em: https://developers.facebook.com/docs/instagram-api/.

8 Disponível em: https://www.gov.br/conecta/catalogo/apis/api-de-servicos.

9 Disponível em: https://spacy.io/universe/project/prodigy.

10 Disponível em: https://labelstud.io/.

11 Disponível em: https://www.imdb.com/.

e, a segunda, com a classificação do sentimento dessa crítica, podendo ter os valores "negativo" e "positivo". Uma primeira impressão da base pode ser vista na Figura 1.

Figura 1 - Exemplo de sentimentos para análise de sentimentos de críticas de filmes.

	id	text_pt	sentiment
0	1	Mais uma vez, o Sr. Costner arrumou um filme por muito mais tempo do que o necessário. Além das terríveis seqüências de resgate no mar, das quais há muito poucas, eu simplesmente não me importei com nenhum dos personagens. A maioria de nós tem fantasmas no armário, e o personagem Costers é realizado logo no início, e depois esquecido até muito mais tarde, quando eu não me importava. O personagem com o qual deveríamos nos importar é muito arrogante e superconfiante, Ashton Kutcher. O problema é que ele sai como um garoto que pensa que é melhor do que qualquer outra pessoa ao seu redor e não mostra sinais de um armário desordenado. Seu único obstáculo parece estar vencendo Costner. Finalmente, quando estamos bem além do meio do caminho, Costner nos conta sobre os fantasmas dos Kutchers. Somos informados de por que Kutcher é levado a ser o melhor sem pressentimentos ou presságios anteriores. Nenhuma mágica aqui, era tudo que eu podia fazer para não desligar uma hora.	negativo
12389	12391	Eu fui e vi este filme ontem à noite depois de ser persuadido por alguns amigos meus. Eu admitiria que estava relutante em vê-lo porque, pelo que eu sabia de Ashton Kutcher, ele só conseguia fazer comédia. Eu estava errado. Kutcher interpretou o personagem de Jake Fischer muito bem, e Kevin Costner interpretou Ben Randall com tal profissionalismo. O sinal de um bom filme é que ele pode brincar com nossas emoções. Este fez exatamente isso. Todo o teatro que foi vendido foi superado pelo riso durante a primeira metade do filme, e foi levado às lágrimas durante o segundo semestre. Ao sair do teatro, eu não só vi muitas mulheres em lágrimas, mas também muitos homens adultos, tentando desesperadamente não deixar ninguém vê-los chorando. Este filme foi ótimo, e eu sugiro que você vá vê-lo antes de julgar.	positivo

Fonte: Captura de tela (2023).

2.3.1 ANÁLISE EXPLORATÓRIA DE DADOS

A primeira etapa consiste em fazer uma análise exploratória de dados para entender melhor o universo que o(a) pesquisador(a) trabalhará. Essa análise busca compreender o aspecto da base de dados, identificar as colunas úteis, os possíveis tratamentos de dados a serem feitos, balanceamento de dados etc. Numa primeira análise é importante responder às seguintes perguntas: (1) Qual o tamanho da base?; (2) Quais colunas compõem essa base e qual a relevância de cada uma delas para o projeto? e; (3) Quais classes de sentimento essa base possui e como ela está distribuída?

A Figura 2 ilustra o comando inicial para carregar a base de dados para análise exploratória. Em seguida, usa-se o comando *df.shape* para obter a resposta da primeira pergunta: a base possui 49459 linhas e 3 colunas.

Figura 2 - Comandos para carregar a base de dados usando o *Pandas*.

```
df = pd.read_csv(
    'https://github.com/guilhermenoronha/sentiment_analysis_chapter/raw/main/dataset/sentiment_analysis.zip',
    sep=',',
    index_col=[0]
)
pd.set_option('display.max_colwidth', None)
```

Fonte: Elaborado pelo autor (2023).

Com o comando *df.head()* obtém-se uma amostra do conteúdo da base para responder à segunda pergunta da análise. O resultado é similar ao mostrado na Figura 1. As colunas da base são: *id*, *text_pt* e *sentiment*. Uma análise inicial nos diz que apenas as colunas *text_pt* e *sentiment* são de interesse para a classificação de sentimento. Para remover a coluna "*id*" (*identity*) usa-se o comando *df.drop(columns=['id'], inplace=True*).

Já, para responder à terceira pergunta, pode-se usar o comando *df['sentiment'].drop_duplicates()* a fim de identificar quantas classes de sentimento essa base possui e, o comando *df['sentiment'].value_counts()*, para realizar a contagem de registros de cada sentimento. Existem 24765 avaliações positivas e 24694 avaliações negativas. É uma proporção de 50,07% para sentimentos positivos e 49,93% para sentimentos negativos.

É fundamental entender como as classes de sentimento estão distribuídas dentro da base de estudos. Se o(a) pesquisador(a) deseja que o *algoritmo* aprenda a classificar igualmente todas as classes, é ideal que a proporção de registros de cada uma seja a mais próxima possível. Se o objetivo for, por exemplo, priorizar apenas a classificação correta de sentimentos negativos, é ideal que a base de dados tenha mais dados da classe de interesse. Na impossibilidade de ter essa proporção de dados, também é possível aplicar pesos diferentes na hora de classificá-los. Ou seja, configura-se o *algoritmo* de aprendizado para dar mais importância às classificações de interesse, ainda que estejam em menor número.

A proporção encontrada anteriormente já seria satisfatória, mas, para exemplo teórico, é possível balancear a base usando os seguintes comandos mostrados na Figura 3. Os comandos balanceiam a base removendo registros das categorias de maior quantidade até que elas tenham a mesma

quantidade da categoria com menor número de registros. O resultado é 24694 registros para ambos os sentimentos.

Figura 3 - Balanceando a base de dados.

```
min_rows = df.groupby('sentiment').apply(lambda x: len(x)).min()
df = df.groupby('sentiment').apply(lambda x: x.sample(min_rows)).reset_index(drop=True)
```

Fonte: Elaborado pelo autor (2023).

Após uma análise exploratória de dados, a próxima etapa é a transformação.

2.3.2 TRANSFORMAÇÃO DE DADOS

O processo de transformação trata-se de preparar os dados para serem consumidos por um usuário e/ou serviço. No exemplo deste capítulo, os dados serão consumidos pelo *pipeline* de aprendizado de máquina que será responsável por treinar e classificar opiniões de filmes. Saber quem (ou o quê) vai consumir os dados é importante para o(a) pesquisador(a), a fim de que ele(a) os transforme da maneira correta.

Existe um ditado cunhado por George Fuechsel, muito importante na análise de dados, que diz: "entra lixo, sai lixo." (*apud* Stenson, 2016, *online*). Fueschel queria dizer que alimentar um sistema com dados ruins traz resultados ruins, ou seja, classificações erradas, tomadas de decisões ruins, dados controversos, políticas imprecisas etc. Nesse sentido, a etapa de transformação mostra-se crucial para garantir a qualidade de entrada dos dados.

A análise de texto possui uma série de metodologias utilizadas para a padronização da linguagem natural. Elas são importantes para que diferentes textos possam ser processados por computadores. A metodologia para análise de sentimentos consiste em identificar quais são os conjuntos de palavras cujo significado é determinante na hora de classificar um texto segundo as classes de sentimentos. Por exemplo, as palavras "adorei" e "horrível" possuem alta significância na classificação de sentimento como positivo ou negativo.

A limpeza de dados é a técnica responsável por eliminar do texto partes cujo significado agrega pouco ou nada na tarefa de classificação. Para os

exemplos deste capítulo, aplicaram-se as seguintes transformações de dados: capitalização; lematização; remoção de *stopwords*, representações numéricas, *URLs*, *emoticons* e espaços em branco. As técnicas são brevemente descritas a seguir.

2.3.2.1 CAPITALIZAÇÃO DE TEXTO

Capitalizar um texto é transformá-lo inteiramente em caixa baixa ou caixa alta. Essa técnica é fundamental para padronizar o texto num mesmo conjunto de caracteres. Os computadores armazenam letras maiúsculas e minúsculas de formas diferentes. A letra 'A' e 'a', por exemplo, são armazenadas usando os códigos ASCII 65 e 97 (*American Standard Code for Information Interchange*) respectivamente. Ou seja, para um *algoritmo*, as palavras "Casa" e "casa" terão códigos ASCII distintos. De acordo com George *et al.* (2016) a capitalização de texto reduz o vocabulário e aumenta o poder estatístico e a validade dos resultados, trazendo benefícios para o aprendizado de máquina. Não há diferença entre capitalização em caixa baixa ou alta, embora haja uma adoção maior para capitalização em caixa baixa.

2.3.2.2 LEMATIZAÇÃO

A *lematização* é uma técnica com o propósito de normalizar a linguagem natural unificando palavras com o mesmo lema, mas flexionadas de formas diferentes. Por exemplo, as palavras "faria", "faz", "faço" pertencem ao mesmo lema: "fazer". Num processo de *lematização* todas as palavras são substituídas pelo seu lema equivalente. O propósito dessa transformação é similar ao proposto pela capitalização de texto: reduzir o vocabulário e aumentar o poder estatístico.

A *lematização* possui uma técnica similar chamada de *stemming*. Essa técnica pretende extrair os radicais das palavras. No exemplo do parágrafo anterior, o *stemming* geraria os seguintes resultados: "far", "faz" e "faç". O resultado seria pior, pois o tamanho do vocabulário ao final do processamento seria maior. No entanto, essa técnica destaca-se por preocupar-se em remover os *afixos*. As palavras "desfazendo" e "fazer" teriam o mesmo radical: "faz". A escolha entre um e outro depende do objetivo da

análise. Quando o contexto importa, usa-se o *lematizador*, caso contrário, o *stemming* (Balakrishnan; Lloyd-Yemoh, 2014).

2.3.2.3 STOP WORDS

Stop words são consideradas palavras extremamente comuns que agregam pouco ou quase nada para a análise de um documento. Geralmente são *stop words* as palavras funcionais que se encaixam nas classes de preposição, artigo, interjeição, pronome e conjunção. A lista de *stop words* pode variar de aplicação para aplicação e pode ser personalizada pelo(a) pesquisador(a), caso necessário. Segundo Hickman *et al.* (2022) as *stop words* aumentam o poder estatístico, mas reduz a capacidade de capturar o estilo de escrita do texto. Para análise de sentimentos, o estilo de escrita não é relevante.

2.3.2.4 REMOÇÃO DE OUTRAS CATEGORIAS DE PALAVRAS

Nesta subseção estão incluídos os tratamentos para remoção de *emoticons*, representações numéricas, pontuações e *URLs*. O objetivo de todos é o mesmo: aumentar o poder estatístico de aprendizado. Os *emoticons* podem identificar sentimentos, mas exige que uma transformação à parte seja aplicada para gerenciar diferentes representações e isso foge do escopo deste capítulo. Além disso, a remoção de *emoticons* aumenta a validade do aprendizado. As *URLs* são como *stop words* e carregam pouco ou nenhum significado para a análise. Ele é válido para representações numéricas e pontuações. Embora alguns autores como Goldbeck *et al.* (2012) defendam o uso da exclamação '!' como identificador de personalidade, isso não se aplica ao contexto de análise de sentimentos.

2.3.2.5 OUTRAS TÉCNICAS

As técnicas apresentadas acima não são as únicas presentes dentro do contexto de PLN para transformação de dados. Pode-se também corrigir erros ortográficos, expandir acrônimos e abreviações e/ou fazer controle do uso de negação nos textos. Cada técnica pode produzir efeitos diferentes no processo de aprendizado de máquina e seu uso deve ser considerado pelo(a)

pesquisador(a). É recomendado que o(a) pesquisador(a) faça testes com diferentes técnicas para chegar ao resultado ótimo. Para mais detalhes sobre técnicas de processamento de linguagem natural, ver (Hickman *et al.*, 2022).

2.3.2.6 TRANSFORMANDO OS DADOS

Para transformar os dados usando as técnicas citadas nas subseções anteriores, usaram-se as bibliotecas *emoji*, *spaCy* e *Pandas*. O *spaCy* foi responsável pela capitalização em caixa baixa, remover pontuações, *stopwords*, representações numéricas e *URLs*. A biblioteca *emoji* removeu os *emoticons*. Todas essas transformações foram adicionadas numa função em *Python* executada pelo *Pandas* em cada um dos textos da base de dados. A Figura 4 mostra o código-fonte usado para a transformação.

Figura 4 - Código-fonte da transformação de dados.

```
def clean_text(sentence):
    doc = nlp(sentence)
    tokens = [token.lemma_.lower() for token in doc
              if  not token.is_punct and # Filter punctuation
                  not token.is_stop  and # Filter stopwords
                  not token.like_num and # Filter numeric representations
                  not token.like_url     # Filter urls
             ]
    cleaned_text = emoji.replace_emoji(' '.join(tokens), replace='') # Remove emoticons
    return cleaned_text.replace("  ", " ") # Remove extra whitespaces
df['processed_text'] = df['text_pt'].apply(clean_text)
```

Fonte: Elaborado pelo autor (2023).

O resultado da transformação é armazenado na coluna *processed_text* apenas para efeito de comparação[12]. A coluna *text_pt* pode ser removida, pois não é usada no processo de aprendizado de máquina. A Figura 5 mostra um exemplo de uma opinião antes e depois da transformação.

12 O processo de transformação é custoso e demorado. É esperado um tempo de execução entre 15 e 30 minutos em um computador com *16GB* de memória *RAM* e um processador *i7* da 11ª geração.

Figura 5 - Resultado da transformação de texto.

text_pt	processed_text
Robin Williams é excelente neste filme e é uma pena que o material não seja páreo para ele. Isso pode funcionar se você comprar o "U-S-A! Número Um!" mentalidade, mas história sábia nada acontece. É uma pena, já que o filme está realmente tentando dizer alguma coisa, e diz sinceramente. Apenas não causa um impacto emocional suficiente.	robin williams excelente filme pena material ser páreo funcionar comprar u-s-a mentalidade história sábio acontecer pena filme realmente tentar algum sinceramente causar impacto emocional suficiente

Fonte: Captura de tela (2023).

2.3.3 CRIAÇÃO DO MODELO DE APRENDIZAGEM

A análise de sentimentos objetiva consumir os dados analisados e transformados previamente por um *algoritmo* de aprendizado de máquina. Esse *algoritmo* usará a base de dados para aprender os padrões que identificam um sentimento na hora de avaliar um filme no *IMDB*. Uma vez treinado, esse *algoritmo* pode ser usado para identificar opiniões novas.

A biblioteca responsável por usar e treinar *algoritmos* de aprendizado de máquina é o *scikit-learn*. Ela possui, por padrão, uma série de *algoritmos* diferentes que podem ser aplicados tanto para análise de sentimentos quanto para outros tipos de PLN. Para criar um modelo de aprendizado, o(a) pesquisador(a) deve escolhê-lo previamente e, depois, aplicar uma última transformação para adequar os dados ao modelo de consumo do *algoritmo*.

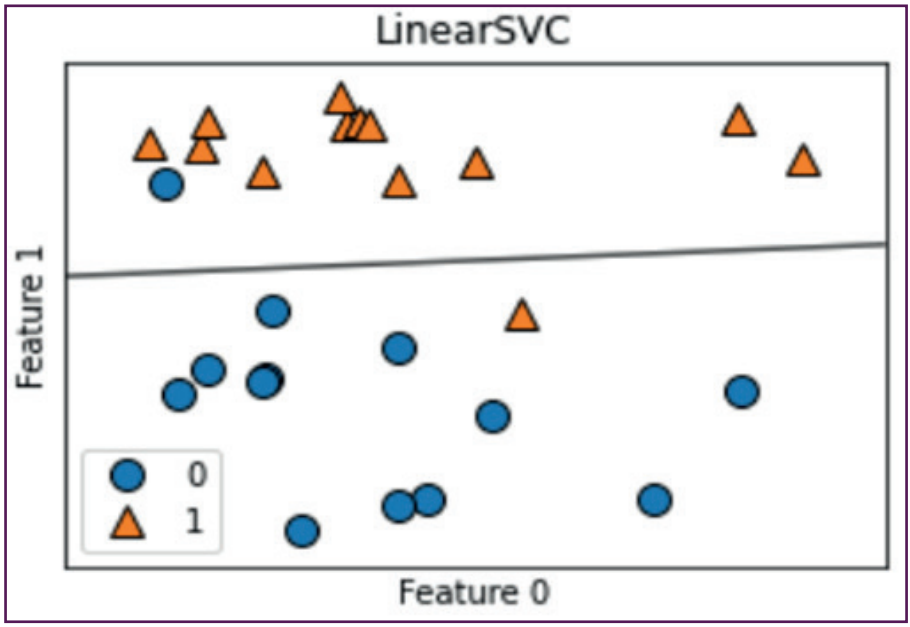
2.3.3.1 ESCOLHENDO O MÉTODO DE APRENDIZADO DE MÁQUINA

O *algoritmo* escolhido para a execução deste capítulo foi *LinearSVC*, mas ele não é o único e não necessariamente o melhor. Cabe ao(à) pesquisador(a) testar e aplicar diferentes métodos para entender qual é aquele que se adequa melhor à base de dados. Aprendizado de máquina não é o escopo deste capítulo. Mais informações sobre o assunto podem ser encontradas em (Burkov, 2019).

O *LinearSVC* é um *algoritmo* da categoria dos classificadores. O *acrônimo SVC* vem do termo *Support Vector Classification*, ou seja, ele faz classificações baseadas em vetores de suporte. Para isso, cada opinião da base de dados, que está em formato de texto, deve ser transformada num vetor para que o *LinearSVC* possa ser aplicado. Em linhas gerais, o *LinearSVC* classifica as classes da base de dados traçando uma *reta ótima* que divide

melhor os vetores em um plano. O *LinearSVC* pressupõe que vetores que estejam próximos um dos outros possuam semelhanças entre si. Logo, espera-se que vetores de classificações positivas e negativas estejam agrupadas em lados opostos do plano (Fan *et al.*, 2008). A Figura 6 ilustra um exemplo de classificação do *LinearSVC*.

Figura 6 - Exemplo de classificação usando o *LinearSVC* num plano de duas dimensões.



Fonte: Captura de tela (2023).

2.3.3.2 VETORIZAÇÃO DOS DADOS

A *vetorização* de textos possui uma técnica simples e funcional. Primeiro, calcula-se o tamanho do vetor sendo igual ao tamanho do vocabulário da base de dados. Cada eixo do vetor corresponde a uma palavra do vocabulário. Se um texto possui uma palavra, o valor do eixo correspondente é preenchido com um valor maior que zero e, zero, caso contrário.

Uma metodologia amplamente usada em PLN para calcular o melhor valor para cada palavra do documento é o *TF-IDF*, do inglês *Term*

Frequency–Inverse Document Frequency. O *TF-IDF* calcula a importância que cada palavra tem em um documento, dando um valor entre 0 e 1. Esse valor é calculado pela frequência que essa palavra aparece no documento e, depois, é multiplicado pela frequência invertida em que essa palavra aparece em diferentes documentos da base de dados. Em outras palavras, quanto mais uma palavra aparece em um documento específico e mais rara ela é entre os demais documentos, mais valor ela possui (Qaiser; Ali, 2018). A *vetorização* da análise de sentimentos aplica os valores calculados de *TF-IDF* para cada palavra em cada documento.

2.3.3.3 CRIANDO E TREINANDO O MODELO DE APRENDIZADO

O código mostrado na Figura 7 descreve uma série de passos para criar e treinar o modelo de aprendizado. As variáveis *tfidf* e *svm* correspondem ao processo de *vetorização* e aprendizado citados nas subseções anteriores. Esses processos são sequenciados dentro da variável *pipe*, que é um *pipeline* de processamento. As variáveis *X* e *y* correspondem aos dados que serão usados e as classes de sentimentos.

Figura 7 - Código-fonte da modelagem e treinamento do algoritmo de aprendizado de máquina.

```
tfidf = TfidfVectorizer()
svm = LinearSVC()
steps = [('tfidf',tfidf),('svm',svm)]
pipe = Pipeline(steps)
X = df['processed_text']
y = df['sentiment']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2)
pipe.fit(X_train,y_train)
y_pred = pipe.predict(X_test)
```

Fonte: Elaborado pelo autor (2023).

Esses dados são divididos em variáveis de treinamento (*X_train, y_train*) e variáveis de teste (*X_test, y_test*). É recomendado, por padrão, que o pesquisador separe 80% dos dados para treinamento e 20% dos dados para teste. Se a base de dados for massiva (*big data*), recomenda-se aumentar o percentual de dados de treinamento.

Por fim, temos a função *fit* que executa o treinamento do modelo baseado nos dados de treinamento fornecidos. Os resultados do treinamento são medidos usando as variáveis de testes e são discutidos na próxima seção.

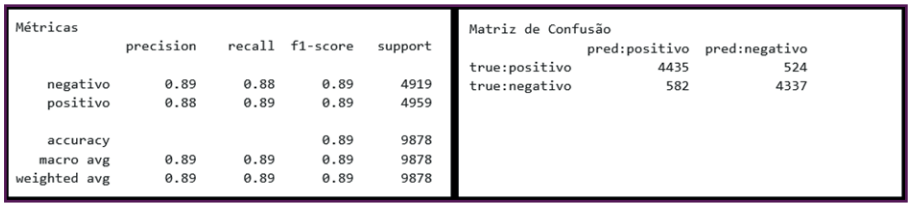
2.3.4 ANALISANDO OS RESULTADOS

Os resultados do treinamento de um *algoritmo* de aprendizado de máquina são obtidos por meio das métricas de precisão, revocação e *F-SCORE*. A precisão mede a proporção de acertos que o modelo teve em relação ao total de predições. Já a revocação mede a proporção de documentos relevantes classificados em relação ao total possível de classificações. Precisão e revocação são excludentes, ou seja, para ter uma boa precisão é preciso renunciar a uma boa revocação e vice-versa. Quando ambas as métricas são importantes, o *F-SCORE* é a melhor escolha, pois trata-se de uma média harmônica entre precisão e revocação (Burkov, 2019).

A escolha do uso entre precisão, revocação e *F-SCORE* depende do objetivo que o(a) pesquisador(a) almeja com a tarefa de análise de sentimentos. Se a prioridade é classificar corretamente os sentimentos, a precisão é mais importante. Caso a prioridade seja recuperar corretamente as classes de interesse, escolhe-se, então, a revocação. O *F-SCORE* é geralmente usado como uma *métrica padrão* de avaliação para comparação de estudos.

As métricas podem ser extraídas do modelo de aprendizado de máquina por meio de uma *matriz de confusão*. Essa matriz resume o quão eficiente foi o teste do modelo. Em um eixo encontram-se as classes de predição e noutro encontra-se a predição feita em relação a essa classe. O *scikit-learn* já calcula as métricas para o(a) pesquisador(a) e a matriz de precisão deve ser usada apenas como um complemento da análise. A Figura 8 mostra o resultado do teste assim como a *matriz de confusão*:

Figura 8 - Métricas e matriz de confusão do modelo.



Métricas

	precision	recall	f1-score	support
negativo	0.89	0.88	0.89	4919
positivo	0.88	0.89	0.89	4959
accuracy			0.89	9878
macro avg	0.89	0.89	0.89	9878
weighted avg	0.89	0.89	0.89	9878

Matriz de Confusão

	pred:positivo	pred:negativo
true:positivo	4435	524
true:negativo	582	4337

Fonte: Captura de tela (2023).

Pode-se tirar algumas conclusões ao analisar as métricas da Figura 8. A precisão para a classe positiva, por exemplo, é de 0,88. Ou seja, espera-se que a predição do *algoritmo* classifique corretamente 88% das vezes. Já a revocação para a mesma classe é de 0,89 que representa que o *algoritmo* consegue recuperar 89% das vezes um sentimento dessa classe. O *F-SCORE* para ambos os sentimentos possui valor de 0,89, refletindo a média harmônica entre precisão e revocação.

As métricas *macro avg* e *weighted avg* são formas diferentes de calcular as métricas. A primeira usa uma média simples entre todas as classes, enquanto a segunda usa uma média ponderada com valores de suporte escolhidos pelo(a) pesquisador(a). O *weighted avg* é importante quando o(a) pesquisador(a) deseja valorizar mais uma classe em detrimento da outra. Como esse valor não foi informado para o *algoritmo*, o *weighted avg* é igual ao *macro avg* (valores de suporte iguais a 1).

Por fim, a acurácia mede a quantidade de sentimentos classificados corretamente pela razão do total de sentimentos classificados. A *matriz de confusão* da Figura 8 é apenas uma forma de validar as métricas acima mencionadas. Ela é interpretada da seguinte maneira: a primeira linha representa a classe de sentimento positivo, sendo que 4435 análises foram classificadas corretamente e 524 incorretamente. Depois que o *algoritmo* foi treinado, ele é capaz de fazer predições conforme mostra a Figura 9.

Figura 9 - Analisando novos sentimentos.

```
sentence_1 = 'Achei esse filme muito bom'
sentence_2 = 'Perdi duas horas da minha vida'
print(f'Sentence: {sentence_1}\nSentiment:{pipe.predict([sentence_1])[0]}')
print(f'Sentence: {sentence_2}\nSentiment:{pipe.predict([sentence_2])[0]}')
```

```
Sentence: Achei esse filme muito bom
Sentiment:positivo
Sentence: Perdi duas horas da minha vida
Sentiment:negativo
```

Fonte: Elaborado pelo autor (2023).

2.4 APLICAÇÕES DE ANÁLISE DE SENTIMENTO EM PESQUISA

Este capítulo trouxe algumas pesquisas que usaram análise de sentimentos nos últimos cinco anos com o objetivo de ilustrar ao leitor as possibilidades dessa técnica dentro de diferentes assuntos das ciências sociais aplicadas. Alsaeedi e Khan (2019) testaram diversas técnicas de aprendizado de máquina para fazer análise de sentimentos no *Twitter*. Os autores encontraram que o *SVM*, da mesma família que o *LinearSVC*, foi a técnica que entregou os melhores resultados. Oliveira *et al*. (2019) analisaram os sentimentos no *Twitter* em relação aos programas públicos do governo Dilma Rousseff. Segundo os autores, as políticas públicas deveriam ser pautadas conforme os discursos da população civil. Os resultados indicaram os programas "Mais Médicos" e "Bolsa Família" como de maior rejeição, enquanto "Pronatec" e "Minha Casa, Minha Vida" tiveram a maior aceitação.

Shaukat *et al*. (2020) se propuseram a usar redes neurais e um dicionário para fazer análise de sentimentos da base do *IMDB* em língua inglesa. Os resultados encontrados por eles foram um F-SCORE de 0,91, pouco melhor que o exemplo usado neste capítulo. Coutinho e Malheiros (2020) usaram o Twitter para fazer a classificação de mensagens homofóbicas. Os autores chegaram a um *F-SCORE* de 0,64. A justificativa para o resultado é a dificuldade de entendimento consensual sobre o que é ou não uma mensagem homofóbica. Os autores utilizaram entrevistas para coleta de dados e obter classificações de cada mensagem. Como não houve consenso quanto à classificação, o modelo não pôde ser treinado corretamente.

Chauhan, Sharma e Sikka (2021) usaram análise de sentimentos para tentar prever os resultados de eleições passadas baseados em dados do *Twitter* e *Facebook*. Embora os resultados tenham sido mistos, os autores concluem que a análise de sentimentos pode ser sim um termômetro para medir as intenções de votos dos eleitores nas redes sociais. Souza, Souza e Meinerz (2021) analisaram sentimentos no mercado de ações em tempo real em uma tentativa de prever a oscilação de preços. A aplicação dos autores atingiu um *F-SCORE* de 0,76. A justificativa para o valor é a baixa captura de *tweets* no período de análise.

Mahyoob *et al*. (2022) usaram análise de sentimentos em *tweets* para medir a percepção da população em relação à emergência causada pela variante *Ômicron* da *COVID-19*. Os pesquisadores dividiram os sentimentos em subclasses que mediam sua força, variando entre o fraco até muito forte. Os resultados indicaram que a população estava preocupada e que esses indicadores poderiam ser usados para a adoção de medidas públicas para acalmá-la. Ainda usando o *Twitter* como fonte de dados, Paes *et al*. (2022) mediram o sentimento de usuários brasileiros em relação ao desmatamento da floresta amazônica, identificando picos de rejeição de até 60%. Os autores identificaram uma correlação entre os picos de rejeição com notícias divulgadas, destacando a importância do papel da mídia para fiscalizar e denunciar atividades nocivas à população brasileira.

2.5 CONSIDERAÇÕES FINAIS

Este capítulo teve como objetivo mostrar ao(à) pesquisador(a) de ciências sociais aplicadas o poder de PLN, mais especificamente a análise de sentimentos e suas múltiplas aplicações. A área de PLN é extensa quanto à variedade de técnicas e metodologias a serem aplicadas em pesquisa. Embora este capítulo traga as principais técnicas usadas na área, ele não é extensivo quanto às possibilidades de aplicação. Um estudo mais detalhado sobre a área pode ser lido em Wankhade, Rao e Kulkarni (2022).

As seções 2 e 3 trouxeram ao leitor um caso de análise de sentimentos na prática. Tanto a base de dados quanto o código-fonte em *Python* foram fornecidos ao(à) leitor(a) para que ele(a) consiga replicá-lo e, se necessário, readaptá-lo ao próprio contexto de pesquisa. As tecnologias usadas são as mais atuais

dentro do contexto da área, fazendo uso de bibliotecas tradicionais tanto de aprendizado de máquina quanto do processamento de textos.

A seção 4 mostrou diferentes aplicações que podem ser desenvolvidas usando a análise de sentimentos nos últimos cinco anos. Embora não seja uma pesquisa exaustiva, é suficiente para apresentar ao leitor ideias de aplicações, bem como possíveis locais para extração e captura de dados (Banks *et al.*, 2018). A análise de sentimentos se beneficia muito das redes sociais e da explosão de dados da *internet*.

Como contribuição, espera-se que este capítulo ajude o(a) leitor(a) a conduzir sua própria pesquisa na área de análise de sentimentos. As pesquisas em PLN dependem muito do desenvolvimento em diferentes idiomas, sendo ainda mais importante o desenvolvimento de pesquisas para o português do Brasil. Segundo a *National Science Board* (2018), o Brasil é apenas o 11º país em termos de pesquisa científica. Ao produzir mais pesquisas para o português, espera-se que esse *ranking* melhore em médio prazo.

REFERÊNCIAS

ALSAEEDI, A.; KHAN, M. Z. A study on sentiment analysis techniques of Twitter data. **International Journal of Advanced Computer Science and Applications - IJACSA**, Cleckheaton, v. 10, n. 2, p. 361-374, 2019.

BALAKRISHNAN, Vimala; LLOYD-YEMOH, Ethel. Stemming and Lemmatization: A Comparison of Retrieval Performances. **Lecture Notes on Software Engineering - LNSE**, [*s. l.*], v. 2, n. 3, p. 262-267, Aug. 2014. DOI 10.7763/LNSE.2014.V2.134. Disponível em: http://www.lnse.org/show-34-165-1.html. Acesso em: 19 set. 2023.

BANKS, G. C.; WOZNYJ, H. M.; WESSLEN, R. S.; ROSS, R. L. A review of best practice recommendations for text analysis in R (and a user-friendly app). **Journal of Business and Psychology**, Berlin, v. 33, n. 4, p. 445-459, Jan. 2018. Disponível em: https://link.springer.com/article/10.1007/s10869-017-9528-3. Acesso em: 19 set. 2023.

BURKOV, A. **The hundred-page machine learning book**. Quebec City: Andriy Burkov, 2019. v. 1, p. 32. ISBN 978-1999579500. Disponível em: http://ema.cri-info.cm/wp-content/uploads/2019/07/2019BurkovTheHundred-pageMachineLearning.pdf. Acesso em: 19 set. 2023.

CHAUHAN, Priyavrat; SHARMA, Nonita; SIKKA, Geeta. The emergence of social media data and sentiment analysis in election prediction. **Journal of Ambient Intelligence and Humanized Computing**, Berlin, v. 12, n. 2, p. 2601-2627, Feb. 2021. DOI 10.1007/s12652-020-02423-y. Disponível em: https://link.springer.com/10.1007/s12652-020-02423-y. Acesso em: 19 set. 2023.

CHOMSKY, N. **Studies on semantics in generative grammar**. Berlin: De Gruyter Mouton, 1972. (Series Janua Linguarum, n. 107). Disponível em: https://www.degruyter.com/document/doi/10.1515/9783110867589/html. Acesso em: 19 set. 2023.

COUTINHO, V. M. M. S.; MALHEIROS, Y. Detecção de mensagens homofóbicas em português no Twitter usando análise de sentimentos. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BraSNAM), 9., Cuiabá, 2020. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1-12. Disponível em: https://sol.sbc.org.br/index.php/brasnam/article/view/11158/11029. Acesso em: 19 set. 2023.

DOMO. Data Never Sleeps 10.0. 2022. Disponível em: https://www.domo.com/data-never-sleeps. Acesso em: 19 set. 2023.

FAN, R.-E.; CHANG, K.-W.; HSIEH, C.-J.; WANG, X.-R.; LIN, C.-J. LIBLINEAR: A library for large linear classification. **Journal of Machine Learning Research**, New York, v. 9, n. 9, p. 1871-1874, Aug. 2008.

GEORGE, Gerard; OSINGA, Ernst C.; LAVIE, Dovev; SCOTT, Brent A. Big Data and Data Science Methods for Management Research. **Academy of Management Journal**, New York, v. 59, n. 5, p. 1493-1507, out. 2016. DOI 10.5465/amj.2016.4005. Disponível em: http://journals.aom.org/doi/10.5465/amj.2016.4005. Acesso em: 19 set. 2023.

GOLBECK, J.; ROBLES, C. G.; EDMONDSON, M.; TURNER, K. Predicting personality from twitter. *In:* IEEE THIRD INTERNATIONAL CONFERENCE ON PRIVACY, SECURITY, RISK AND TRUST AND 2011 IEEE

THIRD INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING, Boston, MA, 2011. **Proceedings** […]. New York: IEEE, 2012. p. 149-156. DOI 10.1109/PASSAT/SocialCom.2011.33. Disponível em: https://ieeexplore.ieee.org/document/6113107. Acesso em: 19 set. 2023.

HICKMAN, L.; THAPA, S.; TAY, L.; CAO, M.; SRINIVASAN, P. Text preprocessing for text mining in organizational research: Review and recommendations. **Organizational Research Methods**, Thousand Oaks, v. 25, n. 1, p. 114-146, 2022. Disponível em: https://doi.org/10.1177/1094428120971683. Acesso em: 19 set. 2023.

HUTCHINS, W. J. The Georgetown-IBM experiment demonstrated in January 1954. *In:* CONFERENCE OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS, 6th, Washington, DC, 2004. **Proceedings** [...]. Berlin: Springer, 2004. p. 102-114. Disponível em: https://link.springer.com/chapter/10.1007/978-3-540-30194-3_12. Acesso em: 19 set. 2023.

MAHYOOB, M.; AL-GARAADY, J.; ALBLWI, A.; ALRAHAILI, M. Sentiment analysis of public tweets towards the emergence of SARS-CoV-2 Omicron variant: a social media analytics framework. **Engineering, Technology & Applied Science Research**, Pátras, v. 12, n. 3, p. 8525-8531, June 2022.

NATIONAL SCIENCE BOARD. Science and Engineering Indicators 2018. Broad-based, objective information on the U.S. and international S7E enterprise. Alexandria, VA: National Science Foundation, 2018. Disponível em: https://www.nsf.gov/statistics/2018/nsb20181/. Acesso em: 19 set. 2023.

OLIVEIRA, D. J. S.; BERMEJO, P. H. S.; PEREIRA, J. R.; BARBOSA, D. A. A aplicação da técnica de análise de sentimentos em mídias sociais como instrumento para as práticas da gestão social em nível governamental. **Revista de Administração Pública**, Rio de Janeiro, v. 53, n. 1, p. 235-251, jan./fev. 2019. DOI 10.1590/0034-7612174204.

PAES, V. J.; ARAÚJO, D.; BRITO, K.; ANDRADE, E. Análise de sentimento em tweets relacionados ao desmatamento da floresta amazônica. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 11., Niterói, 2022. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 61-72. ISSN 2595-6094. Disponível em: https://sol.sbc.org.br/index.php/brasnam/article/view/20517. Acesso em: 19 set. 2023.

QAISER, S.; ALI, R. Text mining: use of TF-IDF to examine the relevance of words to documents. **International Journal of Computer Applications**, New York, v. 181, n. 1, p. 25-29, July 2018.

RAMOS, B.; FREITAS, C. "Sentimento de quê?": uma lista de sentimentos para a Análise de Sentimentos. *In*: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL), Oct. 15-18, 2019. **Anais** [...]. Salvador, BA, 2019. Disponível em: https://www.linguateca.pt/Repositorio/RamosFreitasSTIL2019.pdf. Acesso em: 19 set. 2023.

SHAUKAT, Z.; ZULFIQAR, A. A.; XIAO, C.; AZEEM, M.; MAHMOOD, T. Sentiment analysis on IMDB using lexicon and neural networks. **SN Applied Science,** Switzerland, v. 2, n. 148, p. 1-10, Jan. 2020. Disponível em: https://doi.org/10.1007/s42452-019-1926-x. Acesso em: 19 set. 2023.

SOUZA, Vinicius Augusto de; SOUZA, Érica Ferreira de; MEINERZ, Giovani Volnei. Análise de sentimento em tempo real de notícias do mercado de ações. **Brazilian Journal of Development**, Curitiba, v. 7, n. 1, p. 11084-11091, 2021. DOI 10.34117/bjdv7n1-758. Disponível em: https://www.brazilianjournals.com/index.php/BRJD/article/view/23959/19224. Acesso em: 19 set. 2023.

STENSON, Rob. Is This the First Time Anyone Printed, 'Garbage In, Garbage Out'?. **Atlas Obscura**. 14 Mar. 2016. Disponível em: http://www.atlasobscura.com/articles/is-this-the-first-time-anyone-printed-garbage-in-garbage-out. Acesso em: 19 set. 2023.

WANKHADE, Mayur; RAO, Annavarapu Chandra Sekhara; KULKARNI, Chaitanya. A survey on sentiment analysis methods, applications, and challenges. **Artificial Intelligence Review**, Berlin, v. 55, n. 7, p. 5731–5780, Oct. 2022. DOI 10.1007/s10462-022-10144-1. Disponível em: https://link.springer.com/10.1007/s10462-022-10144-1. Acesso em: 19 set. 2023.

DADOS DO AUTOR:

Guilherme Noronha

Guilherme Noronha é Graduado em Ciência da Computação pela PUC Minas com mestrado e doutorado em Gestão e Organização do Conhecimento pela UFMG. É entusiasta na área de dados com mais de dez anos de experiência e acumula trabalhos em diferentes áreas como proteção à privacidade, processamento de linguagem natural, aprendizado de máquina, modelagem, análise e engenharia de dados. Atualmente atuo como engenheiro de dados.

https://www.linkedin.com/in/noronha2001/
https://orcid.org/0000-0002-1422-2179
guilhermenoronha@2001@gmail.com

Como referenciar o capítulo 2:

NORONHA, Guilherme. Processamento de Linguagem Natural para análise de sentimento utilizando Python. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 2. p. 39-60. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap2

3. PYTHON COMO SUPORTE ÀS PESQUISAS SOCIAIS

Denise Fukumi Tsunoda
Alex Sebastião Constâncio

3.1 INTRODUÇÃO

A história ainda não registrou quantidade tão vasta de dados e informações disponíveis em bases de dados (estruturadas e não estruturadas) os *logs* de transações (registros históricos de operações de negócios), interações sociais, comportamentos, opiniões, dados pessoais e outros.

Se, por um lado, o volume de dados disponível oferece muitas oportunidades para a compreensão de diversos eventos e a sua utilização na tomada de melhores decisões em inúmeros âmbitos, é inegável que a quantidade, a variedade e as complexidades avassaladoras representam, em si mesmas, um desafio para o pesquisador, que precisa encontrar meios para realizar análises e identificar padrões de interesse. É intrigante, no entanto, observar que a tecnologia gerou este oceano de dados e provê os meios para estudá-los.

Este capítulo discorre sobre o uso da linguagem *Python* e algumas bibliotecas (em *Python* também conhecidas por pacotes) de seu ecossistema, como suporte às pesquisas sociais, e pontua locais de convergência entre a ciência social e algumas pesquisas com suporte de tecnologias de ponta.

Python é uma linguagem de programação versátil e poderosa, que com o tempo emergiu como uma ferramenta essencial para os pesquisadores de análise de dados. Apresenta capacidade de manipulação, análise e visualização de dados, e é classificada como uma linguagem de programação de alto nível, interpretada de propósito geral e de código aberto, além de ser reconhecida pela sintaxe simples e aprendizado fácil. Oferece, também, um rico conjunto de ferramentas de produtividade e recursos adicionais disponibilizados na forma de pacotes, que representam um arsenal de ferramentas para aplicação em múltiplos domínios do conhecimento.

A comunidade de *Python*[13] é extremamente ativa e é considerada uma das maiores e mais colaborativas comunidades de desenvolvedores de código aberto do mundo. Quando ela é mencionada, normalmente algumas características são destacadas: grande número de desenvolvedores, inúmeros fóruns de discussão (a exemplo do *Stack Overflow Python* e o *Reddit Python*), algumas conferências que compartilham experiências, como os eventos *PyCon* em todo o mundo, cria e difunde extensa documentação para diversos níveis (iniciantes até avançados), manutenção de repositório de pacotes (PyPI) que abriga milhares de bibliotecas e pacotes *Python* para compartilhamento de código entre desenvolvedores, diversos projetos compartilhados. Além disso, a comunidade é considerada acolhedora, no sentido de criar um ambiente agradável até para os menos experientes.

Especificamente no tópico conferências, o Brasil promove o evento *Python Brasil*[14] que, em 2023, acontecerá em novembro, em Caxias do Sul, que será

> [...] a maior conferência da linguagem de programação Python da América Latina, sendo um evento voltado à educação, treinamento e troca de experiências. Temos como máxima que as pessoas são maiores que tecnologia e queremos que todos que visitem o evento vivam essa máxima. Sob um código de conduta o evento preza por criar um ambiente seguro e convidativo a todas as pessoas (Python Brasil, 2023).

O evento tem como objetivos: difundir a linguagem *Python*; promover a troca de experiências e conhecimentos; incentivar o crescimento da comunidade nacional; incentivar o crescimento da comunidade regional; e impactar econômica e socialmente a região (Python Brasil, 2023).

Este capítulo explora como o *Python* e suas diversas bibliotecas/pacotes podem ser utilizados como suporte às pesquisas sociais. Desde a coleta de dados em tempo real de redes sociais até a análise avançada de texto para extrair *insights* profundos das opiniões públicas, apresenta-se um conjunto de ferramentas disponíveis para pesquisadores que pretendem,

13 Disponível em: https://www.python.org/community/forums/. Acesso em: 21 set. 2023.

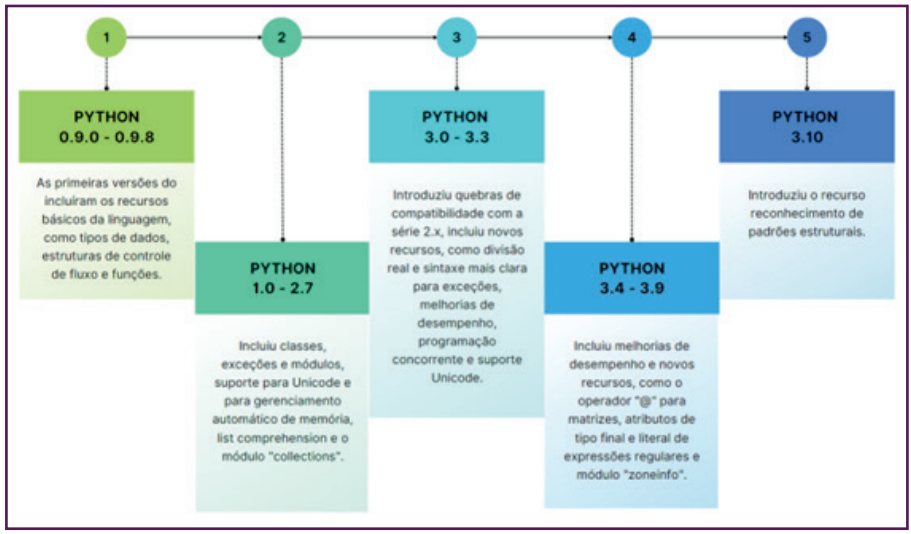
14 Python Brasil 2023: https://2023.pythonbrasil.org.br/#evento. Acesso em: 18 set. 2023.

por exemplo, realizar análise de dados, detectar tendências nos dados e visualizar resultados.

3.2 SOBRE A TECNOLOGIA

Desde a sua primeira versão, a linguagem *Python* evoluiu rápida e significativamente ao longo dos anos. Na Figura 1 podem ser vistas as mais importantes versões da linguagem e seus principais recursos ao longo dos anos.

Figura 1 - Linha do tempo com as versões mais importantes da linguagem



Fonte: Elaborado pelos autores (2023).

Alguns dos principais aspectos e características da linguagem *Python* incluem:

- **Multiplataforma:** é compatível com várias plataformas, incluindo *Windows*, *macOS* e várias distribuições de *Linux*, o que torna o seu código facilmente portável entre diferentes sistemas operacionais, pois algumas decisões dele visam não evidenciar as diferenças entre os ambientes de execução distintos.

- **Domínios:** é usado em uma variedade de domínios, incluindo desenvolvimento web (usando *frameworks* como *Django* e *Flask*), ciência de dados (com pacotes como *NumPy*, *Pandas* e *scikit-learn*), automação de tarefas, desenvolvimento de jogos, aprendizado de máquina etc.

- **Legibilidade:** enfatiza a legibilidade do código, encorajando o uso de uma sintaxe clara e fácil de entender, por meio de convenções devidamente documentadas; isso torna o código *Python* mais próximo da linguagem humana, facilitando sua colaboração e manutenção.

- **Versatilidade:** é usado em uma variedade de domínios, incluindo desenvolvimento web, automação, ciência de dados, aprendizado de máquina, automação de tarefas, desenvolvimento de jogos e muito mais; sua versatilidade é uma das razões para sua popularidade.

- **Interpretada:** possui uma linguagem interpretada, o que significa que o código é executado linha por linha por um interpretador em vez de ser compilado em código de máquina. Isso torna o desenvolvimento e a depuração mais rápidos, embora, em alguns casos, possa ser mais lento quando comparado às linguagens compiladas. No entanto, existe um projeto paralelo chamado de *Cython*, que oferece um compilador compatível com 99% da linguagem; o uso do *Cython* é uma alternativa viável para projetos críticos em tempo de processamento.

- **Extensibilidade:** suporta a criação de módulos em *C* e *C++*, permitindo que os desenvolvedores integrem facilmente códigos daquelas linguagens, quando necessário, para melhorar o desempenho ou acessar recursos específicos do sistema.

- **Pacotes:** dispõe de extensa variedade de pacotes que abrange várias tarefas, desde manipulação de arquivos até desenvolvimento web, análise de dados e muito mais; é o que torna o *Python* uma linguagem versátil e adequada à concepção de variados aplicativos.

- **Comunidade:** possui uma comunidade de desenvolvedores ativa e dedicada, o que resulta em uma grande quantidade de recursos, bibliotecas de terceiros e suporte on-line, facilitando a busca e recuperação de soluções para problemas específicos.

- ***Open Source:*** é distribuído sob uma licença de código aberto, o que significa que é gratuito para uso e pode ser modificado e distribuído livremente.

Python é uma linguagem versátil e poderosa, que atrai desenvolvedores de todas as áreas devido a sua simplicidade e eficácia. A facilidade no aprendizado, a ampla fonte de recursos na internet e o volume enorme de bibliotecas gratuitas à disposição também são atrativos para novos entusiastas e praticantes. A linguagem continua a evoluir e a crescer em popularidade, desempenhando papel significativo em uma ampla gama de campos tecnológicos e científicos.

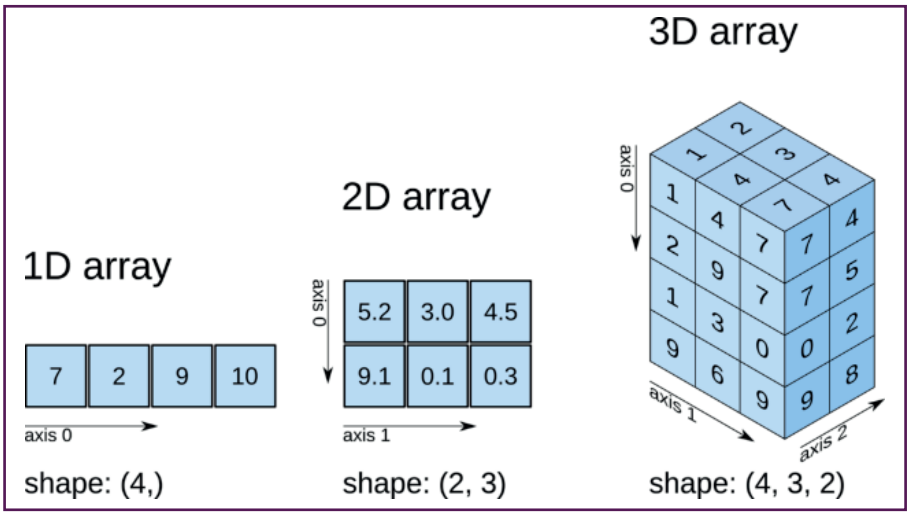
3.3 A TECNOLOGIA E A PESQUISA

Existem vários pacotes do *Python* que são úteis em pesquisas científicas das áreas sociais e humanas (Gholizadeh, 2022). A seguir, são citados alguns exemplos.

O *NumPy*[15] é um pacote essencial para computação científica, uma vez que fornece suporte para operação sobre dados multidimensionais, funções matemáticas avançadas e operações de álgebra linear. A Figura 2 apresenta estruturas em três tipos de dimensões: unidimensional (1D), bidimensional (2D) e tridimensional (3D). Na visão em 3D existe um detalhamento shape (4,3,2) que corresponde, respectivamente, à altura (eixo 0), à largura (eixo 1) e à profundidade (eixo 2).

15 Disponível em: https://numpy.org/. Acesso em: 21 set. 2023.

Figura 2 - Estruturas multidimensionais do *NumPy*



Fonte: NumPy (2023).

Muitos outros pacotes usam o *NumPy* para elaborar seus fundamentos, a exemplo da maioria das bibliotecas de processamento de dados, como *Matplotlib*, *Pandas* e *OpenCV*. Por esse motivo, o domínio do *NumPy* é um fator decisivo no desenvolvimento de qualquer projeto de análise de dados em *Python*. É útil para análise de dados quantitativos em áreas como psicologia, economia e sociologia.

Em agosto de 2023, foi anunciado o *site NumPy* em dois novos idiomas: japonês e português brasileiro[16]. Por ser um projeto de código aberto impulsionado pela comunidade e desenvolvido por colaboradores, diversos brasileiros foram responsáveis por traduzir grande parte do site e auxiliar na disseminação do *NumPy* no Brasil.

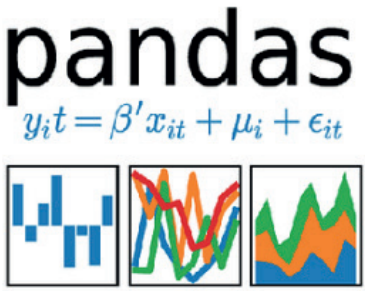
Os autores Galli *et al.* (2022) apresentam uma estrutura de referência para analisar e avaliar técnicas de aprendizado de máquina e aprendizado profundo para detectar notícias falsas, com foco na detecção precoce e na análise do conteúdo das notícias. Os autores abordam os desafios da

16 Título da notícia: "numpy.org agora está disponível em Japonês e Português". Publicado em: 2 ago. 2023. Disponível em: https://numpy.org/pt/news/. Acesso em: 21 set. 2023.

detecção de notícias falsas nas redes sociais, a exemplo da dificuldade de análise do conteúdo e os dados ruidosos e incompletos gerados pelas interações sociais. A implementação do módulo de aprendizagem profunda multimídia foi desenvolvida a partir do *Python 3* no *Jupyter* com as bibliotecas *Keras*[17], *Scikit*[18] e *Numpy*. Os resultados dos experimentos destacam o potencial do aprendizado de máquina e dos modelos de aprendizado profundo na detecção de notícias falsas.

O *Pandas*[19] é um pacote para manipulação e análise de dados que fornece estruturas de dados flexíveis e eficientes para trabalhar com tabelas e séries temporais. É útil para análise de dados qualitativos em áreas como antropologia, história e ciência política. Na Figura 3 podem ser vistos dois dos recursos mais utilizados do *Pandas*, a geração de gráficos de vários tipos (barras, linhas, área, pizza, *boxplot*, dentre muitos outros) e o *DataFrame*, uma estrutura tabular que oferece facilidades para manipulação em diversas dimensões, útil tanto para o processamento quanto para a apresentação de dados.

Figura 3 - Gráfico e DataFrame do Pandas



	BandName	WavelengthMax	WavelengthMin
0	CoastalAerosol	450	430
1	Blue	510	450
2	Green	590	530
3	Red	670	640
4	NearInfrared	880	850
5	ShortWaveInfrared_1	1650	1570
6	ShortWaveInfrared_2	2290	2110
7	Cirrus	1380	1360

Fonte: Towards data science (2019)[20].

17 Disponível em: https://keras.io/. Acesso em: 21 set. 2023.

18 Disponível em: https://scikit-learn.org/stable/. Acesso em: 21 set. 2023.

19 Disponível em: https://pandas.pydata.org/. Acesso em: 21 set. 2023.

20 Disponível em: https://towardsdatascience.com/manipulating-the-data-with-pandas--using-python-be6c5dfabd47. Acesso em: 3 out. 2023.

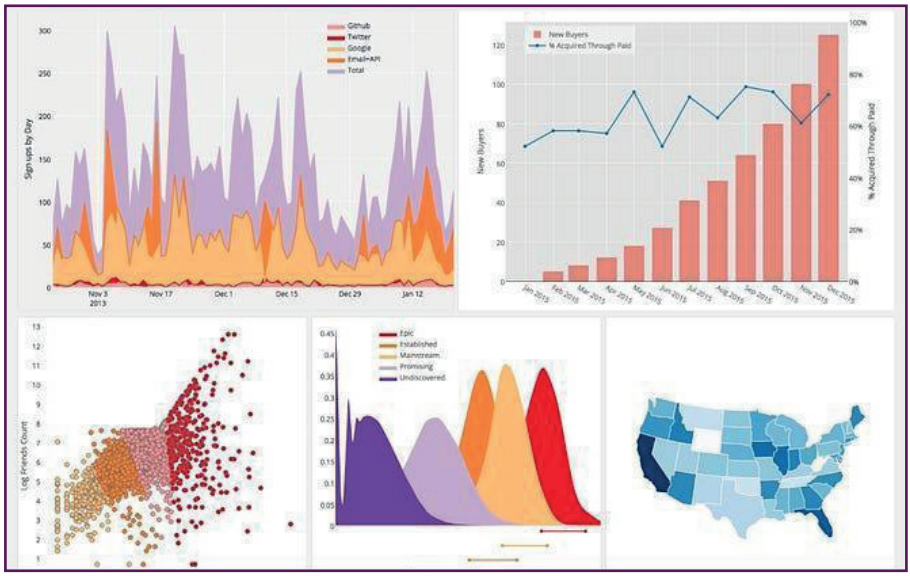
O *Pandas* é bastante utilizado nas tarefas de processamento de dados, tais como limpeza, manipulação e análise, uma vez que dispõe de vários módulos para leitura, processamento e gravação de arquivos *CSV* (Comma-separated values), *JSON* (JavaScript Object Notation) e *Excel* (Microsoft Excel). Obviamente muitas ferramentas de limpeza de dados estão disponíveis, mas, segundo Zia *et al.* (2022, p. 12, tradução nossa), "[...] o gerenciamento e a exploração de dados com a biblioteca Pandas são incrivelmente rápidos e eficazes"[21]. Nessa obra, os autores abordam de forma crítica a mineração de dados médicos baseada em inteligência artificial, a qual definem como "o processo de extrair informações valiosas e *insights* de grandes conjuntos de dados médicos usando algoritmos e técnicas de aprendizado de máquina". O artigo também explora os benefícios e desafios de tal abordagem, bem como suas aplicações práticas.

O *Matplotlib*[22] é um pacote de visualização de dados que fornece ferramentas para criar gráficos em 2D e 3D, histogramas, gráficos de barras e muito mais, conforme ilustrado na Figura 4. É útil para apresentar resultados de pesquisa em áreas como geografia, arqueologia e linguística.

21 Trecho original: [...] *managing and exploring data with the Pandas library is incredibly quick and effective*.

22 Disponível em: https://matplotlib.org/. Acesso em: 21 set. 2023.

Figura 4 - Exemplos de gráficos do *Matplotlib*



Fonte: Medium (2020)[23].

Dentre as diversas aplicações, destacam-se algumas citadas no próprio site do pacote: criação de gráficos com qualidade de publicação, inclusive se a pesquisa envolver a coleta de dados ao longo do tempo; o *Matplotlib*, que pode ser usado na criação de gráficos de séries temporais com destaque para as tendências ao longo do tempo, como mudanças nas opiniões públicas ou padrões de comportamento; criação de figuras interativas, que podem ser ampliadas, deslocadas, atualizadas e utilizadas em relatórios ativos (interativos); concepção de mapas de calor, por exemplo, para destacar áreas com maior criminalidade em um município, ou qualquer outro fenômeno social.

O *Seaborn*[24] é um pacote de visualização de dados baseado no *Matplotlib* que fornece uma interface de alto nível para criar gráficos estatísticos atraentes. É útil para visualizar dados em áreas como psicologia social, educação e estudos culturais.

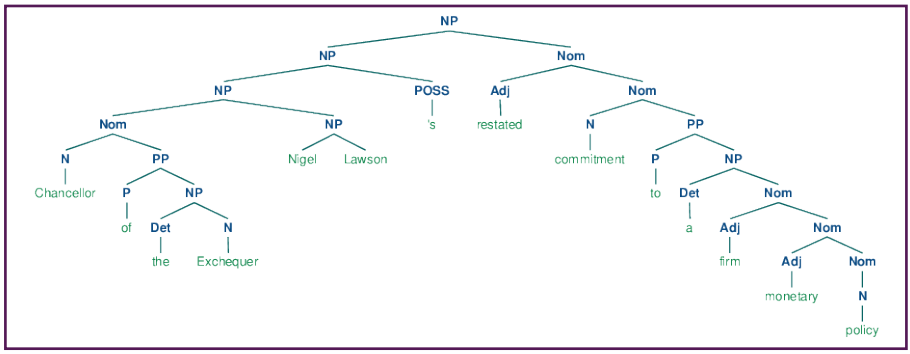
23 Disponível em: https://i.imgur.com/2AE3gch.png. Acesso em: 2 out. 2023.

24 Disponível em: https://seaborn.pydata.org/. Acesso em: 21 set. 2023.

Os autores Liu *et al.* (2022) apresentaram um "modelo de inferioridade" construído a partir de dados coletados em mídia social e aplicados para identificar as causas desses sentimentos. Para o estudo foram usados conjuntos de dados públicos retirados de um site chamado *Weibo* (uma mídia social utilizada na China). Nesse caso, os usuários são informados que podem privatizar ou divulgar suas postagens conforme desejarem, mas as postagens públicas podem ser visualizadas e baixadas pela plataforma por eventuais interessados. O processo de aquisição de dados consistiu em duas etapas: coleta e pré-processamento de dados. Para cada ano, foram extraídas 200 publicações (em um total de 1.400 publicações), que revelavam as causas dos sentimentos de inferioridade. A pesquisa apresenta um código de temas que foi utilizado para rotulação manual realizada por três colaboradores: sentimentos de inferioridade em relação a defeitos físicos; sentimentos de inferioridade em relação a amor e afeto; sentimentos de inferioridade em relação ao histórico familiar; sentimentos de inferioridade em relação à personalidade; sentimentos de inferioridade sobre experiências pessoais; sentimentos de inferioridade sobre interação social; sentimentos de inferioridade sobre aprendizado; e sentimentos de inferioridade sobre habilidades. A ferramenta *Seaborn* foi utilizada para apresentação dos mapas de sentimentos de inferioridade de cada um dos oito códigos utilizados. Os resultados da análise temática mostram que os sentimentos de inferioridade decorrem principalmente da experiência, defeito físico, personalidade, relacionamento amoroso, habilidade, interação social etc.

Os pacotes *NLTK, TextBlob* e *SpaCy* são utilizados para Processamento de Linguagem Natural (PLN) e fornecem ferramentas para pré-processamento (tokenização), análise sintática e semântica, análise de sentimentos e emoção, classificação de texto e muito mais. São úteis para análise de texto em áreas como literatura, ciência da comunicação e ciência política. Na Figura 5 exemplifica-se uma árvore sintática produzida pelo *NLTK* após o processamento do texto.

Figura 5 - Estrutura sintática gerada pelo *NLTK*



Fonte: Update for NLTK 3.0 (2019)[25].

Aline *et al.* (2023) exploraram as crenças dos consumidores sobre os riscos à saúde de alimentos para bebês em análises sobre dados coletados em fóruns disponíveis para os pais, no Reino Unido. Após selecionar um subconjunto de postagens e classificá-las por tópico, de acordo com o produto alimentício discutido e o seu risco à saúde, foram realizados dois tipos de análises: "correlação de Pearson" de ocorrências de termos para destacar pares de "perigo-produto" (harzard-product) predominantes e regressão de Mínimos Quadrados Ordinários (do inglês *Ordinary Least Squares* - OLS) realizada em medidas de sentimentos geradas a partir dos textos que indicou sentimento positivo ou negativo, linguagem objetiva ou subjetiva e modalidade confiante ou não confiante associada a diferentes produtos alimentícios e riscos à saúde. Especificamente nessa etapa, duas ferramentas foram usadas para calcular as métricas de sentimento: o pacote *VADER* (Hutto; Gilbert, 2014) do *NLTK* (Bird; Klein; Loper, 2009) e o módulo *PATTERN* (De Smedt; Daelemans, 2012). Os autores concluem que as métricas de sentimentos foram essenciais para oferecer respostas sobre as percepções dos pais em relação aos riscos de segurança dos produtos alimentícios para bebês.

Os autores Praveen *et al.* (2022) analisaram as percepções dos indianos sobre as vacinas com doses de reforço contra a Covid-19 usando técnicas de processamento de linguagem natural. Os investigadores analisaram

25 *NLTK* (Natural Language Toolkit). Disponível em: https://www.nltk.org/book/ch08-extras.html. Acesso em: 3 out. 2023.

tweets gerados por cidadãos indianos e descobriram que uma parte significativa dos tweets apresentava sentimentos negativos em relação às doses de reforço. O estudo também revelou que as postagens dos indianos nas redes sociais se concentravam na crença de que os mais jovens não precisam de vacinas e que as vacinas não são saudáveis. Para tal estudo, os autores utilizaram, dentre outras ferramentas, o pacote *TextBlob* para a análise de sentimentos (positivos, negativos ou neutros) dos cidadãos.

Outro exemplo de aplicação é a análise de *posts* em português feitos no Twitter e em jornais a respeito da pandemia de Covid-19 (Melo; Figueiredo, 2021). O estudo faz uso de uma biblioteca de terceiros para *Python* chamada *Tweeter-Scrapper*[26], cuja função é localizar e extrair postagens no Twitter. Também incluíram artigos publicados nas páginas da web do jornal Folha de São Paulo. Embora os autores não tenham especificado a biblioteca que foi utilizada, qualquer uma das bibliotecas do *Python* para *web scrapping* (como o *Scrapy*) se prestaria a essa etapa do estudo.

O processo faz uso de análise de sentimentos combinada com identificação de entidades (uma atividade do PLN que analisa automaticamente aquelas palavras e identifica entidades, como pessoas, empresas, países, dentre outros). Para a primeira atividade, foi empregado o *Vader* (Valence Aware Dictionary and Sentiment Reasoner), enquanto para a segunda foi escolhido o *SpaCy*. Os dados foram mostrados em gráficos gerados pelo *Seaborn*, mas algumas análises foram feitas com base em nuvens de palavras construídas com o pacote *Word-Cloud*[27].

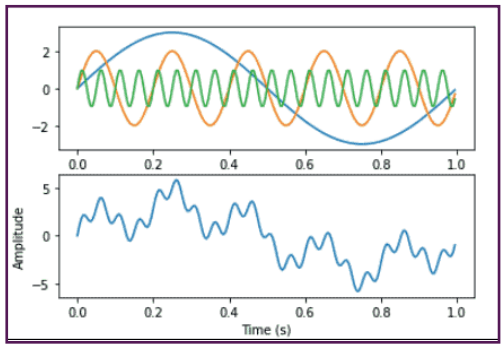
O *SciPy*[28] é um pacote para computação científica que fornece ferramentas de integração numérica, otimização, interpolação, processamento de sinais e muito mais. É útil em áreas como economia, psicologia e geografia. A Figura 6 apresenta uma visualização, usando *Matplotlib*, de um exemplo de transformada de *Fourier* do pacote *SciPy*.

26 Disponível em: https://github.com/bisguzar/twitter-scraper. Acesso em: 21 set. 2023.

27 Disponível em: https://github.com/amueller/word_cloud. Acesso em: 21 set. 2023.

28 Disponível em: https://scipy.org/. Acesso em: 21 set. 2023.

Figura 6 - Visualizações no pacote SciPy



Fonte: Phoenixnap (2021)[29].

Cohen *et al.* (2022) mencionam que os departamentos de emergência (DE) são pontos de identificação de risco de suicídio e poderiam direcionar os pacientes aos cuidados necessários. No entanto, as ferramentas de "triagem" adotadas não estão centradas na pessoa e não utilizam tecnologias inovadoras, a exemplo das técnicas de aprendizado de máquina (inglês *Machine Learning - ML*) baseadas em Processamento de Linguagem Natural (inglês *Natural Language Processing - NLP*), que

> [...] têm se mostrado promissoras para avaliar o risco de suicídio, embora não se saiba se os modelos de NLP têm bom desempenho em diferentes regiões geográficas, em diferentes períodos de tempo ou após eventos de grande escala [...] (Cohen *et al.*, 2022, p. 1, tradução nossa)[30].

Dessa forma, os autores avaliam o desempenho de um modelo *NLP/ML* em um *corpus* coletado no sudeste dos Estados Unidos por meio de modelos previamente testados no centro-oeste dos EUA. Assim, 37 pacientes suicidas e 33 não suicidas de dois departamentos de emergência foram entrevistados para testar o modelo *NLP/ML* de previsão de risco de suicídio desenvolvido anteriormente. Para a análise de dados foi utilizada

29 *SCIPY TUTORIAL*. Disponível em: https://phoenixnap.com/kb/scipy-tutorial. Acesso em: 3 out. 2023.

30 Trecho original: [...] *Natural language processing (NPL) -based machine learning (ML) techniques have shown promise to assess suicide risk, although whether NPL models perform well in differing geographic regions, at different time periods, or after large-scale events* [...].

linguagem *Python* e os pacotes *Pandas*, *Numpy*, *scikit-learn*, *Matplotlib*, *SciPy* e *NLTK*. Os autores concluem que o modelo de risco de suicídio baseado no idioma teve um bom desempenho ao identificar o idioma dos pacientes suicidas de uma parte diferente dos EUA e em um período de tempo posterior àquele em que o modelo foi originalmente desenvolvido e treinado.

O *Statsmodels*[31] é um pacote para modelagem estatística que fornece ferramentas para análise de regressão, análise de séries temporais, testes de hipóteses (Seabold; Perktold, 2010), com suporte específico para modelagem econométrica e estatística. É útil em áreas como sociologia, ciência política e criminologia. Trata-se de mais um pacote que está construído sobre o *NumPy* e *SciPy*, de modo que se integra diretamente a atividades que utilizam recursos destes.

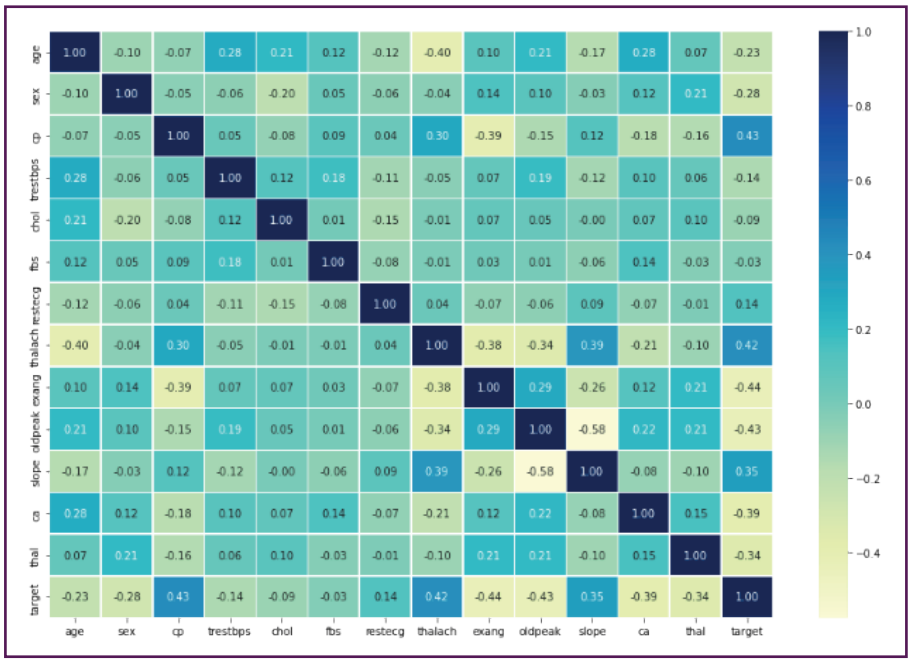
O estudo de Aparício, Romão e Costa (2022) propôs um modelo preditivo para a oscilação de valores de *Bitcoin*. Em uma das etapas, os autores precisaram construir um modelo de estimativa baseado em regressão linear pelo método dos mínimos quadrados utilizando recursos combinados do *Pandas* e do *Statsmodels*.

O *Scikit-learn*[32] oferece recursos para aprendizado de máquina, com ferramentas para classificação, regressão, agrupamento, pré-processamento, visualização (como exemplificado pela matriz de confusão visível na Figura 7, muito comum para avaliar o desempenho de alguns modelos de aprendizado de máquina) e muito mais. É útil em áreas como psicologia, ciência política e economia.

31 Disponível em: https://www.statsmodels.org/stable/index.html. Acesso em: 21 set. 2023.

32 Disponível em: https://scikit-learn.org/stable/. Acesso em: 21 set. 2023.

Figura 7 - Matriz de confusão gerada pelo *Scikit-learn*



Fonte: Infoslack (2023)[33].

O "sentimento do investidor" é fundamental no mercado de ações e, nos últimos anos, vários estudos buscaram prever os preços futuros das ações analisando o sentimento do mercado obtido da mídia social ou por meio das notícias veiculadas. Liu, Leu e Holst (2023) investigam o uso do "sentimento do investidor" nas mídias sociais, com foco no *Stocktwits*, uma plataforma de mídia social para investidores. O estudo propõe uma máquina de vetores de suporte (SVM) com *bagging* para melhorar a precisão das previsões de movimentação de preços de ações e adota uma abordagem que utiliza o *FinBERT*, um modelo de linguagem pré-treinado e projetado especificamente para analisar o sentimento do contexto financeiro. *Bagging* é um tipo de aprendizado que combina múltiplos modelos para fazer previsões mais precisas, e os autores utilizaram o *BaggingClassifier*, disponível no *Scikit-learn* para tal finalidade. O estudo revela que o uso do

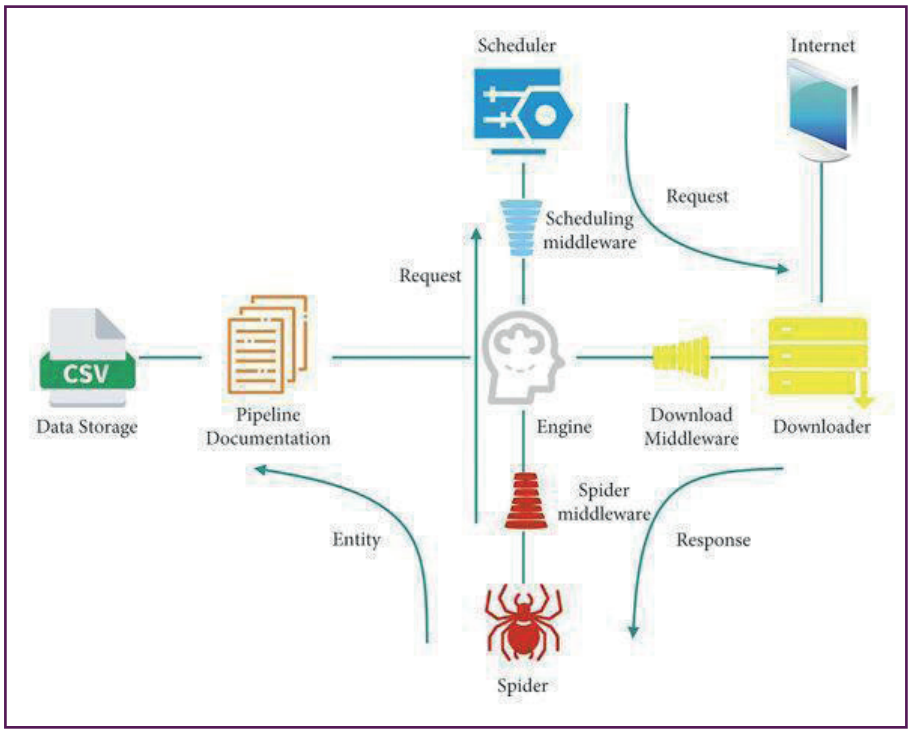
33 *INFOSLACK*. Disponível em: https://infoslack.pro/ml-book/contents/ml-sklearn.html. Acesso em: 3 out. 2023.

modelo *FinBERT* para análise de sentimento produz melhores resultados quando comparado a outras abordagens.

O *Scrapy*[34] é um pacote rastreamento de sites da web e extração de dados estruturados que podem ser usados para uma ampla gama de aplicativos, a exemplo de mineração de dados, processamento de informações ou apenas arquivamento. Os mantenedores mencionam que, embora o *Scrapy* tenha sido originalmente projetado para "raspagem da web", também pode ser usado para extrair dados usando *APIs* (como *Amazon Associates Web Services*) ou como rastreador da Web de uso geral. É útil em áreas como ciência política, sociologia e estudos culturais. A Figura 8 apresenta o modelo de operação do Scrapy, de Wang *et al.* (2022), para rastrear dados relevantes em páginas da web relevantes, como *Wikipedia*, *Baidu Encyclopedia* e *Military News Network* sobre cadeias de destruição militares (que consistem em equipamentos de controle, sensores, ataque e avaliação) com o propósito de construir um gráfico de conhecimento de domínio.

34 Disponível em: https://scrapy.org/. Acesso em: 21 set. 2023.

Figura 8 - Modelo de operação do Scrapy



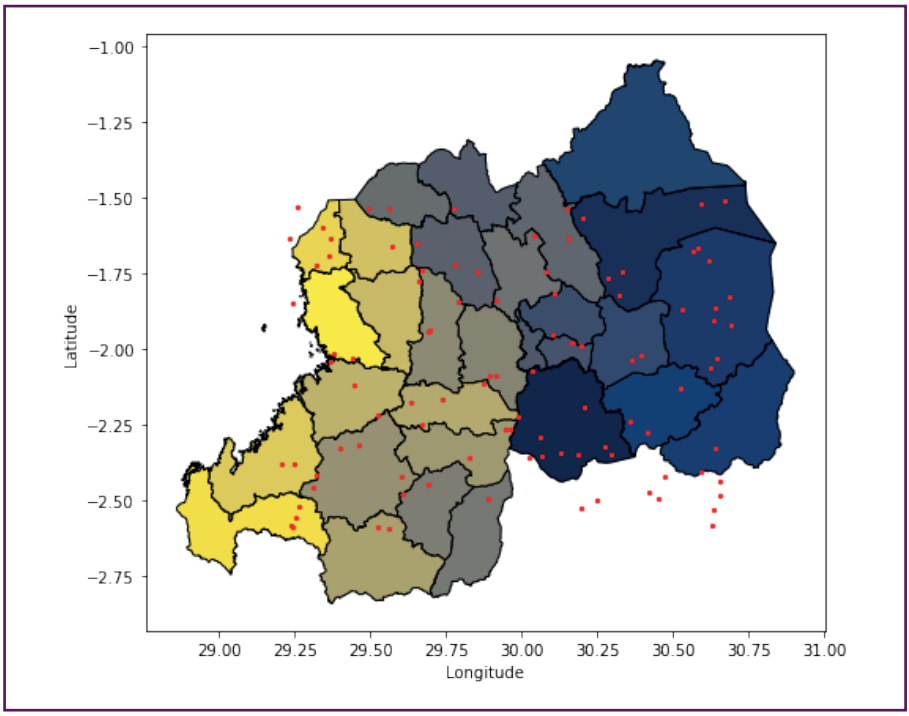
Fonte: Wang *et al.* (2022).

Com a onipresença da internet, muitos anúncios de vagas de empregos são veiculados em sites diversos. No entanto, nem sempre as descrições são claras, não existe análise geral de tendências dos setores ou adoção de métodos de visualização de dados que auxiliem na escolha e submissão de currículos. Considerando tais aspectos, Zihan, Yanhong e Hongtai (2021) rastrearam anúncios de vagas de emprego com base na estrutura do *Scrapy*, armazenam em *Excel* após a limpeza, utilizaram o *algoritmo Apriori* para descoberta de conexões entre diferentes dados e, por fim, sugerem um método de visualização adequado para diferentes tipos de informações de trabalho. Os autores pontuam que o método proposto pode auxiliar os candidatos a emprego a entender "[...] a demanda atual de talentos do setor de forma simples, intuitiva e rápida, mas também tem algum significado orientador para o programa de treinamento de talentos das universidades [...]" (Zihan; Yanhong; Hongtai,

2021, p. 1, tradução nossa)[35]. O estudo destaca o papel do *web scraping* na automação da extração de dados para análise do mercado de trabalho e qualificações dos candidatos.

Folium e *Geopandas*: são pacotes que permitem a criação de mapas interativos e a manipulação de dados geoespaciais, envolvendo localização de eventos e padrões de mobilidade demográfica. São úteis em áreas como ciência política, sociologia e marketing. Na Figura 9 está exemplificado um mapa gerado pelo GeoPandas, com pontos de interesse sinalizados em vermelho.

Figura 9 - Mapa gerado pelo *Geopandas*



Fonte: GEOPANDAS (2020)[36].

35 Trecho original: [...] *the current talent demand of the industry in a simple, intuitive and fast way, but also has some guiding significance for the talent training program of universities* [...].

36 GEOPANDAS. Disponível em: https://www.linkedin.com/pulse/geopandas-plotting-data-points-map-using-python-r%C3%A9gis-nisengwe/. Acesso em: 3 out. 2023.

Segundo García-Madurga, Grilló-Méndez e Esteban-Navarro (2020), a área conhecida por Inteligência Territorial é uma prática dedicada a obter, analisar e valorizar informações e conhecimentos sobre um território e seu ambiente, com o objetivo de projetar e implementar planos territoriais em questões estratégicas para tomada de decisão. Os autores mencionam que os primeiros registros de pesquisa na temática surgiram na França, como uma aplicação da Inteligência Econômica, mas já é considerada uma disciplina autônoma, que origina aplicações específicas, como a Inteligência Turística em diversos países.

Um exemplo de aplicação nessa área são os sistemas de compartilhamento gratuito de bicicletas, que podem ter influência positiva na mobilidade dos centros urbanos, desde que exista a preocupação com o desenvolvimento de estratégias de localização eficientes a fim de evitar aglomerações nos horários de pico e aumentar a disponibilidade do serviço. Rojas *et al.* (2023) destacaram como resolver a localização de estações virtuais de bicicletas em uma cidade latino-americana virtual por meio de uma metodologia de organização de dados geoespaciais. A solução foi implementada em *Python* com o uso das bibliotecas *Geopandas* e *LocalSolver* para determinar os locais das estações de bicicletas virtuais que maximizam a demanda potencial prevista para a cidade. O protótipo do sistema de suporte à decisão fornece uma recomendação sobre onde as estações de bicicletas virtuais devem ser localizadas durante os horários de pico e, conforme relato dos autores, melhora a disponibilidade geral em mais de 37%.

Pesquisadores frequentemente usam *Python* para analisar o sentimento de postagens em mídias sociais, identificando a polaridade das opiniões em relação a tópicos específicos. Assim, a análise de polaridade envolve a classificação dos termos em positivos, negativos ou neutros, para que seja possível determinar o sentimento/opinião geral do texto. Essa classificação é feita por meio de modelos estatísticos que utilizam técnicas de processamento de linguagem natural.

Algumas das já explicadas bibliotecas, tais como *spaCY*, *NLTK* e *TextBlob*, são frequentemente utilizadas para analisar o sentimento expresso em *tweets*, postagens no *Facebook* e outros dados de mídias sociais para entender os sentimentos das pessoas em relação aos mais diversos assuntos: marcas, artigos, produtos, serviços, política, religião, esportes, medicamentos, suplementos, cursos etc.

Por sua vez, Pereira (2021) publicou um trabalho com o título *A Survey of Sentiment Analysis in the Portuguese Language*, no qual destaca 11 ferramentas que podem ser utilizadas para análise de sentimentos em língua portuguesa (além de outros idiomas), a exemplo de:

a) *NLPnet*[37], a biblioteca *Python* para tarefas de processamento de linguagem natural (PLN) que utiliza redes neurais e apresenta funcionalidades, tais como: marcação de parte da fala, rotulagem de função semântica e análise de dependência para realizar a análise de sentimentos, incluindo o pré-processamento de textos, a identificação de palavras-chaves e a análise de polaridade.

b) *spaCY*: a biblioteca de PLN em *Python* utilizada para construção de aplicativos de reconhecimento de linguagem natural (Honnibal, 2016);

c) *NLTK* (Natural Language Toolkit): a biblioteca em *Python* que fornece ferramentas para PLN, como *tokenização*, *stemming*, *lematização*, etiquetagem de partes do discurso, análise sintática, entre outras (Hardeniya *et al.*, 2016). Em 2017, Barbosa *et al.* (2017) apresentaram, na III Escola Regional de Informática do Piauí, o minicurso *Introdução ao Processamento de Linguagem Natural Usando Python* e detalharam o uso do *NLTK* em todas as etapas do PLN.

Por outro lado, em diversas situações, existe o uso combinado de pacotes. Por exemplo, caso a pesquisa envolva a análise de dados textuais, o *NumPy* pode ser utilizado em conjunto com outras bibliotecas de processamento de texto (como *NLTK* ou *spaCy*) para preparar e analisar dados de texto coletados em pesquisas sociais.

Ainda em análise de redes sociais, a linguagem *Python* pode ser utilizada para identificar influenciadores, calcular métricas de centralidade e detectar comunidades em redes como o Twitter e o Facebook com pacotes tais como o *NetworkX*[38] (Harbeg; Schult; Swart, 2008).

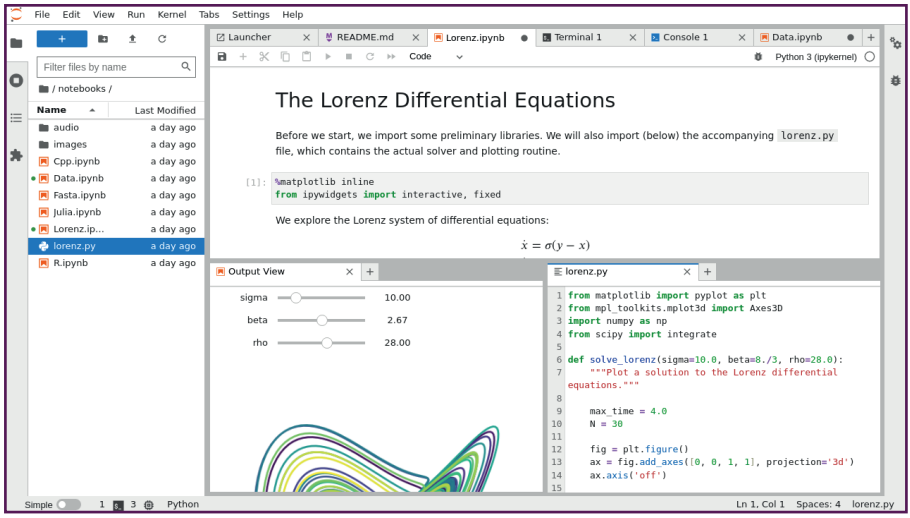
37 Disponível em: https://pypi.org/project/nlpnet/. Acesso em: 21 set. 2023.

38 Disponível em: https://networkx.org/. Acesso em: 21 set. 2023.

Para aplicações específicas, a exemplo de inteligência esportiva, os pesquisadores criam soluções próprias em *Python*, como Mallepalle *et al.* (2019), que desenvolveram um *software* para processamento de imagem, de código aberto, e projetado especificamente para extrair dados de rastreamento da *National Football League* (NFL) a partir de imagens, a fim de avaliar *quarterbacks* e defesas de passes. A ferramenta chamada de *next-gen-scraPy* permite extrair dados brutos dos gráficos de passes da *NFL*, incluindo o resultado do passe e a localização em campo e viabiliza a realização de análises mais aprofundadas sobre a eficiência dos *quarterbacks* e as defesas de passes na *NFL*.

Além dos pacotes, o ecossistema do *Python* é composto por diversas ferramentas de apoio, como, por exemplo, o *Jupyter* (Figura 10). Trata-se de um ambiente de programação interativo em que o programador consegue executar comandos com resposta imediata, envolvendo a manipulação e visualização de dados de forma instantânea.

Figura 10 - Exemplo de interface de um notebook de programação do *Jupyter*



Fonte: Jupyter (2023).

O *Jupyter* acelera a análise exploratória de dados, pois permite um fluxo recorrente de comandos e respostas que retroalimenta o estudo. Os dados

podem ser manipulados e o *Jupyter* permite que os resultados sejam salvos para observação ou aprimoramento posterior.

Todos esses recursos são disponibilizados gratuitamente ao pesquisador no *PyPi*[39], o repositório global do *Python*, que conta com dezenas de milhares de pacotes prontos para serem instalados e aproveitados. No entanto, muito tempo pode ser economizado com o *Anaconda*[40], uma distribuição de milhares de pacotes previamente selecionados para tarefas de análise de dados.

Com o *Anaconda*, os pacotes já são instalados e um ambiente pronto com as ferramentas de produtividade mais comumente utilizadas é preparado para o pesquisador. Já os pacotes específicos não presentes na seleção do *Anaconda* podem ser instalados a qualquer momento a partir do *PyPi*.

Uma alternativa para evitar a instalação do *Anaconda* ou do *Python* no computador local é fazer uso do *Google Colab*[41], plataforma gratuita fornecida pela *Google* e de operação baseada em navegador da internet que é muito semelhante ao *Jupyter*. Com o *Google Colab* é possível fazer o *upload* de arquivos de dados e de código *Python* e escrever trechos que são executados imediatamente.

O *Google Colab* oferece o mesmo tipo de fluxo de trabalho que o *Jupyter*, com a vantagem de não necessitar instalações locais e com a disponibilização de *GPUs* (Graphics Processing Units) para suportar cálculos mais exigentes.

3.4 CONSIDERAÇÕES FINAIS

À medida que encerramos este capítulo sobre o uso de *Python* como suporte às pesquisas sociais, esperamos ter evidenciado que tal linguagem

39 Disponível em: https://pypi.org/. Acesso em: 21 set. 2023.

40 Disponível em: https://www.anaconda.com/. Acesso em: 21 set. 2023.

41 Disponível em: https://colab.research.google.com/?utm_source=scs-index. Acesso em: 21 set. 2023.

de programação se tornou uma ferramenta indispensável para os pesquisadores que buscam compreender os intrincados meandros da sociedade contemporânea. Esperamos ter relevado uma estrutura Python com sua simplicidade, versatilidade e robustez, a qual pode servir como ferramenta para pesquisadores de diversos campos sociais. O poder dessa linguagem de programação se estende muito além do desenvolvimento de software convencional, e sua influência é profunda na análise e compreensão dos complexos aspectos das sociedades humanas.

Neste capítulo exploramos as várias maneiras pelas quais *Python* pode ser aplicado, desde a coleta e a análise de dados de mídias sociais até a compreensão de padrões demográficos para inteligência territorial, análises de opiniões e aplicações específicas, como avaliação de passe de jogadores. Apresentamos como as bibliotecas de processamento de linguagem natural, visualização de dados e aprendizado de máquina ampliam as capacidades de análise e permitem que os pesquisadores extraiam *insights* profundos de conjuntos de dados sociais cada vez maiores.

A riqueza da biblioteca padrão e a abordagem *open source* da linguagem *Python* proporcionam um ambiente propício para a colaboração e a inovação. Além disso, as inúmeras bibliotecas de terceiros criadas pela comunidade expandem ainda mais as capacidades de *Python* em áreas específicas da pesquisa social. Ainda sobre a comunidade, destacamos o crescimento e a colaboração contínuos de que compartilham aplicativos, bibliotecas, tutoriais e suporte técnico, tornando *Python* acessível para todos, mesmo para iniciantes interessados no seu aprendizado. Por isso, encorajamos pesquisadores interessados a explorar e dominar tal linguagem, pois ela pode abrir portas para descobertas e inovações que moldarão o futuro da nossa sociedade.

Finalmente, pontuamos que, à medida que a tecnologia evolui, novas questões/inquietações éticas e de privacidade aparecem de forma mais complexa. Ainda que o desenvolvimento desenfreado de aplicações esteja ocorrendo, é imperativo que os pesquisadores sociais utilizem qualquer ferramenta tecnológica com responsabilidade e considerem cuidadosamente as consequências e implicações éticas de suas pesquisas.

REFERÊNCIAS

ALINE, Sherman *et al.* Infant food users´ perceptions of safety: a web-based analysis approach. **Frontiers in Artificial Intelligence**, [*s. l.*], v. 6, 2023. DOI: https://doi.org/10.3389/frai.2023.1080950. Disponível em: https://www.frontiersin.org/articles/10.3389/frai.2023.1080950/full. Acesso em: 3 out. 2023.

APARICIO, João Tiago; ROMAO, Mario; COSTA, Carlos J. Predicting Bitcoin prices: the effect of interest rate, search on the internet, and energy prices. *In*: IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES, 17th., 2022, Madrid, Spain. **Proceedings [...].** [*S. l.*]: IEEE, 2022. DOI: https://doi.org/10.23919/CISTI54924.2022.9820085. Disponível em: https://ieeexplore.ieee.org/document/9820085. Acesso em: 21 set. 2023.

BARBOSA, Jardeson Leandro Nascimento *et al.* Introdução ao processamento de linguagem natural usando Python. *In*: **III Escola Regional de Informática do Piauí**: Anais, Artigos e Minicursos, v. 1, n. 1, p. 336-360, jun. 2017. Disponível em: https://www.facom.ufu.br/~wendelmelo/terceiros/tutorial_nltk.pdf. Acesso em: 21 set. 2023.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with python**: analyzing text with the natural language toolkit. Sebastopol, CA: O'Reilly Media, Inc., 2009.

COHEN, Joshua *et al.* Integration and validation of a natural language processing machine learning suicide risk prediction model based on open-ended interview language in the emergency department. **Frontiers in Digital Health**, [*s. l.*], Feb., 2022. DOI: https://doi.org/10.3389/fdgth.2022.818705. Disponível em: https://pubmed.ncbi.nlm.nih.gov/35187527/. Acesso em: 17 set. 2023.

DE SMEDT, Tom; DAELEMANS, Walter. Pattern for python. **Journal of Machine Learning Research**, [*s. l.*], v. 13, p. 2063-2067, 2012. Disponível em: https://jmlr.org/papers/v13/desmedt12a.html. Acesso em: 17 set. 2023.

GALLI, Antonio *et al.* A comprehensive benchmark for fake news detection. **Journal of Intelligent Information Systems**, [*s. l.*], v. 59, n. 1, p. 237-261, Aug. 2022. DOI: https://doi.org/10.1007/s10844-021-00646-9.

Disponível em: https://link.springer.com/content/pdf/10.1007/s10844-021-00646-9. Acesso em: 17 set. 2023.

GARCÍA-MADURGA, Miguel-Ángel; GRILLÓ-MÉNDEZ, Ana-Julia; ESTEBAN-NAVARRO, Miguel-Ángelo. Territorial intelligence, a collective challenge for sustainable development: a scoping review. **Social Sciences, MDPI**, Basel, v. 9, n. 7, 2020. DOI: https://doi.org/10.3390/socsci9070126. Disponível em: https://ideas.repec.org/a/gam/jscscx/v9y2020i7p126-d387607.html. Acesso em: 18 set. 2023.

GHOLIZADEH, Samira. **Top popular Python libraries in research**. [*S. l.*]: Authorea, Feb. 25, 2022. [8 p.] DOI: https://doi.org/10.22541/au.164580055.55493761/v1. Disponível em: https://www.authorea.com/doi/full/10.22541/au.164580055.55493761/v1. Acesso em: 21 set. 2023.

HAGBERG, Aric A.; SCHULT, Daniel A.; SWART, Pieter J. Exploring network structure, dynamics, and function using NetworkX. *In*: VAROQUAUX, Gaël; VAUGHT, Travis; MILLMAN, Jarrod (ed.). PYTHON IN SCIENCE CONFERENCE (SCIPY2008), 7th, Pasadena, 2008. **Proceedings [...]**. Pasadena, CA: SciPy, 2008. p. 11-15. Disponível em: https://conference.scipy.org/proceedings/SciPy2008/paper_2/full_text.pdf. Acesso em: 21 set. 2023.

HARDENIYA, Nitin *et al.* **Natural language processing**: python and NLTK. Birmingham: Packt Publishing, 2016.

HONNIBAL, Matthew. Introducing spaCy. **Explosion.ai:** Feb. 18, 2015, update: Oct. 3, 2016. [*online*]. Disponível em: https://explosion.ai/blog/introducing-spacy. Acesso em: 21 set. 2021.

HUTTO, C.; GILBERT, Eric. VADER: a parsimonious rule-based model for sentiment analysis of social media text. *In*: INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 8th, 2014, Ann Arbor, Michigan. **Proceedings of the International AAAI Conference on Web and Social Media**, [*s. l.*] v. 8, n. 1, p. 216-225, 2014. DOI: https://doi.org/10.1609/icwsm.v8i1.14550. Alto: 2014. Disponível em: https://ojs.aaai.org/index.php/ICWSM/article/view/14550. Acesso em: 17 set. 2023.

LIU, Jin-Xian; LEU, Jenq-Shiou; HOLST, Stefan. Stock price movement prediction based on Stocktwits investor sentiment using FinBERT and

ensemble SVM. **PeerJ Computer Science**, [*s. l.*], v. 9, 2023. DOI: https://doi.org/10.7717/peerj-cs.1403. Disponível em: https://peerj.com/articles/cs-1403.pdf. Acesso em: 17 set. 2023.

LIU, Yu *et al.* Analysis of the causes of inferiority feelings based on social media data with Word2Vec. **Scientific Reports**, [*s. l.*], v. 12, n. 5218, 2022. DOI: https://doi.org/10.1038/s41598-022-09075-2. Disponível em: https://www.nature.com/articles/s41598-022-09075-2. Acesso em: 17 set. 2023.

MALLEPALLE, Sarah *et al.* Extracting NFL tracking data from images to evaluate quarterbacks and pass defenses. **Journal of Quantitative Analysis in Sports**, [*s. l.*], v. 16, n. 2, p. 95-120, 2020. DOI: https://doi.org/10.1515/jqas-2019-0052. Disponível em: https://www.degruyter.com/document/doi/10.1515/jqas-2019-0052/html. Acesso em: 18 set. 2023.

MELO, Tiago de; FIGUEIREDO, Carlos M. S. Comparing news articles and tweets about COVID-19 in Brazil: sentiment analysis and topic modeling approach. **JMIR Public Health and Surveillance**, [*s. l.*], v. 7, n. 2, e24585, 2021. DOI: https://doi.org/10.2196/24585. Disponível em: https://publichealth.jmir.org/2021/2/e24585/. Acesso em: 21 set. 2023.

PEREIRA, Denilson Alves. A survey of sentiment analysis in the Portuguese language. **Artificial Intelligence Review**, [*s. l.*], v. 54, n. 2, p. 1087-1115, Feb. 2021. DOI: https://doi.org/10.1007/s10462-020-09870-1. Disponível em: https://link.springer.com/article/10.1007/s10462-020-09870-1. Acesso em: 21 set. 2023.

PRAVEEN, S. V. *et al.* Twitter-based sentiment analysis and topic modeling of social media posts using natural language processing, to understand people's perspectives regarding COVID-19 booster vaccine shots in India: crucial to expanding vaccination coverage. **Vaccines**, [*s. l.*], v. 10, n. 11, 2022. DOI: https://doi.org/10.3390/vaccines10111929. Disponível em: https://www.mdpi.com/2076-393X/10/11/1929. Acesso em: 21 set. 2023.

PYTHON BRASIL. **Python Brasil 2023**. [*S. l.*: *s. n.*], 2023. Disponível em: https://2023.pythonbrasil.org.br/#inicio. Acesso em: 28 set. 2023.

ROJAS, Claudio *et al.* Using Geopandas for locating virtual stations in a free-floating bike sharing system. **Heliyon**, [*s. l.*], v. 9, n. 1, 2023. DOI:

https://doi.org/10.1016/j.heliyon.2022.e12749. Disponível em: https://www.cell.com/heliyon/fulltext/S2405-8440(22)04037-3?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2405844022040373%3Fshowall%3Dtrue. Acesso em: 18 set. 2023.

SEABOLD, Skipper; PERKTOLD, Josef. Statsmodels: econometric and statistical modeling with python. *In*: WALT, Stéfan van der; MILLMAN, Jarrod (ed). PYTHON IN SCIENCE CONFERENCE (SCIPY), 9th, Austin, 2010. **Proceedings [...]** Austin, Texas: SciPy, 2010. p. 92-96. DOI: https://doi.org/10.25080/Majora-92bf1922-011. Disponível em: https://conference.scipy.org/proceedings/scipy2010/seabold.html. Acesso em: 21 set. 2023.

WANG, Yanfeng *et al.* Military chain: construction of domain knowledge graph of kill chain based on natural language model. **Hindawi, Mobile Information Systems**, [*s. l.*], v. 22, article ID 7097385, 2022. Disponível em: https://www.hindawi.com/journals/misy/2022/7097385/. Acesso em: 21 set. 2023.

ZIA, Amjad *et al*. Artificial intelligence-based medical data mining. **Journal of Personalized Medicine**, [*s. l.*], v. 12, n. 9, 1359, 2022. DOI: https://doi.org/10.3390/jpm12091359. Disponível em: https://www.mdpi.com/2075-4426/12/9/1359. Acesso em: 17 set. 2023.

ZIHAN, Song; YANHONG, Yang; HONGTAI, Guo. Analysis of data crawling and visualization methods for recruitment industry information. **Journal of Physics:** Conference Series, [*s. l.*], v. 1971, n. 1, 2021. DOI: https://doi.org/10.1088/1742-6596/1971/1/012092. Disponível em: https://iopscience.iop.org/article/10.1088/1742-6596/1971/1/012092/pdf. Acesso em: 18 set. 2023.

DADOS DOS AUTORES:

Denise Fukumi Tsunoda

Denise Fukumi Tsunoda é Professora titular na UFPR, Departamento de Ciência e Gestão da Informação com atuação no curso de graduação em Gestão da Informação, Programa de Pós-Graduação em Gestão da Informação e Mestrado Profissional em Economia. Possui graduação em Bacharelado em Informática pela Universidade Federal do Paraná, mestrado e doutorado em Engenharia Elétrica e Informática Industrial pela UTFPR com estágio pós-doutoral em Ciência da Informação pela UFSC. Atua principalmente nos seguintes temas: inteligência artificial, machine learning, deep learning, mineração de dados, mineração de processos, mineração de textos, computação evolucionária, algoritmos genéticos e análise de dados.

https://orcid.org/0000-0002-5663-4534
dtsunoda@ufpr.br

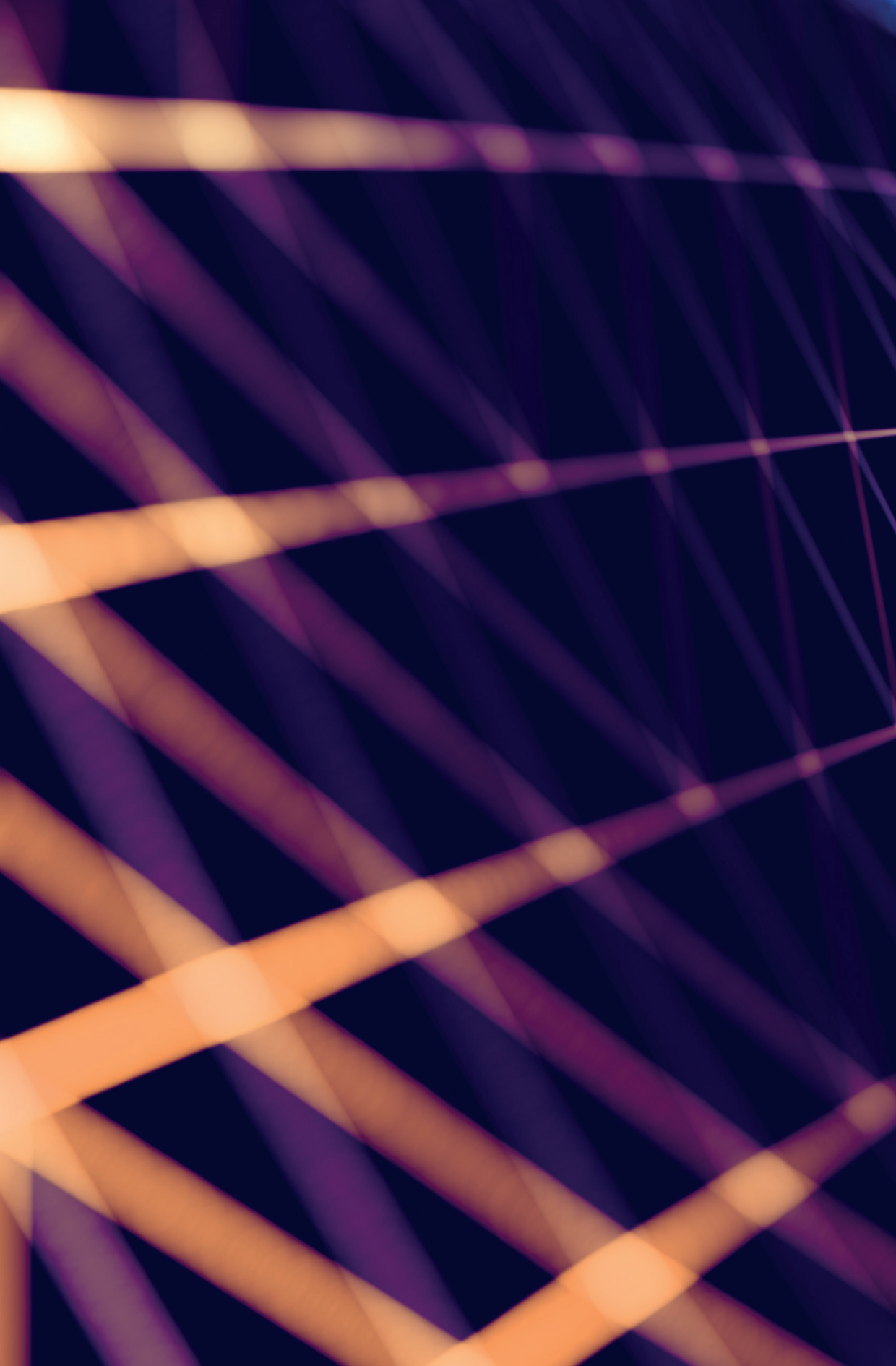
Alex Sebastião Constâncio

Alex Sebastião Constâncio é Analista de tecnologia da informação na UFPR, atua na área de engenharia de software e gestão e administração de banco de dados Oracle. Tem experiência na área de desenvolvimento de software em várias linguagens de programação (C, C++, Delphi, Java, C#, Python e Javascprit) e sistemas de banco de dados (Oracle, Sybase, Microsoft SQL Server e Postgres SQL), tendo atuado nas áreas de software básico, desenvolvimento de frameworks, aplicações web, computação gráfica, compiladores e outras. Graduado em Bacharelado em Informática pela UFPR e mestrado em Ciência, Gestão e Tecnologia da Informação, também pela UFPR. Atualmente é aluno de doutorado no PPG em Gestão da Informação na UFPR e pesquisa a detecção automática de mentiras por meio de métodos de inteligência artificial.

https://orcid.org/0000-0001-8725-4481
alex.constancio@ufpr.br

Como referenciar o capítulo 3:

TSUNODA, Denise Fukumi; CONSTÂNCIO, Alex Sebastião. Python como suporte às pesquisas sociais. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 3. p. 61-89. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap3.



4. LINGUAGEM DE PROGRAMAÇÃO R APLICADA ÀS CIÊNCIAS SOCIAIS APLICADAS

Luciano Heitor Gallegos Marin

4.1 INTRODUÇÃO

As Ciências Sociais Computacionais, uma disciplina interdisciplinar que combina princípios das ciências sociais com técnicas computacionais, tais como ciência de dados e de inteligência artificial, têm ganhado destaque na pesquisa acadêmica e aplicada. Um dos pilares dessa abordagem é a análise de dados massivos gerados em plataformas digitais, como em redes sociais, fóruns on-line e sites de notícias. Essa análise de dados é frequentemente realizada com o uso de algoritmos de aprendizado de máquina e técnicas de mineração de texto para entender padrões de comportamento, opiniões públicas e dinâmicas sociais em escala global (Lazer *et al.*, 2009).

A análise de redes sociais, na qual investiga-se como as conexões e interações entre indivíduos influenciam o comportamento coletivo e a disseminação de informações, também faz parte das Ciências Sociais Computacionais. O uso de métodos computacionais nesse contexto permite uma análise detalhada de redes complexas e sua dinâmica (Wasserman; Faust, 1994) e, além disso, desempenha um papel crucial na compreensão e enfrentamento da desinformação e da propagação de notícias falsas nas redes sociais. Pesquisadores usam esses métodos para rastrear a disseminação de informações enganosas, identificar redes de desinformação e desenvolver estratégias para mitigar seu impacto (Friggeri *et al.*, 2014). Normalmente, os métodos computacionais são utilizados e aplicados por meio de softwares, aplicativos especializados e, também, por linguagens de programação.

A linguagem de programação *R*, criada nos anos 70 como "linguagem de programação S", foi inicialmente concebida para auxiliar no processamento estatístico e na análise de dados. O *R* é de código aberto e, atualmente, se tornou uma escolha popular entre cientistas de dados, estatísticos e

pesquisadores de diversas áreas devido à sua flexibilidade e extensibilidade (Ihaka; Gentleman, 1996), oferecendo uma vasta gama de pacotes e bibliotecas que facilitam a manipulação, visualização e modelagem de dados, tornando-o uma escolha essencial para análise estatística. Uma das principais características que torna o *R* atraente é sua capacidade de produzir gráficos e visualizações de alta qualidade, pois oferece várias opções para criar gráficos personalizados, o que é essencial para a apresentação eficaz de resultados de análise de dados (Wickham, 2016). Além disso, a comunidade de usuários do *R* é ativa e contribui constantemente com novos pacotes e soluções, mantendo-o atualizado e relevante para as demandas em permanente evolução da análise de dados.

Neste capítulo, a linguagem de programação *R* será, primeiramente, pormenorizada como ferramenta de apoio para o processamento e análise de dados na seção "Sobre a Tecnologia". Na sequência, essa mesma linguagem de programação será descrita como apoio dentro das pesquisas em Ciências Sociais Computacionais na seção "A Tecnologia e a Pesquisa". Exemplos de pesquisas em Ciências Sociais Computacionais, utilizando a linguagem de programação em *R*, serão descritas na seção "Exemplos de Pesquisas que Utilizaram a Tecnologia". Finalmente, na seção "Considerações Finais" serão explorados os principais aspectos do presente capítulo.

4.2 SOBRE A LINGUAGEM DE PROGRAMAÇÃO R

O *R* é uma linguagem de programação de código aberto amplamente utilizada para análise de dados e estatísticas. Nesta seção, tem-se como objetivo fornecer uma introdução abrangente à essa linguagem, desde os conceitos básicos até tópicos avançados.

A linguagem de *Programação R* pode ser facilmente instalada por meio de sua página oficial, conhecida por "*The R Project for Statistical Computing*[42]". Normalmente, na página oficial da linguagem de programação *R*, estarão disponibilizadas opções para instalação do *R* nos mais diversos tipos de sistemas operacionais, tais como *Microsoft Windows*, *MacOS* e

42 Página oficial do "*The R Project for Statisitical Computing*": https://www.r-project.org/.

distribuições para plataformas baseadas em UNIX. O *R* tem suas atualizações e instalações de bibliotecas baseadas em repositório conhecidos por "*Comprehensive R Archive Network - CRAN*" e, no Brasil, temos 2 universidades que apoiam essa iniciativa: Universidade de São Paulo[43] e Universidade Federal do Paraná[44]. Uma vez instalada a linguagem de programação *R*, e escolhida a sua *CRAN*, sua utilização dá-se por meio de comandos e funções no console do sistema operacional utilizado pelo usuário, o que pode limitar o seu uso. Opcionalmente, pode-se instalar outros ambientes para utilizar o *R* de forma facilitada, como é o caso do *RStudio*.

O *RStudio*[45] é um ambiente de desenvolvimento integrado (*Integrated Development Environment - IDE*) amplamente utilizado por programadores e cientistas de dados que trabalham com a linguagem de programação *R*. Esse ambiente fornece uma interface amigável e altamente funcional para o *R*, tornando-o mais acessível, produtivo e eficiente. O *RStudio* inclui um console *R* interativo, uma área de script para escrever código, além de recursos de depuração, visualização de gráficos e gerenciamento de projetos, em uma única interface. O *RStudio* publica e oferece diversas bibliotecas para facilitar o processamento, análise e divulgação de resultados.

A biblioteca "*dplyr*" é uma poderosa ferramenta para manipulação e transformação de dados em *R*. Desenvolvida por Hadley Wickham, o *dplyr* oferece um conjunto de funções simples e consistentes que facilitam a limpeza, filtragem, agrupamento e resumo de dados de forma eficiente. Uma das características do *dplyr* é sua sintaxe clara e concisa, que torna o código mais legível e fácil de manter. As principais funções do *dplyr* incluem "*filter()*" para filtrar linhas de dados com base em critérios específicos, "*select()*" para escolher colunas relevantes, "*mutate()*" para criar novas variáveis e "*group_by()*" para agrupar dados por uma ou mais variáveis (Wickham *et al.*, 2021).

O "*ggplot2*" é uma biblioteca para criação de gráficos e visualização de dados em *R*. Desenvolvida por Hadley Wickham, a *ggplot2* é baseada

43 Repositório *CRAN* da Universidade de São Paulo: https://vps.fmvz.usp.br/CRAN/.

44 Repositório *CRAN* da Universidade Federal do Paraná: https://cran-r.c3sl.ufpr.br/.

45 Página do *RStudio*: https://posit.co/products/open-source/rstudio/.

no conceito de "gramática dos gráficos", o que significa que permite aos usuários criarem gráficos de alta qualidade de maneira intuitiva e flexível. Os gráficos gerados pelo *ggplot2* são altamente customizáveis, o que permite aos usuários controlarem praticamente todos os aspectos do visual, desde cores até títulos, resultando em visualizações de dados claras e informativas. A principal vantagem da *ggplot2* é sua sintaxe coerente e intuitiva, que simplifica a criação de gráficos complexos. Os usuários podem começar com uma função *ggplot()* e, em seguida, adicionar camadas estéticas e geométricas para construir gráficos personalizados. Por exemplo, para criar um gráfico de dispersão, pode-se usar "*ggplot(data = dados, aes(x = variavel1, y = variavel2)) + geom_point()*". A flexibilidade e extensibilidade do *ggplot2* também permitem a criação de gráficos facetados (Wickham, 2016).

A *Linguagem de Programação R*, e o *RStudio*, também suportam bibliotecas envolvendo aprendizado de máquina, processamento de linguagem natural e aprendizado profundo. A biblioteca "*TensorFlow*" é amplamente reconhecida por suas capacidades avançadas de aprendizado de máquina e redes neurais e, agora, também está disponível para a linguagem de programação *R* por meio do pacote '*tensorflow*'. Esse pacote permite que os usuários do *R* possam realizar tarefas de aprendizado profundo e criação de modelos de redes neurais. Ele oferece suporte à construção, treinamento e implantação de modelos de aprendizado de máquina complexos, incluindo redes neurais convolucionais e redes recorrentes. Além disso, os usuários podem aproveitar a capacidade de processamento paralelo e a escalabilidade do *TensorFlow* enquanto continuam a utilizar o ambiente familiar do *R* para análise de dados e visualização (RStudio, 2021).

O "*tidyverse*", do *RStudio*, é uma coleção abrangente de pacotes e bibliotecas para a linguagem de programação *R* comumente utilizada como o "*ggplot2*", o "*dplyr*", dentre outras, embora dependa de mais processamento por condensar tais bibliotecas em uma única. Uma das características do *Tidyverse* é a ênfase na "*tidy data*", um formato organizado e padronizado de dados que facilita a análise. Isso é alcançado usando convenções consistentes para nomes de funções e argumentos, o que torna o código mais legível e fácil de entender. Além disso, o *Tidyverse* promove a utilização de *pipelines* para encadear operações de transformação de dados de forma clara e eficiente (Wickham, 2017).

A *Integrated Development Environment - IDE RStudio* possui a capacidade de suportar a criação de relatórios dinâmicos e apresentações por meio de ferramentas como o *R Markdown*, que permite combinar texto, código *R* e gráficos em um único documento. O *R Markdown* é bastante utilizado na comunicação de resultados de análises de dados e relatórios técnicos (Allaire *et al.*, 2021). Outro recurso bastante utilizado para a publicação de resultados em painéis estáticos e dinâmicos ocorre por meio da ferramenta *R Shiny*.

A ferramenta *R Shiny*, do *RStudio*, possibilita a criação de aplicativos *web* interativos com interface gráfica de usuário (*Graphic User Interface - GUI*) sem a necessidade de conhecimento avançado em desenvolvimento *web*. Ele é útil para cientistas de dados e analistas que desejam compartilhar análises de dados de forma interativa com outras pessoas, por meio de *URL*. A principal característica do *Shiny* é a capacidade de transformar códigos escritos em *R* em aplicativos *web* funcionais, tais como: botões, caixas de seleção e gráficos, que respondem às ações do usuário. A biblioteca lida com a comunicação entre o navegador do usuário e o servidor *R*, permitindo que os aplicativos *web* gerem saídas dinâmicas e interajam com os dados em tempo real (Chang *et al.*, 2021).

4.3 A LINGUAGEM DE PROGRAMAÇÃO R E A PESQUISA EM CIÊNCIAS SOCIAIS APLICADAS

A linguagem de programação *R* tem sido amplamente utilizada nas ciências sociais aplicadas devido à sua versatilidade e capacidade de lidar com análises de dados complexas. Aqui estão algumas maneiras pelas quais o *R* é aplicado nessas áreas:

- Análise de Dados Estatísticos: o *R* oferece uma ampla gama de pacotes e funções estatísticas que são essenciais para a análise de dados nas ciências sociais. Os pesquisadores podem realizar testes de hipóteses, análises de variância, regressões e outras análises estatísticas avançadas para examinar dados em áreas como psicologia, sociologia e economia.

- Visualização de Dados: a biblioteca *ggplot2*, mencionada anteriormente, é altamente valorizada nas ciências sociais aplicadas por sua capacidade

de criar gráficos estatísticos de alta qualidade. A visualização de dados desempenha um papel crucial na apresentação de resultados de pesquisa e na comunicação eficaz de descobertas.

- Análise de Texto e Mineração de Dados Sociais: nas ciências sociais, a análise de texto e a mineração de dados sociais são cada vez mais relevantes. O *R* oferece pacotes como *tm* (*text mining*) e quanteda para lidar com dados de texto, tornando possível a análise de documentos, redes sociais e outras fontes de dados não estruturados.

- Modelagem e Previsão: a modelagem estatística e a previsão são aspectos fundamentais das ciências sociais aplicadas. O *R* oferece uma variedade de técnicas de modelagem, incluindo modelos de regressão, séries temporais e modelos de séries temporais espaciais, que são aplicados em estudos de economia, demografia e outras áreas.

- Pesquisa Reprodutível: a capacidade de criar documentos dinâmicos usando o *R Markdown* permite que os pesquisadores documentem e compartilhem suas análises de maneira clara e reprodutível. Isso é essencial para a transparência e validade da pesquisa nas ciências sociais.

- Bibliotecas e Recursos Específicos: existem pacotes específicos do *R* desenvolvidos para áreas particulares das ciências sociais, como psicometria, análise de redes sociais, demografia, entre outros, tornando o *R* uma escolha flexível para uma ampla gama de aplicações.

O *R* desempenha um papel fundamental nas ciências sociais aplicadas, fornecendo ferramentas poderosas para coleta, análise e interpretação de dados. Sua capacidade de lidar com uma variedade de tarefas e sua comunidade ativa de usuários e desenvolvedores o tornam uma escolha valiosa para pesquisadores e profissionais que buscam *insights* nas ciências sociais.

4.4 EXEMPLOS DE PESQUISAS QUE UTILIZAM A LINGUAGEM DE PROGRAMAÇÃO R EM CIÊNCIAS SOCIAIS APLICADAS

Nesta seção, apresentam-se pesquisas e seus respectivos resultados sobre dados de redes e mídias sociais, que vêm sendo largamente utilizados para

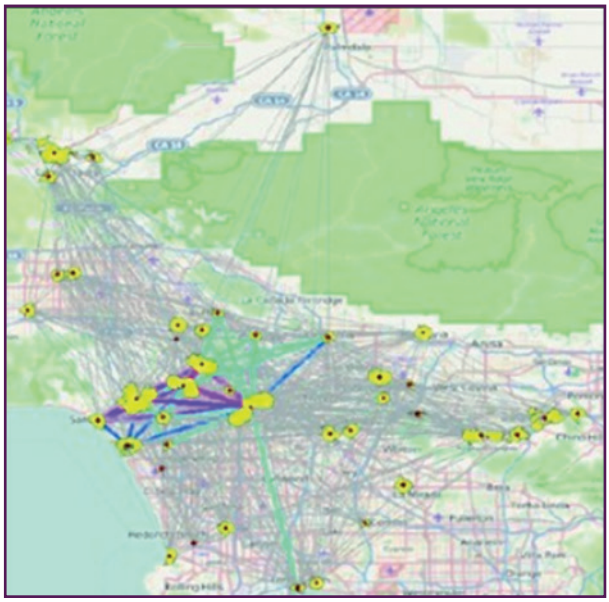
estudos, pesquisas, construção de modelos e análises para entender padrões de comportamento, opiniões públicas e dinâmicas sociais em escala global. Dessa forma são apresentados, sem a pretensão de serem descritas de forma exaustiva, trabalhos de pesquisadores brasileiros e internacionais envolvendo as Ciências Sociais Computacionais e o uso da linguagem de programação *R* para a coleta, processamento e análise desses trabalhos.

Os serviços de partilha de localização fornecidos pelas redes sociais, atualmente, proporcionam um acesso sem precedentes a dados geolocalizados para estudar a interação entre esses fatores em uma escala muito maior. Torna-se, então, possível utilizar dados de serviço de partilha de geolocalização pessoal (exemplo: *Foursquare*)[46] atrelados a plataformas de redes sociais de comunicação, como o *Twitter* (atualmente, renomeado com o nome de *X*)[47] para analisar propriedades dos indivíduos que estão em uma região, tais como: felicidade, estresse, depressão e ansiedade. Normalmente, áreas com mensagens mais positivas incentivam as pessoas que vivem nelas a compartilhar mensagens positivas, ou de outras áreas a se deslocarem até esses locais, enquanto o contrário, afastam as pessoas. Essas informações lançam luzes sobre a influência que certos lugares desempenham em relação às emoções e à mobilidade das pessoas, o que, por sua vez, pode ser usado pelos planejadores urbanos para conceberem cidades mais felizes e mais equitativas (Gallegos *et al.*, 2016). Nesse sentido, a escolha pelos destinos de consumo e compras também é influenciada por mensagens mais positivas (Huang; Gallegos; Lerman, 2017), incentivando a abertura de nichos de negócios, como apresentado na Figura 1. Além disso, podem ser utilizadas para a aproximação de tendências em mensagens massivas e geolocalizadas que emitam emoções sobre epidemias e pandemias (Maia; Oliveira; Gallegos, 2021), auxiliando na compreensão de eventos impactantes em regiões escolhidas (Figura 2). Os resultados apresentados nas Figuras 1 e 2 utilizam as bibliotecas *ggplot2* e *leaflet*, da linguagem de programação *R*, para a apresentação dos resultados:

46 *Site* do *Foursquare*: https://foursquare.com/. Acesso em: 18 set. 2023.

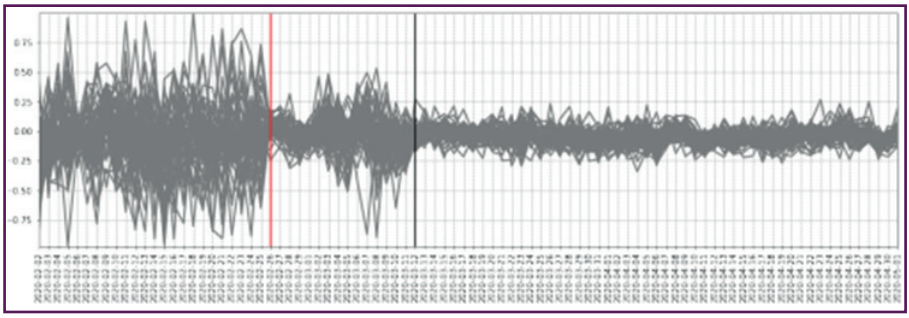
47 *Site* de *X*: https://twitter.com/X?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor. Acesso em: 18 set. 2023.

Figura 1 - Agrupamento de áreas na cidade de Los Angeles por mensagens geolocalizadas mais positivas. Tais áreas possuem, normalmente, comércio e lazer.



Fonte: Huang, Gallegos e Lerman (2017).

Figura 2 - Médias diárias de 86 cidades brasileiras, após análise georeferenciada de mensagens, com pontuações de análise de sentimento neutras (0), positivas (máximo +1) e negativas (mínimo -1), demarcados pelo primeiro caso de Covid-19 (barra vermelha) e primeira morte (barra preta) no Brasil, segundo informações do Ministério da Saúde do Brasil.



Fonte: Maia, Oliveira e Gallegos (2021).

As ciências sociais aplicadas debruçam-se, também, sobre aspectos que envolvem polarização e discurso de ódio, que são "ideias que incitem a discriminação racial, social ou religiosa em determinados grupos, na maioria das vezes, às minorias". De fato, essas manifestações, nos mais diversos cenários, são cada vez mais frequentes nas plataformas sociais e, para a detecção de tais discursos, são utilizados anotadores (exemplo: pessoas que anotam palavras e seus sentidos, manualmente) que podem criar rótulos para diferentes tarefas de classificação, com definições divergentes de discurso de ódio, esquemas binários ou multi-rótulos e diversas metodologias para recolha de dados. Pode-se, então, examinar os principais conjuntos de dados disponíveis publicamente para investigação desses discursos por tipo (por exemplo, etnia, religião, orientação sexual) presentes na sua composição, revelando detalhes para a compreensão do fenômeno desse tipo de discurso e para melhorar a sua deteção em plataformas sociais (Guimarães *et al.*, 2023). A coleta de dados, para esse trabalho voltado ao estudo do discurso do ódio em redes sociais, foi realizada com a biblioteca *selenium* da linguagem de programação *R*.

4.5 CONSIDERAÇÕES FINAIS

As Ciências Sociais Computacionais são uma disciplina interdisciplinar que combina princípios das ciências sociais com técnicas computacionais, tais como ciência de dados e de inteligência artificial. Visando auxiliar na exploração e no trabalho com iniciativas em ciências sociais computacionais, neste capítulo, foram exploradas as principais características e funcionalidades da linguagem de programação *R* como ferramenta de apoio para o processamento e a análise de dados, bem como de demonstração de resultados de diferentes formas. Dessa forma, o *R* foi apresentado como ferramenta de apoio nas ciências sociais aplicadas devido à sua versatilidade e capacidade de lidar com análises de dados complexas.

A linguagem de programação *R*, devido a essas características, foi descrita como uma ferramenta capaz de auxiliar na análise de dados estatísticos, na visualização de dados, na análise de textos e mineração de dados sociais, na modelagem e na previsão de dados, na pesquisa reprodutível e, por meio de novas bibliotecas criadas e disponibilizadas por pesquisadores, contribuir com profissionais e interessados em ciências sociais computacionais.

REFERÊNCIAS

ALLAIRE, J. J.; XIE, Y.; MCPHERSON, J.; LURASCHI, J.; USHEY, K.; ATKINS, A.; IANNONE, R. **R Markdown:** the definitive guide. Flórida: Chapman and Hall/CRC, 2021.

CHANG, W.; CHENG, J.; ALLAIRE, J. J.; SIEVERT, C.; SCHLOERKE, B.; XIE, Y.; ALLEN, J.; MCPHERSON, J; DIPERT, A.; BORGES, B. **shiny:** Web Application Framework for R. R package version 1.7.1. 2021. Disponível em: https://CRAN.R-project.org/package=shiny. Acesso em: 28 set. 2023.

FRIGGERI, A.; ADAMIC, L. A.; ECKLES, D.; KLEINBERG, J. Rumor cascades. *In*: INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 8th, 2014, Michigan. **Proceedings** [...]. Michigan: University of Michigan, 2014. p. 101-110.

GALLEGOS, L.; LERMAN, K.; HUANG, A.; GARCIA, D. Geography of emotion: Where in a city are people happier? *In*: INTERNATIONAL CONFERENCE COMPANION ON WORLD WIDE WEB, 25th, 2016, Québec. **Proceedings** [...]. Québec: IW3C2, 2016.

GUIMARÃES, S.; KAKIZAKI, G.; MELO, P.; SILVA, M.; MURAI, F.; REIS, J. C.; BENEVENUTO, F. Anatomy of Hate Speech Datasets: composition analysis and cross-dataset classification. *In*: ACM CONFERENCE ON HYPERTEXT AND SOCIAL MEDIA, 34th, 2023, Roma. **Proceedings** [...]. Roma: SIGWEB, 2023.

HUANG, A.; GALLEGOS, L.; KRISTINA, L. Travel analytics: understanding how destination choice and business clusters are connected based on social media data. **Transportation Research Part C**: emerging Technologies, Oxford, v. 77, p. 245-256, 2017.

IHAKA, R.; GENTLEMAN, R. R: A language for data analysis and graphics. **Journal of Computational and Graphical Statistics**, Alexandria, VA, v. 5, n. 3, p. 299-314, 1996.

LAZER, D.; PENTLAND, A. S.; ADAMIC, L.; ARAL, S.; BARABASI, A. L.; BREWER, D.; JEBARA, T. *et al.* Computational social science. **Science**, Washington, v. 323, n. 5915, p. 721-723, 2009.

MAIA, M.; OLIVEIRA, M.; GALLEGOS, L. Covid-19 e tweets no brasil: coleta, tratamento e análise de textos com evidências de estados afetivos alterados em momentos impactantes. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 10th, 2021, João Pessoa. **Proceedings** [...]. João Pessoa: SBC, 2021.

RSTUDIO. **TensorFlow for R**. 2021. Disponível em: https://tensorflow.rstudio.com/. Acesso em: 28 set. 2023.

WASSERMAN, S.; FAUST, K. **Social network analysis**: methods and applications. Cambridge: Cambridge University Press, 1994. v. 8.

WICKHAM, H. **ggplot2:** elegant graphics for data analysis. New York: Springer, 2016.

WICKHAM, H. **Tidyverse:** easily install and load 'tidyverse' packages. R package version 1.2.1. 2017. Disponível em: https://CRAN.R-project.org/package=tidyverse. Acesso em: 28 set. 2023.

WICKHAM, H.; FRANÇOIS, R.; HENRY, L.; MÜLLER, K.; VAUGHAN, D. **dplyr**: a grammar of data manipulation. R package version 1.0.7. 2021. Disponível em: https://CRAN.R-project.org/package=dplyr. Acesso em: 28 set. 2023.

DADOS DO AUTOR:

Luciano Heitor Gallegos Marin

Luciano Heitor Gallegos Marin é professor da Universidade Federal do Paraná - UFPR, alocado no Departamento de Ciência e Gestão da Informação. Possui bacharelado em Análise de Sistemas, mestrado em Engenharia da Computação pelo Instituto Tecnológico de Aeronáutica, doutorado em Engenharia Elétrica pela Université de Rennes I com período sanduíche pela Northeastern University, e pós-doutorado em Ciências Sociais Computacionais pela University of Southern California. Atua como coordenador do bacharelado em Gestão da Informação, e do eixo de Informação e Tecnologia da Pós-graduação em Gestão da Informação da UFPR. Como pesquisador, vem atuando em trabalhos, projetos e publicações envolvendo Ciências Sociais Computacionais, Ciência de Dados Comportamentais, Sistemas Colaborativos, Agentes Conversacionais, e Ontologias e Web Semântica.

https://orcid.org/0000-0002-4331-6588
luciano.gallegos@ufpr.br

Como referenciar o capítulo 4:

MARIN, Luciano Heitor Gallegos. A linguagem de programação R aplicada às Ciências Sociais Aplicadas. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 4. p. 91-102. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap4.

5. EXTRAÇÃO E ANÁLISE DE DADOS REGISTRADOS EM TEXTO LIVRE DE PRONTUÁRIO ELETRÔNICO DO PACIENTE POR MEIO DE PROCESSAMENTO DE LINGUAGEM NATURAL

Amanda Damasceno de Souza
Eduardo Ribeiro Felipe
Fernanda Farinelli

5.1 INTRODUÇÃO

Na área de saúde, a informação de qualidade para tomada de decisão é essencial por prestar melhor assistência ao paciente. Com isso, as Tecnologias de Informação (TI) impactaram consideravelmente a prática clínica ao se integrarem como peças fundamentais às rotinas dos profissionais da área de saúde (Shortliffe, 2014). A informação de assistência em saúde é registrada no Prontuário Eletrônico de Paciente (PEP), conhecido por vários outros termos: prontuário médico, prontuário nosológico do paciente, prontuário médico do paciente etc., e por termos que dizem respeito à sua documentação: laudo médico, relatório médico, exame médico e registro de saúde (Blobel, 2018; Conselho Regional de Medicina do Distrito Federal, 2006).

Nesse contexto, as terminologias clínicas padronizadas são importantes por realizarem a interface dos dados clínicos com os sistemas de atenção à saúde, entre eles o PEP (Rogers, 2005). Além disso, as terminologias padronizadas são recursos valorosos para possibilitar a interoperabilidade no PEP, ao colaborar na realização de auditoria, pesquisa, *benchmarking* e gerenciamento de resultados para o hospital (Miñarro-Giménez *et al.*, 2019). Schulz *et al.* (2017) citam três tipos de terminologias em saúde e propõem uma metodologia para realizar conexão entre elas: Terminologias de Interface (texto clínico do prontuário ou jargão médico), Terminologias de Referência (vocabulários controlados e/ou ontologias) e Terminologias

de Agregação (CID, SNOMED-CT)[48, 49]. Uma lacuna indicada por Schulz e*t al.* (2017) é encontrar uma forma de promover a conexão entre os dados clínicos presentes nos textos clínicos do PEP e as terminologias clínicas padronizadas. As terminologias clínicas padronizadas podem ser usadas para identificar o significado semântico dos conceitos lexicais em um texto clínico dos prontuários. Além disso, terminologias são importantes ao expandir o conceito, com um ou mais sinônimos, na identificação de abreviações e acrônimos em texto clínico e também para mapear conceitos para terminologias de outros sistemas (Dalianis, 2018a, p. 35). Para possibilitar a interoperabilidade entre as terminologias clínicas é preciso analisar os termos e dados clínicos do PEP e, com isso, buscar uma conexão destes termos para outros artefatos terminológicos[50] da área da saúde. Para solucionar este problema semântico, a Ciência da Informação (CI) e a Ciência da Computação (CC)[51] precisam caminhar juntas.

A CI tem um papel importante na área de saúde, por sua contribuição nas áreas de recuperação, representação e organização da informação. A CI lida com a representação, a armazenagem e a recuperação da informação, com o tratamento da informação na tarefa de "[...] construir interfaces entre os acervos de documentos e informações e seus usuários" (Alvarenga, 2003, p. 25). O bibliotecário que atua na saúde precisa desenvolver habilidades para manipular diferentes suportes de busca e recuperação da informação e, com isso, contribuir para a melhoria do tratamento da informação e o cuidado com a saúde. Cada vez mais a CI necessita caminhar junto com a TI na Recuperação da Informação (RI) de registros médicos presentes no PEP, a fim de possibilitar ao usuário o acesso à informação (Souza, 2021; Campos, 2001).

Os registros médicos clínicos são ricos em informação, e em grande parte são concebidos em formato de texto livre (dados clínicos). Os meios para extrair informação estruturada desses registros em texto livre é um significativo esforço de pesquisa (Zhou *et al.*, 2006). Nesse contexto, o volume

48 CID: Classificação Internacional de Doenças.

49 *SNOMED-CT: Systematized Nomenclature of Medical - Clinical Terms*.

50 Conjunto estruturado de termos por meio de uma metodologia no estabelecimento de relações hierárquicas ou semânticas.

51 Também representada pela TI neste contexto.

de informação produzido na pesquisa acadêmica e na prática clínica há muito exige tratamento computacional. Uma importante fonte de dados de pacientes, relevante para a pesquisa, além de essencial para a gestão das unidades de saúde, é o PEP. Dessa forma, técnicas de processamento de linguagem natural (PLN) são alternativas importantes para lidar com os campos de texto livre dessa fonte dinâmica, onde constantemente se registram novos dados. O PLN envolve processamento inteligente de texto, no qual o computador busca interpretar o que foi escrito em linguagem natural, valendo-se de métodos computacionais linguísticos. Essa abordagem de PLN visa à extração de informação específica de documentos ou coleções de documentos, a fim de serem aplicadas em campos de texto livre dos PEPs (Dalianis, 2018b).

A Mineração de Texto (*Text Mining*) consiste na aplicação de técnicas de garimpagem de dados para obtenção de informações importantes. É um processo que utiliza algoritmos capazes de analisar coleções de documentos escritos em linguagem natural, com o objetivo de extrair conhecimento e identificar padrões. Dentre as técnicas utilizadas, destaca-se o PLN (Dalianis, 2018b).

O processamento de textos clínicos e biomédicos, no âmbito da informática médica, envolve a utilização de métodos baseados em PLN. O *Text Mining* (TM) objetiva encontrar, previamente, fatos desconhecidos no texto ou em coleções de textos, assim como criar hipóteses para serem provadas. O PLN se refere ao processamento inteligente de texto, no qual o computador busca interpretar o que foi escrito em linguagem natural, utilizando, para isso, métodos computacionais linguísticos. Essas duas abordagens, de TM e PLN, se referem à Extração de Informação (EI), que busca encontrar informação específica em um documento ou em coleções de documentos (Dalianis, 2018b). Na análise de informação lexical (textos) do domínio biomédico, Bodenreider (2006) sugere utilizar soluções de *Machine Learning Technique* (ML). Para Dalianis (2018b), o Machine Learning (ML), ou seja, o aprendizado de máquina, é um conjunto de técnicas que usa dois padrões de comportamento, a saber: os supervisionados e os não supervisionados.

Como exemplos de métodos supervisionados, citamos o *Conditional Random Field* (CRF)/Campo Aleatório Condicional (CAC) e o *Support Vector Machine* (SVM)/ Máquina de Vetor de Suporte (MVS). Como métodos não

supervisionados, citamos o *Latent Semantic Analysis* (LSA)[52], o *Latent Semantic Indexing* (LSI)[53], o *Random Indexing*[54] e text clustering[55].

A recuperação do conhecimento, presente nos textos em linguagem natural, é uma tarefa árdua, que envolve técnicas como o PLN. O PLN diz respeito ao processamento inteligente de texto em linguagem natural, com utilização de métodos computacionais linguísticos (Manning; Schütze, 1999).

Com o desenvolvimento crescente das TI, uma equipe médica produz, hoje, uma quantidade de informação maior do que em qualquer outro momento da história. Grande parte dessa informação está em formato texto e digital. A sobrecarga de informação resultante de tanto material disponível impacta na tomada de decisão, sendo necessário utilizar recursos tecnológicos para recuperar o conteúdo relevante. Diante disso, são necessárias ferramentas para extração, análises e organização dos dados e informações (Blake, 2011). Assim, as Ciências Sociais Aplicadas com áreas como a CI assumem papel importante, ao considerar a informação como força construtiva na sociedade, por sua intensa e crescente aplicação em áreas computacionais em Saúde.

O presente estudo se insere nesse contexto, e descreve tecnologias de informação e comunicação (TICs) utilizadas na extração e análise dos dados, nas etapas metodológicas de pesquisa acadêmica. O objetivo deste capítulo é demonstrar o uso de TICs, a saber, PLN e ferramentas de gestão de projeto, na metodologia de pesquisa *stricto sensu* (doutorado) no âmbito da CI aplicada em textos livres de PEP (intitulado Terminologia de Interface) no campo da saúde, especificamente na área de Ginecologia e Obstetrícia. A seguir, o tópico dois aborda o detalhamento da tecnologia que foi empregada na pesquisa, o tópico três aborda um exemplo prático da metodologia da pesquisa *stricto sensu*, ao abordar etapas de extração e análise de dados da Terminologia de Interface por meio de PLN no contexto da Ginecologia e Obstetrícia; o tópico quatro faz uma breve revisão

52 Em português: Análise Semântica Latente (ASL).

53 Em português: Indexação semântica latente.

54 Em português: indexação aleatória.

55 Em português: agrupamento de texto.

de literatura de pesquisas que utilizaram o PNL em terminologias clínicas. Por fim, o tópico cinco consiste nas considerações finais, seguidas das referências e dos anexos.

5.2 SOBRE A TECNOLOGIA EMPREGADA NA PESQUISA

A análise do problema, bem como do ambiente informacional, é a etapa inicial que fundamenta as escolhas e estratégias do modelo tecnológico. Com o foco na extração de dados em texto livre no contexto da saúde de Ginecologia e Obstetrícia, identificou-se que o modelo de registro das informações ultrapassa a complexidade da linguagem natural enquanto representação do conhecimento. A inserção de códigos médicos, abreviações, siglas e, por muitas vezes, o uso do jargão médico, demonstrou a dificuldade de análise do conjunto de dados. Alinhado a esse cenário, o cuidado com a exposição de dados sensíveis elevou o cuidado com a extração das informações de seu banco de dados original.

5.2.1 PROCESSO DE AQUISIÇÃO DOS DADOS

Optou-se, portanto, por uma estratégia de extração e tratamento dos dados autorizados pelo hospital. A Figura 1 permite visualizar graficamente o processo de Aquisição das informações para o experimento, onde:

1. A instituição de saúde avalia o experimento e, por meio do registro de processo administrativo, aprova o uso de dados reais em sua comissão de ética em pesquisa. Cabe ressaltar que pesquisas realizadas com dados de PEP necessitam de aprovação do Comitê de Ética em Pesquisa (CEP). Na área das Ciências da Saúde, as pesquisas, envolvendo seres humanos, são subordinadas à Resolução CNS nº 466, de 12 dezembro de 2012, e às especificidades da área de Ciências Humanas e Sociais, e contempladas pela Resolução nº 510, de 07 de abril de 2016 (Souza, 2022). Os dados utilizados neste estudo foram coletados de PEP de Hospital Privado, aprovados pelo CEP local e pelo número do CAAE: 03384418.0.0000.51259. A comissão instituiu regras a serem seguidas pelo departamento de TI da instituição.

2. Após a aprovação da comissão, a pesquisadora analisou o banco de dados da instituição a fim de identificar as informações passíveis de análise por meio de critérios alinhados ao objetivo da pesquisa.

3. Alinha-se com a gerência de tecnologia da informação uma estratégia de extração das informações do banco de dados oficial. Realizam-se processos de tratamento para alinhamento com a Lei Geral de Proteção de Dados[56] (LGPD) e a desidentificação de dados sensíveis, incapacitando a base de dados extraída na identificação de pacientes.

4. Este conjunto de dados extraído do banco de dados principal foi gravado em um banco denominado *PostgreSQL*, compactado para transferência e posterior uso no computador pessoal da pesquisadora.

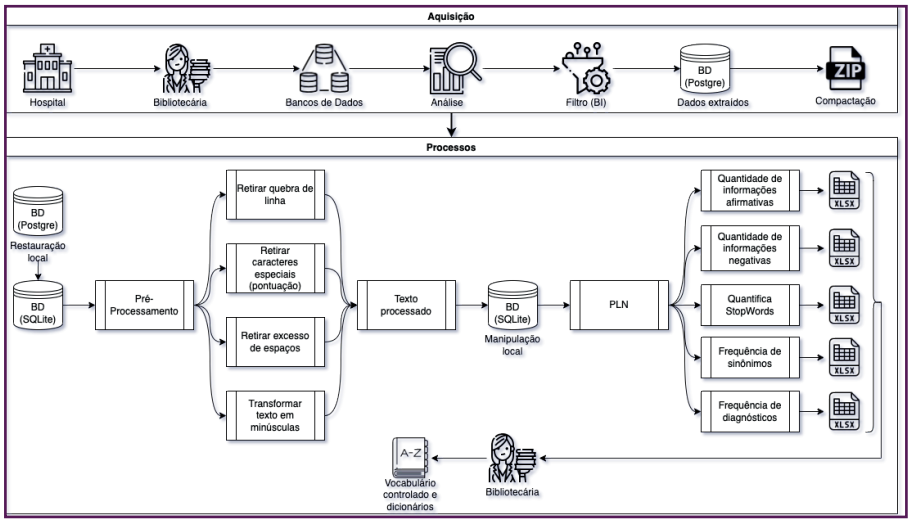
Pode-se ressaltar, nesse processo, a dinâmica tecnológica alinhada aos requisitos da pesquisa. Para a pesquisa no banco de dados principal da instituição foi usada a linguagem SQL[57]. Identificados os dados principais, foram desenvolvidas *queries*[58] no ambiente de *Business Intelligence* (BI) para conexão e seleção das informações. O conjunto de dados resultante (*dataset*) foi exportado para um banco de menor porte, como citado anteriormente no item IV.

56 Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709compilado.htm. Acesso em: 3 out. 2023.

57 Disponível em: https://www.ibm.com/docs/en/i/7.4?topic=concepts-structured-query-language. Acesso em: 3 out. 2023.

58 Formulações em linguagem computacional SQL para consultas ao banco de dados.

Figura 1 - Principais processos informacionais da metodologia



Fonte: Elaborado pelos autores (2023).

5.2.2 PROCESSOS TÉCNICOS

De posse das informações, definiu-se uma abordagem local para a pesquisa. A Figura 1 também demonstra os grupos "Aquisição" e "Processos", responsáveis pela manipulação dos dados. Como processos iniciais, destacam-se, de forma sintetizada: a) a extração da informação do banco de dados do Hospital (aquisição); b) a restauração dos dados em ambiente local; e c) a adequação do formato do banco de dados para manipulação.

5.2.2.1 AQUISIÇÃO

Os dados do Hospital estão armazenados em um sistema de grande porte. Após cuidadosa análise, para preservar as normas de sigilo de dados pessoais dos pacientes e exclusão de dados sensíveis, a realização da extração deu-se a partir de um recorte conceitual (anamnese e evolução). Os registros foram exportados para um banco de dados *PostgreSQL*[59]. Esse artefato (banco de

59 Disponível em: https://www.postgresql.org/. Acesso em: 3 out. 2023.

dados) foi o resultado do processo de identificação dos registros de interesse do projeto (anamneses e evoluções), constituindo o recorte a ser analisado. Esse arquivo foi enviado à pesquisadora em formato compactado, no formato ZIP, como fonte de dados principal para fomento da pesquisa. Tal etapa corresponde ao grupo "aquisição" da Figura 1.

5.2.2.2 RESTAURAÇÃO E ADEQUAÇÃO DO BANCO DE DADOS EM AMBIENTE LOCAL

Por se tratar de um banco de dados dependente de um *software* "servidor", optou-se por migrar os dados para um banco mais simples, denominado *SQLite*. O formato permitiu grande flexibilidade no acesso às informações por não precisar de um "serviço servidor" instalado no sistema operacional do computador pessoal da pesquisadora.

5.2.2.3 PRÉ-PROCESSAMENTO

De posse da estrutura, iniciou-se o tratamento dos dados por meio do desenvolvimento de algoritmos na linguagem de programação *Python*. Destaca-se o uso de bibliotecas para tarefas específicas, como: i) *openpyxl*[60] para criação/manipulação de planilhas eletrônicas; ii) *wordcloud*[61] para geração de nuvem de palavras e apresentação gráfica de informações textuais em destaque terminológico; e iii) *nltk*[62], uma reconhecida biblioteca para diversas funcionalidades em PLN.

Além disso, destaca-se o desenvolvimento dos algoritmos para tratamento e extração de dados específicos à pesquisa, que utilizaram as técnicas de PLN. Visto que os dados foram exportados em seu formato real, sem tratamento, alguns processos foram fundamentais para adequar um padrão textual e permitir o estabelecimento de regras para que os algoritmos pudessem se comportar de maneira esperada. Ou seja, a linguagem

60 Disponível em: https://openpyxl.readthedocs.io/. Acesso em: 3 out. 2023.

61 Disponível em: https://github.com/amueller/word_cloud. Acesso em: 3 out. 2023.

62 Disponível em: https://www.nltk.org/. Acesso em: 3 out. 2023.

natural, usada na descrição médica, proporciona um texto despadronizado de forma e sintaxe, o qual necessita de uma intervenção antes de realizar a análise, identificação e extração dos dados.

Existem diversas iniciativas para o mapeamento de caracteres e símbolos, como o *American Standard Code for Information Interchange* (ASCII), o *Unicode Transformation Format* (UTF-8), o *American National Standards Institute* (ANSI), Windows-1252, e ISO-8859-1, os quais são padrões que viabilizam a correta exibição de fontes (letras e símbolos) para textos eletrônicos. A maneira como o texto é codificado – denominada *charset* – é importante para proporcionar uma versão final adequada, compreensível e compatível com os mesmos caracteres ou símbolos, codificados pelas terminologias clínicas. O padrão de codificação define a forma como o texto resultante da conversão deve ser representado, além de permitir a leitura e o entendimento humano. Considerando o idioma do projeto, onde tanto a base de dados como as listas terminológicas foram codificadas em português, não houve necessidade de conversão da codificação. Elencam-se, no Quadro 1, os principais processos necessários para adequação dos dados antes da aplicação dos algoritmos de extração.

Quadro 1 - Processos de tratamento textual

Técnica	Objetivo	Referência
stopWords	Retirar termos de pouca ou nenhuma carga semântica à pesquisa	https://github.com/stopwords-iso/stopwords-pt
acentuação	retirar todos os caracteres acentuados e substituir pelo respectivo caracter não acentuado	"á", "à", "ã", "é", "í", "ô", "ç"

Técnica	Objetivo	Referência
excesso de espaços	retirar excesso de espaços em branco a fim de padronizar a tokenização[63] dos termos	expressões regulares
caracteres especiais	retirar caracteres da constante "punctuation" em Python	""!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~"""
quebra de linha	Os textos originais foram formatados com quebras de linhas para facilitar o entendimento humano, mas não são necessários ao processamento computacional. Os caracteres para esta formatação "/n" foram excluídos, tornando o texto uma sequência de caracteres (string) para padronizar o parse[64] do texto pelas funções	método replace para cada \n
minúsculas / maiúsculas	padronizar todo o texto em minúscula para permitir matching[65] perfeitos	método lower() - case-folding[66]

Fonte: Elaborado pelos autores (2023).

63 Sequência de caracteres em grupos separados por espaço.

64 Processo de análise sintática pelo *algoritmo* a fim de identificar estruturas textuais.

65 Processo de encontrar combinações de textos (*strings*) idênticas.

66 Disponível em: https://nlp.stanford.edu/IR-book/html/htmledition/capitalizationcase-folding-1.html. Acesso em: 3 out. 2023.

No processo de PLN, porém, destaca-se a *tokenização*. Trata-se de uma técnica de análise textual que permite a quebra de texto, composto por muitas palavras (ou *strings*), em palavras e símbolos separados em *substrings*, resultando em uma estrutura de dados em forma de lista (*array*), a qual permite uma análise individualizada. O resultado dessa etapa é um texto armazenado em nova coluna no banco de dados. A coluna será usada para a próxima etapa, de extração, processamento e análise de informação. O Quadro 2 exibe um exemplo desse tratamento em relação ao registro original.

Quadro 2 – Tratamento dos textos para processamento

Texto original	Texto tratado com pré-processamento
## GINECOLOGIA ## PACIENTE SUBMETIDO A RESSECÇÃO DE TUMOR DE PAREDE ABDOMINAL COM EXERESE DE SEGMENTO DE APONEUROSE , SOB RAQUE ANESTESIA. ATO CIRURGICO SEM INTERCORRENCIAS ENVIADO MATERIAL PARA EXAME ANATOMO-PATOLOGICO	ginecologia paciente submetido a ressecção de tumor de parede abdominal com exerese de segmento de aponeurose sob raque anestesia ato cirurgico sem intercorrencias enviado material para exame anatomo patológico

Texto original	Texto tratado com pré-processamento
PACIENTE 46 ANOS , G1P1A0 , APRESENTANDO METRORRAGIA REFRATARIO AO TRATAMENTO CONSERVADOR. AO ULTRASSOM TRANSVAGINAL PRESENÇA DE VOLUMOSO MIOMA SUBMUCOSO. APRESENTOU ANEMIA AGUDA COM HB- 7,6G/DL PRE-OP—ASA 1 PRE-ANESTESICO—ASA1 PA- 120X80MMHG MUCOSAS HIPOCORADAS E HIDRATADA AFEBRIL BCNRNF SONS RESP NORMAIS ABDOME LIVRE , SEM SINAIS DE IRRITAÇÃO PERITONEAL UTERO AUMENTADO DE VOLUME COM SANGRAMENTO PERSISTENTE CD- HISTERECTOMIA TOTAL COM ANEXECTOMIA UNILATERAL	paciente 46 anos g1p1a0 apresentando metrorragia refratario ao tratamento conservador ao ultrassom transvaginal presença de volumoso mioma submucoso apresentou anemia aguda com hb 7 6g dl pre op asa 1 pre anestesico asa1 pa 120x80mmhg mucosas hipocoradas e hidratada afebril bcnrnf sons resp normais abdome livre sem sinais de irritação peritoneal utero aumentado de volume com sangramento persistente cd histerectomia total com anexectomia unilateral

Fonte: Dados da pesquisa (2021).

Como padrão de saída, para facilitar a análise especialista, os dados foram gravados em planilhas eletrônicas independentes (Figura 1).

O código-fonte do projeto está gravado no repositório digital *GitHub* para fins de fomento à pesquisa e discussões com outros pesquisadores. As

ferramentas utilizadas no desenvolvimento técnico da pesquisa foram relacionadas a seguir (Quadro 3):

Quadro 3 – Ferramentas utilizadas na condução técnica da metodologia da pesquisa

Ferramenta	Utilização	Acesso
GitHub	Repositório digital para salvar os algoritmos e o conjunto de resultados da extração de dados utilizando as técnicas de PLN foram salvas no repositório digital	https://github.com/amandadsouza/RiLN
PostgreSQL	Servidor de banco de dados para restaurar os dados	https://www.postgresql.org/
Visual Studio Code	Ambiente de desenvolvimento para desenvolvimento de software	https://code.visualstudio.com/
DBeaver	Interface para interação com os diferentes formatos de bancos de dados	https://dbeaver.com/

Fonte: Elaborado pelos autores (2023).

Esses processos relacionados ao tratamento textual constituem o pré-processamento dos dados recebidos pelo hospital.

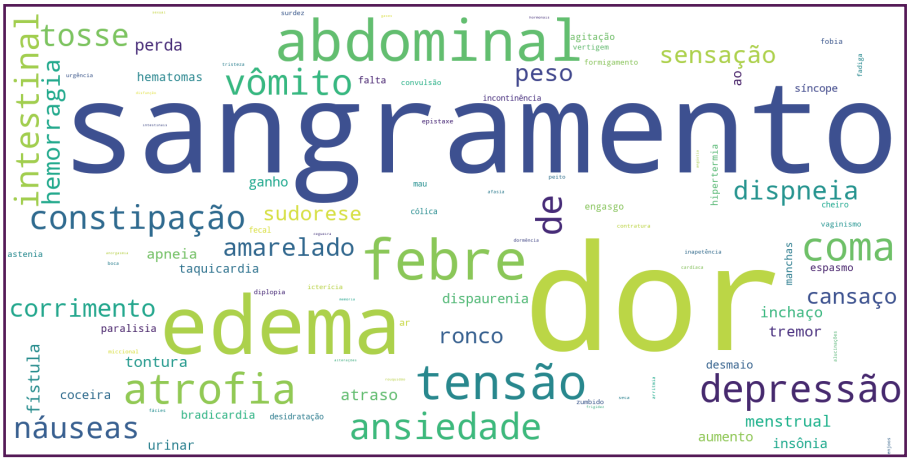
Após a realização deste ajuste textual, procedeu-se com a análise dos dados e o desenvolvimento de algoritmos específicos para o processamento das informações. Na medida que os dados foram manipulados, dois modelos de saída para análise foram produzidos.

5.2.3 PROCESSOS DE SAÍDA (APRESENTAÇÃO)

O modelo de saída para análise se deu por meio de apresentação analítica com gravação de dados tabulares em planilha eletrônica. As análises quantitativas e modelos de frequência foram analisados de forma individual (por registro no modelo de banco de dados) e por agrupamento, permitindo ao especialista um modo de rastreio e percepção de validação do algoritmo que, por vezes, foi aperfeiçoado.

Outro modelo de saída foi a representação das frequências terminológicas no formato nuvem de palavras, conforme a Figura 2. Este modo de visualização permite destacar os termos mais frequentes em determinados contextos, além de facilitar o entendimento do usuário final. Neste processo foi usada a biblioteca *Python word_cloud*[67].

Figura 2 – Nuvem de Palavras para representação dos termos da Terminologia de Interface



Fonte: Dados da pesquisa de doutorado de Souza (2021).

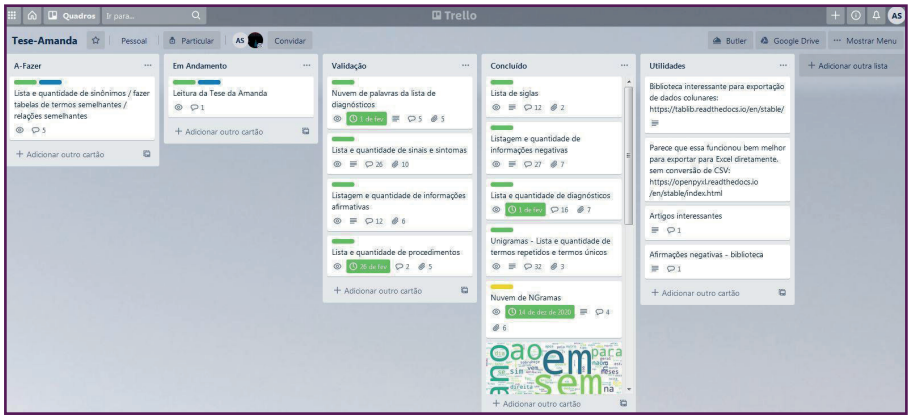
67 Disponível em: https://github.com/amueller/word_cloud. Acesso em: 3 out. 2023.

5.2.4 PROCESSOS ADMINISTRATIVOS

Na condução da pesquisa, diferentes especialistas se uniram a fim de contribuir para o propósito do trabalho. Nesse aspecto, outras ferramentas tecnológicas foram usadas para a gestão de tarefas e documentos.

Para o controle de tarefas em equipe, utilizou-se o *software Trello*[68]. Este programa permitiu estabelecer grupo de tarefas, sua cronologia e *status* para fins de alinhamento de tarefas e responsabilidades. Por meio do método Kanban[69] é possível identificar facilmente tarefas em andamento, bem como seu status e responsável.

Figura 3 - Divisão de tarefas da metodologia pesquisa no *Trello*



Fonte: Captura de tela do Trello utilizada para a pesquisa de doutorado de Souza (2021).

68 Disponível em: https://trello.com/pt-BR. Acesso em: 3 out. 2023.

69 Disponível em: https://blog.trello.com/br/metodo-kanban. Acesso em: 3 out. 2023.

Para o controle dos diversos arquivos gerados (excetuando o código fonte), foi usado a ferramenta *Google Drive* (Quadro 4).

Quadro 4 – Ferramentas utilizadas nos processos administrativos

Ferramenta	Utilização	Acesso
Trello	Software para o controle de tarefas em equipe necessárias na condução da metodologia da pesquisa	https://trello.com/b/ur2kJ7Vm/tese-amanda
Google Drive	Software para centralizar arquivos "em nuvem" para compartilhamento de dados e backup	disponível para usuários do gmail

Fonte: Elaborado pelos autores (2023).

Os processos mencionados nesta seção configuram os aspectos técnicos e operacionais para manipulação dos dados em face de uma necessidade intelectual. A próxima seção abordará os problemas e motivações na adoção das tecnologias para a solução dos problemas específicos no ambiente da pesquisa.

5.3 A TECNOLOGIA E A PESQUISA

Nesta seção, apresentam-se as etapas metodológicas aplicadas de extração e análise de dados por meio de PLN no contexto da Ginecologia e Obstetrícia na pesquisa *stricto sensu*, sendo elas: 1) Definir lista preliminar de termos para delimitar algoritmo; 2) Extrair de dados a partir de ferramenta e técnicas automáticas de PLN: Algoritmos desenvolvido utilizando linguagem *Python*; 3) Analisar os termos extraídos: frequências absolutas e relativas; e 4) Apresentação dos termos por meio de nuvem de palavras e tabelas.

1. Definição das listas preliminares de termos para delimitar algoritmo do PLN

Para delimitar os algoritmos na busca por dados, foram criadas listas preliminares de termos da Ginecologia e Obstetrícia em relação a: sinais e sintomas e diagnóstico, entre outros tipos de termos. As listas de termos preliminares foram criadas junto à equipe do Núcleo Integrado de Pesquisa e Tratamento da Endometriose (NIPTE)[70] e com auxílio de membros da Clínica de Ginecologia e Obstetrícia do local da pesquisa. As listas foram desenvolvidas também com base em terminologias utilizadas em formulários criados pelo NIPTE, formulários da clínica de Ginecologia e Obstetrícia de coleta de dados no REDCap[71], do sistema MV-PEP, anamnese, evolução, com validação e correção de médicos da Ginecologia e Obstetrícia. A listagem de termos do contexto específico da Obstetrícia, foram utilizadas as terminologias de protocolos clínicos e manuais, por exemplo: Secretaria de Estado de Saúde de Minas Gerais e Associação de Ginecologistas e Obstetras de Minas Gerais (2013), Peixoto (2014), Brasil (2004), Brasil (2016), Comissão Nacional de Incorporação de Tecnologias no Sistema Único de Saúde (2016) e Matos *et al.* (2017).

Em relação a doenças/diagnóstico, a CID-10 nas letras N, Q e Z, classifica as doenças relacionadas à especialidade ginecológica, assim na lista preliminar de diagnósticos foi utilizada a CID-10 e indicação de termos pelos ginecologistas do hospital (Organização Mundial da Saúde, 1994).

Em relação aos sinais e sintomas, Souza e Teixeira (2017) abordam que a consulta ginecológica está relacionada a três queixas principais: sangramentos anormais, corrimentos patológicos e dores pélvicas. Assim, os sinais e sintomas referentes a essas queixas deverão ser notificados na anamnese do prontuário. Para a lista de sinais e sintomas também foi utilizado: *National Library of Medicine* (NLM) *Classification 2020 Summer Edition* (Willis, 2019), *Wikipédia*[72], Falcão Júnior *et al.* (2017) e a CID-10 (Organização Mundial da Saúde, 1994) (ANEXO A). Para a pré-lista de

70 Disponível em: https://www.feliciorocho.org.br/servicos/endometriose. Acesso em: 3 out. 2023.

71 Disponível em: https://redcap.feliciorocho.org.br/redcap/index.php. Acesso em: 3 out. 2023.

72 SINAL MÉDICO. *In*: WIKIPÉDIA: a enciclopédia livre. [*S. l.*]: Wikimedia Foundation, 17 ago. 2018. Disponível em: https://pt.wikipedia.org/wiki/Sinal_m%C3%A9dico. Acesso em: 12 out. 2020.

sinais e sintomas é importante incluir sobre os sistemas: circulatório e respiratório; digestivo e abdômen; pele e tecido subcutâneo; nervoso e músculo esquelético; e urinário. Incluir também termos sobre: cognição, percepção, estado emocional e comportamento; fala e voz; sinais e sintomas gerais. Esta pré-lista deve ser verificada por um ginecologista, ou seja, especialista de domínio.

Após projetar os algoritmos, foram realizadas diferentes interações, para extração de termos em relação aos dois tipos documentos da Terminologia de Interface (anamnese e evolução). Foram detectados: a) a presença de sinais e sintomas; b) diagnósticos; e c) termos mais frequentes e termo únicos.

2. Extrair de dados a partir de ferramenta e técnicas automáticas de PLN: Algoritmos desenvolvido utilizando linguagem Python

Passo 1 – Frequência de diagnósticos: Para extrair quais eram diagnósticos e sua quantidade nos documentos eletrônicos (PEPs), foi criada uma lista dos diagnósticos em arquivo texto, que por sua vez foram lidos pelo algoritmo, a fim de criar uma lista (*array*). A leitura dos documentos no banco de dados foi segmentada por tipo de análise ("Anamnese" e "Evolução"). Percorre-se, portanto, cada lista do banco de dados e para cada documento, verifica-se se os diagnósticos (já disponíveis na memória) estão presentes. Uma estrutura de dados organizada por chave: valor, denominada dicionário, em linguagem de programação *Python*, permite armazenar diagnóstico (chave) e sua quantidade (valor) encontrada. Esta estrutura foi usada para posterior gravação, em arquivo de formato planilha eletrônica.

Passo 2 – Frequência de sinais e sintomas: de forma semelhante ao processo de diagnóstico, uma lista de termos sinais e sintomas ginecológicos foi desenvolvida pela pesquisadora. A lista de sinais e sintomas permitiu identificar estes termos nos documentos e levantar sua quantidade para armazenar o resultado em arquivo.

Passo 3 – Quantidade de termos repetidos/únicos: A estratégia para identificar os termos repetidos e únicos deu-se pela estratificação de cada documento em tokens e, a partir dessa grande lista, indicar para cada item da lista, sua repetição ou participação única.

Após a realização dos passos 1, 2 e 3 foi necessário analisar os termos extraídos por meio do PLN.

3. Análise dos termos extraídos: frequências absolutas e relativas

A etapa de extração de termos, descrita anteriormente permitiu extrair: a) frequência de diagnósticos, (para essa tarefa foi construída uma lista de termos para delimitar o *algoritmo*); b) frequência de sinais e sintomas, (para essa tarefa foi construída uma lista de termos para delimitar o *algoritmo*); c) frequência de termos únicos e repetidos, (para essa tarefa não foi necessária uma lista de termos para delimitar o *algoritmo*).

Posteriormente ao PLN foi realizada estatística dos termos extraídos pelos algoritmos, para fins de análise de frequência absoluta (f_a quantidade de vezes que cada termo aparece) e frequência relativa (f_r é o percentual de vezes que cada f_a aparece em relação ao total da amostra n). Foram avaliadas as seguintes variáveis:

- Frequência absoluta (total de termos da Terminologia de Interface) e frequência relativa de termos sobre sinais e sintomas (total) na denominada Terminologia de Interface, ou seja, texto clínico do prontuário ou jargão médico (Schulz *et al.*, 2017);

- Frequência absoluta (total de termos da Terminologia de Interface) e frequência relativa de termos sobre diagnósticos (total) na Terminologia de Interface;

- Frequência absoluta (total de termos da Terminologia de Interface) e frequência relativa de termos repetidos na Terminologia de Interface;

- Frequência absoluta (total de termos da Terminologia de Interface) e frequência relativa de termos únicos na Terminologia de Interface;

- Listar frequência relativa dos termos únicos, termos repetidos, sinais e sintomas, e diagnóstico, em formato de tabelas e nuvem de palavras.

Os cálculos de frequência absoluta e frequência relativa foram realizados utilizando a fórmula (Figura 5).

Figura 5 – Fórmula para cálculo da frequência

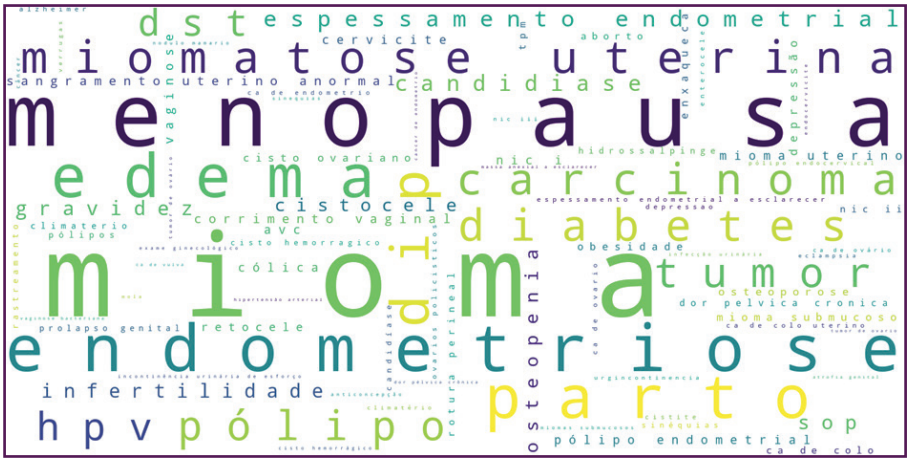
$$x = \frac{fa * 100}{n}$$

Fonte: Elaborado pelos autores (2023).

Apresentação dos termos por meio de nuvem de palavras e tabelas

Após a análise das frequências dos termos (Terminologia de Interface), esses os resultados foram ilustrados conforme exemplos da Figura 6 e Tabela 1.

Figura 6 – Nuvem de palavras de diagnóstico da Terminologia de Interface



Fonte: Souza (2021, p. 158).

Esta representação em nuvem de palavras ilustra quais os termos mais representativos nas notificações em campos abertos do PEP, ou seja, na

Terminologia de Interface. Já a Tabela 1 representa, de forma quantitativa, os termos mais frequentes.

Tabela 1 – Exemplo da frequência absoluta de diagnóstico da Terminologia de Interface

Termo	*f*a	Termo	*f*a
mioma	1986	espessamento endometrial	274
menopausa	1199	candidiase	257
endometriose	948	gravidez	207
parto	806	sop	206
edema	800	infertilidade	200
dip	602	cistocele	180
miomatose uterina	419	osteopenia	178
diabetes	402	sangramento uterino anormal	169
carcinoma	390	corrimento vaginal	168
pólipo	351	NIC I	158
tumor	349	vaginose	151
hpv	306	avc	150
dst	277	cólica	150

Fonte: Souza (2021).

A listagem de termos únicos foi importante para identificar os termos não recuperados pelo algoritmo. A Tabela 2 descreve as frequências absolutas e relativas da Terminologia de Interface.

Tabela 2 – Frequência absoluta (n) e frequência relativa (%) de termos únicos e repetidos na Terminologia de Interface

Variáveis	Terminologia de Interface	
	n	%
Termos únicos	16.248	1,19
Termos repetidos	1.348.116	**98,81**
Total	**1.364.364**	**100**

Fonte: Souza (2021).

O tópico quatro a seguir aborda estudos semelhantes na literatura.

5.4 ANÁLISE TERMINOLÓGICA DE TEXTO CLÍNICO POR MEIO DE PROCESSAMENTO DE LINGUAGEM NATURAL: UM BREVE RELATO DA LITERATURA

Pesquisas envolvendo o processamento de texto clínico (*Text Mining*) para conectar terminologias de referência e agregação já foram realizadas. Soderland *et al.* (1995) cita um trabalho da *The National Center for Intelligent Information Retrieval* (CIIR) da *University of Massachusetts* na cidade de Amherst, na mineração de textos de PEP para realizar uma classificação automatizada com a utilização da terminologia de agregação, CID-9. A pesquisa buscou automatizar os códigos CID-9 para sumários de alta hospitalar, na qual foi realizada a tarefa de automatizar a "compreensão do conteúdo dos sumários para rotular frases que contenham informação relativa a (1) diagnóstico e (2) sinais ou sintomas de doença" (Soderland *et al.*, 1995, p. 1, tradução nossa). Pesquisas que envolvem extração de informação de texto clínico de PEP foram realizadas por Kim *et al.* (2017), Wang *et al.* (2012), Zhou *et al.* (2006), Meystre *et al.* (2010), entre outros. No estudo de Kim *et al.* (2017), utilizou-se o TM para extrair informação sobre intervenção coronária percutânea. Wang *et al.* (2012) utilizaram técnica de ML por meio do *Support Vector Machine* (SVM), para extrair resultados

de diagnóstico de textos clínicos do PEP sobre angiografia coronariana e câncer de ovário.

Um estudo de Zhou *et al.* (2006) descreveu um sistema de extração de Informação Médica (MedIE), que extraiu uma variedade de informações e registros clínicos de textos clínicos do paciente sobre queixas de doenças da mama. Meystre *et al.* (2010) fizeram uma revisão da literatura sobre pesquisas recentes em desidentificação de documentos de texto clínico narrativo em PEP. Estudos como esses demonstram a importância de se analisar a necessidade da conexão de dados entre Sistemas de Informação em Saúde (SISs).

Outro trabalho relevante que retrata um conjunto de ações direcionado à aplicação de PLN em textos biomédicos e mineração de textos, está em (Kafkas; Toonsi; Alsaedi, 2023). Destaca-se o uso dos corpus derivados do algoritmo da *Google Bidirectional Encoder Representations from Transformers*[73] (BERT) e os processos de transformação, tratamento, *tokenização*, reconhecimento de entidades, normalização e identificação das relações no texto em linguagem natural.

Por fim o estudo de Souza *et al.* (2022) abordou o trabalho do Bibliotecário da Saúde, junto à equipe médica, em pesquisas sobre as terminologias clínicas em prontuários na área de saúde, no qual este profissional precisa saber recuperar dados a partir da PLN e técnicas de inteligência artificial com a finalidade de contribuir para a melhoria do tratamento da informação, no cuidado em saúde. Na pesquisa foram identificados por meio de vocabulários terminológicos com apoio de algoritmos em *Python* termos relacionados que envolviam: presença de sinais, sintomas, expressões negativas e afirmativas, termos únicos e mais frequentes, siglas e abreviaturas, entre outros.

73 Disponível em: https://arxiv.org/pdf/1810.04805.pdf. Acesso em: 3 out. 2023.

5.5 CONSIDERAÇÕES FINAIS

Este estudo relatou uma metodologia de PLN e ferramentas de gestão de projeto, aplicada a pesquisa *stricto sensu* (doutorado) no âmbito da CI e no campo da saúde, especificamente na área de Ginecologia e Obstetrícia. Para isso foi demonstrado as etapas para se realizar análise terminológica de textos clínicos de campo aberto do PEP, denominada Terminologia de Interface, com dados reais de hospital privado.

A análise de textos clínicos é uma área de grande importância no contexto da Medicina. Neste trabalho, observou-se que a análise das informações registradas no PEP permitiu traçar as características da terminologia clínica, que retrataram as principais características (diagnósticos, sinais e sintomas) em um público específico da área de saúde (Ginecologia e Obstetrícia). Essa informação foi disponibilizada à instituição hospitalar a fim de embasar decisões futuras que impactem em investimentos estruturais, na contratação de especialistas, na seleção de pacientes para estudos clínicos, entre outros. Os resultados também permitiram a análise dos artefatos terminológicos (Terminologia de Interface), evidenciando a necessidade de atualização e aproximação da realidade de representação do diagnóstico, sinais e sintomas em terminologia clínicas como a CID-10 e em linguagem natural (jargão médico). Ou seja, os artefatos terminológicos precisam evoluir na medida que os profissionais da área necessitam representar seu conhecimento e informações na descrição do diagnóstico, sinais e sintomas e evolução do paciente.

Reconhece-se, contudo, que as limitações no processamento e análise da linguagem natural ainda são grandes no contexto da sintática e semântica. Problemas como erros ortográficos, abreviações, mnemônicos, pontuações, quebras de linha, entre outros citados no trabalho original, foram evidências de desafios que precisam ser tratados para uma análise confiável. A principal dificuldade em analisar o jargão médico utilizado no PEP, referiu-se aos seus aspectos epistemológicos que dependem fortemente do contexto médico. A intenção desta análise terminológica não foi criar terminologias ou guidelines para as mesmas, mas sim entender suas características, possibilidades de extração e de análises.

Uma das principais contribuições da pesquisa, foi indicar formas de delimitar o algoritmo no domínio da Ginecologia e Obstetrícia, na língua

portuguesa, e a partir da extração de termos da Terminologia de Interface, possibilitar futuramente o enriquecimento de artefatos terminológicos como as Ontologias Biomédicas e Vocabulários Controlados.

REFERÊNCIAS

ALVARENGA, Lídia. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. **Encontros Bibli**, Florianópolis, v. 8, n. 15, p. 18-40, 1. sem. 2003. DOI: https://doi.org/10.5007/1518-2924.2003v8n15p18. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2003v8n15p18. Acesso em: 10 out 2019.

BLAKE, Catherine. Text mining. **Annual Review of Information Science and Technology**, [*s. l.*], v. 45, n. 1, p. 121-155, jan. 2011. DOI: https://doi-org.ez106.periodicos.capes.gov.br/10.1002/aris.2011.1440450110. Disponível em: https://asistdl-onlinelibrary-wiley.ez106.periodicos.capes.gov.br/doi/10.1002/aris.2011.1440450110. Acesso em: 3 out. 2023.

BLOBEL, Bernd. Interoperable EHR Systems: challenges, standards and solutions. **European Journal for Biomedical Informatics**, [*s. l.*], v. 14, n. 2, p. 10-19, 2018. DOI: https://doi.org/10.24105/ejbi.2018.14.2.3. Disponível em: https://www.ejbi.org/scholarly-articles/interoperable-ehr-systems--challenges-standards-and-solutions.pdf. Acesso em: 2 out. 2023.

BODENREIDER, Olivier. Lexical, terminological and ontological resources for biological text mining. *In*: ANANIDOU, Sophia; MCNAUGHT, John (ed.). **Text mining for biology and biomedicine**. London, UK: Artech House, 2006. Chapter 3, p. 43-66. Disponível em: https://lhncbc.nlm.nih.gov/LHC-publications/PDF/pub2006007.pdf. Acesso em: 6 out. 2023.

BRASIL. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Ações Programáticas Estratégicas. **Política nacional de atenção integral à saúde da mulher**: princípios e diretrizes. Brasília: Ministério da Saúde, 2004. 82 p. (Série C. Projetos, Programas e Relatórios). Disponível em : http://bvsms.saude.gov.br/bvs/publicacoes/politica_nac_atencao_mulher.pdf. Acesso em: 8 jan. 2020.

BRASIL. Ministério da Saúde; INSTITUTO SÍRIO-LIBANÊS DE ENSINO E PESQUISA. **Protocolos da Atenção Básica**: saúde das mulheres. Brasília: Ministério da Saúde; Instituto Sírio-Libanês de Ensino e Pesquisa, 2016. 230 p. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/protocolos_atencao_basica_saude_mulheres.pdf. Acesso em: 23 set 2023.

CAMPOS, Maria Luiza de Almeida. **Linguagem documentária**: teorias que fundamentam sua elaboração. Niterói, RJ: EdUFF, 2001. 133p.

COMISSÃO NACIONAL DE INCORPORAÇÃO DE TECNOLOGIAS (Brasil). **Diretrizes Nacionais de Assistência ao Parto Normal**. Brasília: Ministério da Saúde, maio 2016. 399 p. (Relatório de recomendação, nº 211).

CONSELHO REGIONAL DE MEDICINA DO DISTRITO FEDERAL. **Prontuário médico do paciente**: guia para uso prático. Brasília: CRM-DF, 2006. Disponível em: https://crmdf.org.br/wp-content/uploads/2021/05/prontuario-medico-do-paciente-1.pdf. Acesso em: 2 out. 2023.

DALIANIS, Hercules. Characteristics of patient records and clinical corpora. *In*: DALIANIS, Hercules. **Clinical text mining**: secondary use of electronic patient records. [*S. l.*]: Springer Cham, 2018b. Chapter 4, p. 21-34. DOI: https://doi.org/10.1007/978-3-319-78503-5_4. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-78503-5_4. Acesso em: 2 jan. 2019.

DALIANIS, Hercules. Medical classifications and terminologies. *In*: DALIANIS, Hercules. **Clinical text mining**: secondary use of electronic patient records. [*S. l.*]: Springer Cham, 2018a. Chapter 5, p. 35-43. DOI: https://doi.org/10.1007/978-3-319-78503-5_5. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-78503-5_5. Acesso em: 2 jan. 2019.

FALCÃO JÚNIOR, João Oscar Almeida *et al*. **Ginecologia e obstetrícia**: assistência primária e saúde da família. Rio de Janeiro: MedBook, 2017.

KAFKAS, Senay; TOONSI, Sumyyah; ALSAEDI, Sakhaa. T1: Natural Language Processing Tutorial for Biomedical Text Mining (Half-day). *In*: INTERNATIONAL CONFERENCE ON BIOMEDICAL ONTOLOGY, 14.; SEMINAR ON ONTOLOGY RESEARCH IN BRAZIL JOINT CONFERENCE, 16., 2023, Brasília. [Tutorial]. Brasília, DF: Faculdade de Ciência da Informação, UnB,

2023. Disponível em: https://github.com/stoonsi/ICBO-NLP-for-Biomedical--Text-Mining-tutorial/tree/main. Acesso em: 19 set. 2023.

KIM, Yoon Seob *et al.* Extracting information from free-text electronic patient records to identify practice-based evidence of the performance of coronary stents. **PLoS One**, San Francisco, v. 12, n. 8, e0182889, Aug. 2017. DOI: https://doi.org/10.1371/journal.pone.0182889. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182889. Acesso em: 5 out. 2023.

MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. Cambridge, Massachusetts: MIT Press, 1999. 620 p.

MATOS, Margarida Santos *et al.* **Manual de ginecologia**. Salvador: EBMSP, 2017.

MEYSTRE, Stephane M. *et al*. Automatic de-identification of textual documents in the electronic health record: a review of recent research. **BMC Medical Research Methodology**, [London], v. 10, article number 70, 2010.

MINAS GERAIS. Secretaria de Estado de Saúde; ASSOCIAÇÃO DE GINECOLOGISTAS E OBSTETRAS DE MINAS GERAIS. **Atenção à saúde da gestante**: novos critérios para estratificação de risco e acompanhamento da gestante: Programa Viva Vida: Projeto Mães de Minas. Belo Horizonte: SES-MG, maio 2013. [Nota Técnica Conjunta]. Disponível em: https://www.conass.org.br/liacc/wp-content/uploads/2015/02/Oficina--3-Estratificacao-de-Risco-GESTANTE.pdf. Acesso em: 5 out. 2023.

MIÑARRO-GIMÉNEZ, Jose A. *et al*. Quantitative analysis of manual annotation of clinical text samples. **International Journal of Medical Informatics**, [*s. l.*], v. 123, p. 37-48, Mar. 2019. DOI: https://doi.org/10.1016/j.ijmedinf.2018.12.011. Disponível em: https://www.sciencedirect.com/science/article/pii/S1386505618305446. Acesso em: 2 out. 2023.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **CID-10**: Classificação Estatística Internacional de Doenças. v. 1. [*S. l.*]: Edusp, 1994. Disponível em: https://www.medicinanet.com.br/cid10.htm. Acesso em: 19 set. 2023.

PEIXOTO, Sérgio. **Manual de assistência pré-natal**. 2. ed. São Paulo: FEBRASGO, 2014. Disponível em: https://www.abenforj.com.br/site/arquivos/manuais/304_Manual_Pre_natal_25SET.pdf. Acesso em: 5 out. 2023.

ROGERS, Jeremy. **Using medical terminologies**. 2005. Disponível em: http://www.cs.man.ac.uk/~jeremy/HealthInf/RCSEd/termi nologyusing. Htm. Acesso em: 5 mar. 2019.

SCHULZ, Stefan *et al*. Interface terminologies, reference terminologies and aggregation terminologies: a strategy for better integration. *In*: GUNDLAPALLI, Adi V.; JAULENT, Marie-Christine (ed.). **MEDINFO 2017**: precision healthcare through informatics. Proceeding of the 16th World Congress on Medical and Health Informatics. Amsterdam: IOS Press; International Medical Informatics Associations, c2017. p. 940-944. (Studies in Health Technology and Informatics, v. 245). DOI: https://doi.org/10.3233/978-1-61499-830-3-940. Disponível em: https://ebooks.iospress.nl/publication/48291. Acesso em: 2 out. 2023.

SHORTLIFFE, Edward H. Biomedical informatics: the science and the pragmatics. *In*: SHORTLIFFE, Edward H.; CIMINO, James J. (ed.). **Biomedical informatics**: computer applications in health care and biomedicine. 4th ed. London: Springer-Verlag, 2014. Cap. 1, p. 3-37. DOI: https://doi.org/10.1007/978-1-4471-4474-8_1.

SODERLAND, Stephen *et al*. **Machine learning of text analysis rules for clinical records**. [*S. l.: s. n.*], 1995. Disponível em : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.1340. Acesso em: 5 mar. 2019.

SOUZA, Amanda Damasceno de. Exigências éticas da pesquisa. *In*: CASTELLANO, Elisabete Gabriela; ASSIS, Orly Zucatto Mantovani de (org.). **Metodologia do trabalho e da pesquisa científica**. São Carlos: Diagrama, 2022. p. 487-507.

SOUZA, Amanda Damasceno de *et al.* O bibliotecário e a pesquisa terminológica em prontuários na área de saúde. *In*: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 29., 26 a 30 de setembro de 2022, [evento *online*]. **Anais** [...]. São Paulo: FEBAB, 2022. v. 1. n. 1. [Eixo 4 - Ciência da Informação: diálogos e conexões].

Disponível em: https://portal.febab.org.br/cbbd2022/article/view/2550. Acesso em: 23 set. 2023.

SOUZA, Amanda Damasceno de. **O discurso na prática clínica e as terminologias de padronização**: investigando a conexão. 2021. Tese (Doutorado em Gestão e Organização do Conhecimento) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: http://hdl.handle.net/1843/38044. Acesso em: 23 set. 2023.

SOUZA, José Helvécio Kalil de; TEIXEIRA, Ivana Vilela. Anamnese e exame físico em ginecologia: propedêutica em ginecologia: aspectos atuais. *In*: FALCÃO JÚNIOR, João Oscar de Almeida *et al*. **Ginecologia e obstetrícia**: assistência primária e saúde da família. Rio de Janeiro: MedBook, 2017. cap. 18, p. 249-256.

WANG, Zhuoran *et al*. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. **PLoS One**, San Francisco, v. 7, n. 1, e30412, Jan. 2012. DOI: https://doi.org/10.1371/journal.pone.0030412. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030412. Acesso em: 5 out. 2023.

WILLIS, Sharon R. NLM Classification 2019 Summer Edition Now Available. Posted: 2019 Oct. 1. **NLM Technical Bulletin**, Bethesda, n. 430, e4, Sep-Oct 2019. Disponível em: https://www.nlm.nih.gov/pubs/techbull/so19/so19_nlm_classification_summer_2019.html. Acesso em: 21 jul. 2020.

ZHOU, Xiaohua *et al*. Approaches to text mining for clinical medical records. *In*: ANNUAL ACM SYMPOSIUM ON APPLIED COMPUTING, 21st, 2006, Dijon, France, April 23-27. **Proceedings** [...]. New York: ACM, April, 2006. p. 235-239. DOI: https://doi.org/10.1145/1141277.1141330. Disponível em: http://www.ischool.drexel.edu/faculty/hhan/SAC2006_CAHC.pdf. Acesso em: 20 jun. 2019.

ANEXO A – EXEMPLO DE CLASSIFICAÇÃO DAS DOENÇAS GINECOLÓGICAS E DE SINAIS E SINTOMA DA CID-10 EM PORTUGUÊS

Fração da Tabela R – Sinais e Sintomas
R00 - Anormalidades do Batimento Cardíaco
R01 - Sopros e Outros Ruídos Cardíacos
R02 - Gangrena Não Classificada em Outra Parte
R03 - Valor Anormal da Pressão Arterial Sem Diagnóstico
R04 - Hemorragia Das Vias Respiratórias
R05 - Tosse
R06 - Anormalidades da Respiração
R07 - Dor de Garganta e no Peito
R09 - Outros Sintomas e Sinais Relativos Aos Aparelhos Circulatório e Respiratório
R10 - Dor Abdominal e Pélvica
R11 - Náusea e Vômitos
R12 - Pirose
R13 - Disfagia
R14 - Flatulência e Afecções Correlatas
R15 - Incontinência Fecal
R16 - Hepatomegalia e Esplenomegalia Não Classificadas em Outra Parte

Fração da Tabela R – Sinais e Sintomas
R17 - Icterícia Não Especificada
R18 - Ascite
R19 - Outros Sintomas e Sinais Relativos ao Aparelho Digestivo e ao Abdome
R20 - Distúrbios da Sensibilidade Cutânea
R21 - Eritema e Outras Erupções Cutâneas Não Especificadas
R22 - Tumefação, Massa ou Tumoração Localizadas da Pele e do Tecido Subcutâneo
R23 - Outras Alterações Cutâneas
R25 - Movimentos Involuntários Anormais
R26 - Anormalidades da Marcha e da Mobilidade
R27 - Outros Distúrbios da Coordenação
R29 - Outros Sintomas e Sinais Relativos Aos Sistemas Nervoso e Osteomuscular
R30 - Dor Associada à Micção
R31 - Hematúria Não Especificada
R32 - Incontinência Urinária Não Especificada
R33 - Retenção Urinária
R34 - Anúria e Oligúria

Fração da Tabela R – Sinais e Sintomas
R35 - Poliúria
R36 - Secreção Uretral
R39 - Outros Sintomas e Sinais Relativos ao Aparelho Urinário
R40 - Sonolência, Estupor e Coma
R41 - Outros Sintomas e Sinais Relativos à Função Cognitiva e à Consciência
R42 - Tontura e Instabilidade
R43 - Distúrbios do Olfato e do Paladar
R44 - Outros Sintomas e Sinais Relativos às Sensações e às Percepções Gerais
R45 - Sintomas e Sinais Relativos ao Estado Emocional
R46 - Sintomas e Sinais Relativos à Aparência e ao Comportamento
R47 - Distúrbios da Fala Não Classificados em Outra Parte
R48 - Dislexia e Outras Disfunções Simbólicas, Não Classificadas em Outra Parte
R49 - Distúrbios da Voz
[...]

Fonte: Tabela CID-10 de sinais e sintomas[74].

74 Disponível em: https://www.medicinanet.com.br/cid10/r.htm. Acesso em: 15 fev. 2020.

DADOS DOS AUTORES:

Amanda Damasceno de Souza

Amanda Damasceno de Souza é doutora em Gestão e Organização do Conhecimento e mestre em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG) e graduada em Biblioteconomia também pela UFMG.Atuou como Bibliotecária Clínica no Hospital Felício Rocho e na Oncologia do Hospital Belo Horizonte e da Santa Casa de Belo Horizonte. Atualmente é Coordenadora do Comitê de Ética em Pesquisa da Universidade FUMEC e é membro dos Grupos de Pesquisa Núcleo de Estudos e Pesquisas sobre Recursos, Serviços e Práxis Informacionais (NERSI), do grupo Representação do Conhecimento, Ontologias e Linguagem (ReCOL), do NCOR-BR, do Comitê ABNT CE 021:002.032 e da Comissão de Pesquisa e Iniciação Científica (CoPIC) da Universidade FUMEC. Atua como docente no Bacharelado em Estética, Bacharelado em Administração e no Programa de Pós-Graduação em Tecnologia da Informação Comunicação e Gestão do Conhecimento (PPGTICGC) da Universidade FUMEC e editora das Revistas Estética em Movimento e Código-31.

https://orcid.org/0000-0001-6859-4333
amanda.dsouza@fumec.br

Eduardo Ribeiro Felipe

Eduardo Felipe é professor Adjunto do curso de Engenharia de Computação na Universidade Federal de Itajubá. Doutor em Gestão e Organização do Conhecimento pela UFMG. Mestre em Ciência da Informação pela UFMG e Pós-graduado em Engenharia de Software pela PUC-Minas. Possui graduação como Tecnólogo em Processamento de Dados pelo Centro Universitário Newton Paiva. Membro do grupo de pesquisa Representação do Conhecimento, Ontologias e Linguagem (ReCOL). Membro do Grupo de pesquisa; Laboratório de Robótica, Sistemas Inteligentes e Complexos - RobSIC. Membro do Conselho Universitário CONSUNI, Membro do Núcleo Docente Estruturante NDE (Computação). Atua nas áreas de Linguagens de Programação, Desenvolvimento Web e Mobile, Recuperação da Informação e Ontologias.

https://orcid.org/0000-0003-1690-2044
eduardo.felipe@unifei.edu.br

Fernanda Farinelli

Fernanda Farinelli é Professora Adjunta na Faculdade de Ciência da Informação da UnB. Doutora em Gestão e Organização do Conhecimento pela Escola de Ciência da Informação da UFMG pesquisando o tema ontologias formais realistas como solução de integração semântica de dados. Ontologista responsável pelo projeto da OntONeo (Ontologia do domínio obstétrico e neonatal). Pesquisadora visitante no Departamento de Filosofia e no Departamento de Informática Biomédica da Universidade Estadual de Nova York em Buffalo entre 05/2015 e 04/2017. Mestre em Administração de Empresas com ênfase em Gestão estratégica da informação (Fundação Pedro Leopoldo/MG). Especialista em Banco de Dados (UNI-BH). Bacharel em Ciência da Computação (PUC-MG). Possui mais de 15 anos de experiência em Gestão de Dados atuando com administração de banco de dados, arquitetura e administração de dados e implantação de governança de dados em grandes empresas como Unisys Brasil, Cedro Têxtil, Prodemge. Atua há cerca de 15 anos como docente em cursos de graduação e pós-graduação em renomadas instituições de ensino no estado de Minas Gerais como PUC-MG, IEC, Fundação Pedro Leopoldo, Universidade de Itaúna, Faculdade Cotemig, Unipac e IGTI. Possui as certificações CDMP, CBIP, CDP e OCP.

https://orcid.org/0000-0003-2338-8872
fernanda.farinelli@unb.br

Como referenciar o capítulo 5:

SOUZA, Amanda Damasceno de; FELIPE, Eduardo Ribeiro; FARINELLI, Fernanda. Extração e análise de dados registrados em texto livre de prontuário eletrônico do paciente por meio de processamento de linguagem natural. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 5. p. 103-138. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap5.

6. POTENCIALIDADES INVESTIGATIVAS UTILIZANDO ANÁLISE DE REDES SOCIAIS[75]

Alex Fabianne de Paulo

6.1 DEFINIÇÃO E CONTEXTO HISTÓRICO

A Análise de Redes Sociais (ARS) é uma abordagem interdisciplinar que se concentra no estudo das relações e conexões entre atores em uma rede social para revelar padrões e dinâmicas subjacentes. Essa disciplina procura compreender como os indivíduos, grupos, organizações ou entidades interagem, comunicam-se e influenciam-se mutuamente por meio dessas relações, fornecendo uma visão única das complexidades das interações sociais. A Análise de Redes Sociais (ARS) não apenas descreve as redes, mas também analisa e interpreta seu significado, impacto e implicações em várias áreas do conhecimento (1), (2).

A história da ARS remonta ao início do século XX (3), quando os primeiros estudos sobre a sociometria, uma precursora da ARS, foram conduzidos por Jacob Moreno que desenvolveu métodos para avaliar as interações sociais em grupos, criando gráficos sociométricos para representar visualmente as relações entre indivíduos em contextos como escolas e organizações (4), (5). Seu trabalho pioneiro contribuiu significativamente para o entendimento das dinâmicas sociais em grupos humanos. No entanto, foi nas décadas de 1960 e 1970 que a ARS começou a ganhar reconhecimento acadêmico e a se expandir para outras disciplinas. Nesse período, sociólogos como Harrison White e John Barnes desenvolveram abordagens teóricas sólidas para a análise de redes sociais, aplicando conceitos matemáticos e estatísticos para medir e descrever as características das

75 Este capítulo foi desenvolvido com assistência de ferramentas de inteligência artificial.

redes sociais. Mark Granovetter introduziu o conceito de "laços fracos" na análise dessas redes, destacando a importância das conexões menos íntimas nas redes sociais para a difusão de informações e oportunidades (6). Essa ideia revolucionou a forma como é entendida a influência das redes sociais em nossas vidas cotidianas.

A ARS também encontrou aplicações em campos diversos, incluindo antropologia, psicologia, administração, epidemiologia, biologia, genética e ciência da informação. Em cada uma dessas áreas, a ARS fornece uma perspectiva única que permite aos interessados analisar os sistemas sociais subjacentes e entender melhor como as informações, influências e recursos fluem por meio das conexões interpessoais (7). Além disso, com o advento da era digital e das mídias sociais, a ARS encontrou novas oportunidades de pesquisa, permitindo que os cientistas sociais explorem redes sociais on-line, comportamentos de compartilhamento de informações e influência digital em escala global. Essa evolução continua a impulsionar a expansão e a relevância da Análise de Redes Sociais nos dias de hoje, constituindo um método poderoso a ser usado além das fronteiras das análises e estatísticas tradicionais (8), (9).

6.2 ARS E SUA UTILIZAÇÃO EM CIÊNCIAS SOCIAIS APLICADAS

A Análise de Redes Sociais (ARS) desempenha um papel relevante na área de Ciências Sociais Aplicadas, proporcionando uma perspectiva única e uma metodologia robusta para investigar e compreender as complexidades das interações, estruturas de relacionamentos e dinâmicas sociais em uma variedade de contextos. Essa abordagem interdisciplinar tem se mostrado crucial nessa área por várias razões conceituais e práticas, tais como:

1. Compreensão das estruturas sociais: ARS permite uma análise sistemática das redes sociais que sustentam as sociedades e as organizações. Isso ajuda a desvelar as estruturas subjacentes das comunidades, grupos e instituições, fornecendo *insights* sobre como as conexões entre os indivíduos moldam o comportamento, as interações e a disseminação de informações. Por meio da identificação de líderes, influenciadores e *gatekeepers* nas redes, a ARS revela as hierarquias e as interações de poder dentro dessas estruturas (10).

2. Aplicações práticas em diversas áreas: ARS encontra aplicações em uma ampla gama de campos das Ciências Sociais Aplicadas, incluindo sociologia, psicologia, administração, *marketing*, saúde pública, antropologia, arquivologia, ciência política, comunicação, biblioteconomia, gestão da informação e ciência da informação. Essa versatilidade permite que profissionais apliquem a ARS para abordar questões específicas em seus respectivos domínios. Por exemplo, ARS pode ser usada para estudar redes de amizade em contextos escolares, redes de colaboração em empresas, redes de transmissão de doenças e redes políticas (10), (12).

3. Identificação de padrões e tendências: ARS permite a detecção de padrões e tendências que podem não ser aparentes por meio de métodos de pesquisa tradicionais. Por intervenção da análise de métricas de centralidade, coesão de grupos, difusão de informações e outros conceitos-chave advindo da Teoria do Grafos (7), os pesquisadores podem identificar nós críticos, comunidades, laços fortes e fracos, bem como prever o comportamento futuro das redes. Isso contribui para uma compreensão mais profunda dos fenômenos sociais em estudo (11), (13), (14).

4. Informações sobre comportamento individual e coletivo: ARS fornece informações valiosas sobre o comportamento individual e coletivo. Ela ajuda a entender como as redes sociais influenciam as decisões e as ações dos indivíduos, bem como a disseminação de informações, inovações e comportamentos que impactam grupos e comunidades. Isso é fundamental para o desenvolvimento de estratégias eficazes de intervenção, políticas públicas e tomada de decisões em vários campos (15), (16).

5. Estudos interdisciplinares e colaboração: ARS promove a colaboração interdisciplinar, permitindo que profissionais de diferentes áreas se unam para abordar questões complexas. Ela proporciona uma linguagem comum e estrutura analítica que facilita a comunicação e a colaboração entre cientistas sociais, matemáticos, epidemiologistas, economistas e outros profissionais. Essa colaboração enriquece a pesquisa e a aplicação prática da ARS (17), (18).

O Quadro 1 mostra alguns trabalhos científicos desenvolvidos na área de Ciências Sociais Aplicadas utilizando ARS.

Quadro 1 – Exemplos de estudos e aplicações

Aplicação	Contexto	Artigos
Análise de coautoria	Estudos que mapeiam a colaboração entre pesquisadores em determinadas áreas, identificando quem são os atores principais, quais são as sub-redes mais ativas e como a colaboração se desenvolveu ao longo do tempo.	(19)–(26)
Redes de inovação	Pesquisas que analisam como as empresas e instituições colaboram em ecossistemas de inovação. Esses estudos podem revelar quais organizações são centrais para a disseminação de inovações ou tecnologias.	(27)–(37)
Análise de redes políticas	Estudos que mapeiam as relações entre políticos, doadores de campanha, lobistas e outras entidades para entender padrões de influência e poder.	(38)–(46)
Redes de comércio e negócios	Investiga como os negócios estão interconectados, seja em termos de comércio bilateral, cadeias de fornecimento globais ou outras formas de relacionamento comercial.	(47)–(57)
Redes de comunicação em movimentos sociais	Usando dados de mídias sociais ou outras fontes, pesquisadores analisam como as informações são disseminadas dentro de movimentos sociais ou como os movimentos se organizam em rede.	(58)–(70)

Aplicação	Contexto	Artigos
Interação em plataformas online	Análise de redes de interação em plataformas como X (Twitter), Facebook, Instagram, Tiktok ou fóruns on-line para entender padrões de comunicação, disseminação de informações ou formação de comunidades.	(71)–(79)
Análise de redes em educação	Estudos que observam como estudantes ou educadores interagem em ambientes de aprendizado colaborativos, buscando entender padrões de colaboração, mentorias ou trocas de informações.	(80)–(93)
Relações interorganizacionais	Examina como diferentes organizações (por exemplo, ONGs, empresas, governos) interagem em projetos conjuntos ou em contextos específicos, como resposta a desastres.	(94)–(105)
Redes de confiança e capital social	Explora como as redes de confiança são formadas em comunidades e como elas influenciam variáveis como cooperação, solidariedade e desenvolvimento local.	(106)–(118)
Prospecção tecnológica	Estudos que avaliam tendências tecnológicas, especialmente baseado em dados de patentes, utilizando algoritmos e técnicas diversas bem como ferramentas de análise de redes sociais para investigação e representação dos resultados.	(119)–(134)

Aplicação	Contexto	Artigos
Cooperação e parcerias entre organizações	Avalia aspectos inerentes aos acordos de cooperação entre organizações e relacionados à gestão da inovação, Pesquisa e Desenvolvimento (P&D), hélice-tripla, cooperação empresa-universidade e inovação aberta e desenvolvimento tecnológico.	(135)–(153)
Mapeamento e tendências de pesquisas e inovação	Métricas (grau, proximidade, intermediação, densidade, menor caminho, modularidade) e a própria representação visual provida pela análise de redes sociais são utilizadas para complementar estudos bibliométricos que anteriormente eram mais descritivos, o que torna tais estudos mais sofisticados e com maior potencial de análise e descoberta de resultados.	(154)–(166)

Fonte: elaborado pelo autor (2023).

6.3 101 PARA ARS: CONCEITOS FUNDAMENTAIS

Alguns conceitos são essenciais para a compreensão e modelagem das interações em redes sociais. Graças a eles, os pesquisadores e analistas podem descobrir padrões, identificar atores e relações relevantes, entender a disseminação de informações e muito mais. A leitura e interpretação de grafos em redes sociais (ou qualquer tipo de rede modelada como um grafo) suportam os *insights* obtidos da análise. Dentre os conceitos essenciais (3), (167), (168), destacam-se os seguintes:

- Nó (*Node* ou *Vertex*): também chamado por Vértice, os nós representam as entidades individuais dentro da rede. Dependendo do contexto, um nó pode representar uma pessoa, uma organização, uma palavra, um equipamento, entre outros tipos de entidades ou objeto que possam

ter algum tipo de relação. Em uma rede social de amizades, por exemplo, cada pessoa seria um nó. Não raro, os nós também podem ser referenciados como atores, a depender do contexto da rede.

- Ligação (*Edge* ou *Link*): podem ser referenciadas como arestas ou conexões, e representam as relações entre os nós. Em uma rede social, uma ligação poderia representar uma amizade, uma mensagem enviada, um *like* ou qualquer outra forma de interação entre atores.

- Tipo de grafo: É crucial reconhecer se o grafo é direcionado ou não direcionado. Em um grafo direcionado, as ligações têm uma direção, indicando uma relação unidirecional (por exemplo, "A segue B" ou "A vende produtos para B"). Em grafos não direcionados, as ligações não têm sentido direcional, indicando uma relação mútua (por exemplo, "A é amigo de B e vice-versa" ou "A vende para B e vice-versa"). A definição sobre o tipo de grafo é diretamente relacionada ao contexto da rede e das relações definidas entre os nós.

- Peso das ligações: algumas redes são ponderadas, o que significa que as ligações têm um valor ou peso associado a elas, podendo indicar a força ou intensidade da relação. Por exemplo, o número de mensagens trocadas entre duas pessoas pode ser usado como um peso. Ou ainda, os valores (ou quantidade) de transações comerciais entre duas empresas pode ser um ponderador, dando maior (ou menor) relevância para as diferentes ligações entre as empresas que compõem a rede.

- Caminho (*Path*): é a sequência de nós e ligações entre dois nós. O encadeamento de ligações entre nós representa caminhos em que ocorre o fluxo de transações ou troca de mensagens, por exemplo. Esse conceito é central para entender como a informação ou influência se propaga por meio de uma rede.

- Caminho geodésico: refere-se ao caminho mais curto entre dois nós em uma rede, medido pelo número de arestas que conectam esses nós. Esse conceito também é crucial para entender a distância e acessibilidade a diferentes partes da rede. Por exemplo, se estiver analisando uma rede de amizades e desejar encontrar o caminho mais curto entre duas pessoas, procura-se o número mínimo de conexões (amigos

em comum, amigos de amigos) necessárias para conectar essas duas pessoas, sendo esse o caminho geodésico entre elas.

- Centralidade: sustenta um conjunto de métricas que buscam identificar os nós mais "importantes" ou "influentes" em uma rede. Também denominadas com estatísticas da rede, tais medidas incluem de grau, de intermediação e de proximidade, que serão discutidas mais adiante.

- Homofilia: diz respeito à tendência dos nós de formar ligações com outros que são semelhantes a eles em algum aspecto. Por exemplo, em redes sociais, as pessoas muitas vezes formam amizades com outras que compartilham interesses, origens ou opiniões semelhantes. Em redes profissionais como o *LinkedIn*, indivíduos de campos profissionais semelhantes tendem a se conectar em uma rede mais densa. A homofilia opera em diferentes contextos e reconhecer tais padrões é crucial para entender a estrutura e dinâmica de uma rede, podendo ter implicações para tópicos como disseminação de informações, formação de opinião e dinâmicas de grupo (168), (169).

- Dinâmica temporal: algumas redes são dinâmicas, com nós e ligações sendo adicionados ou removidos ao longo do tempo. A temporalidade busca entender como as interações e os relacionamentos modificados com o passar do tempo afetam a estrutura e a função da rede. Por exemplo, em uma rede de colaboração científica, onde os nós representam pesquisadores e os laços representam coautoria em trabalhos acadêmicos, a temporalidade pode ser observada nas alterações das colaborações ao longo dos anos, revelando períodos de intensa colaboração interdisciplinar, novos rearranjos colaborativos ou o surgimento de novas áreas de pesquisa.

- Componente gigante: refere-se a um subgrafo dentro da rede maior que contém uma proporção significativa de todos os nós da rede. Em outras palavras, é o maior componente conectado em uma rede, no qual há um caminho entre cada par de nós, possibilitando a comunicação ou a transferência de informações, recursos ou qualquer outra entidade considerada na análise da rede entre todos os nós desse componente. O componente gigante é caracterizado por uma alta densidade de nós interconectados, o que significa que há uma rota ou caminho entre quase todos os pares de nós dentro desse componente. Além

disso, nós que fazem parte do componente gigante tendem a ter maior centralidade e, assim, maior influência ou importância dentro da rede. Eles possivelmente serão atores chave para a dinâmica da rede, como a disseminação de informações ou doenças.

- Rede ego: também se refere a uma sub-rede cujo foco principal está em um único nó, chamado de "ego", juntamente com os nós com os quais está diretamente conectado, denominados "alter". Os laços entre os "alters" também são incluídos, formando uma estrutura de rede centrada no nó "ego". Uma rede ego pode ser utilizada para analisar a influência e o suporte social que o nó "ego" recebe de seus "alters", fato essencial para entender dinâmicas em contextos como desbalanceamento de mercado por meio de monopólio (ou risco que ele oferece). Ou ainda, pode ser apropriada para estudar o capital social do nó "ego", ou seja, os recursos, informações ou suporte que o "ego" pode acessar primeiro ou preferencialmente por intermédio de seus "alters".

- Rede bipartida (ou bimodal): tipo especial de rede onde os nós têm significados diferentes, sendo divididos em dois conjuntos distintos, e as arestas (conexões) só podem existir entre nós de conjuntos diferentes. Não há arestas entre nós dentro do mesmo conjunto. Para exemplificar, uma rede bipartida tem nós "autores" e nós "publicações científicas", sendo que um conjunto de nós representa os autores e o outro conjunto representa as publicações. As arestas conectam os autores às suas respectivas publicações, mas não há conexões diretas entre autores ou entre publicações. Podendo assumir ainda o formato tripartido (três tipos de nós diferentes), esse tipo de arranjo permite a análise de relações ainda mais complexas. Elas proporcionam uma representação mais rica e detalhada dos atores e suas interações, sendo essenciais para explorar a estrutura e a dinâmica de sistemas complexos em contextos como ciência, comércio, gestão, mídias sociais, entre outros.

A compreensão de tais conceitos é quesito-chave para a utilização da ARS em pesquisas na área de Ciências Sociais Aplicadas, bem como para a correta interpretação dos resultados das métricas, que serão apresentadas a seguir.

6.4 MÉTRICAS EM ARS: DIFERENCIAL ANALÍTICO VALIOSO

Métricas, em sua essência, são medidas quantitativas que proporcionam uma compreensão objetiva e sistemática sobre determinados aspectos de um sistema ou fenômeno. Em diferentes campos de estudo, as métricas servem como ferramentas para avaliar, comparar e predizer comportamentos, performances ou tendências. Seja no mundo dos negócios, na ciência ou na engenharia, as métricas são fundamentais para direcionar decisões, validar hipóteses e monitorar progressos (170), (171).

No contexto da análise de redes, as mensurações assumem um papel singular. Enquanto em muitos outros campos as métricas podem focar em variáveis isoladas, como lucro, eficiência ou frequência, as métricas de análise de redes estão profundamente enraizadas na interconexão e relação entre entidades. Elas capturam a complexidade e a dinâmica das interações, fornecendo indícios mais claros sobre padrões de comunicação, influência e coesão dentro de uma rede. A seguir, apontam-se algumas métricas estatísticas da ARS comumente utilizadas nos estudos em Ciências Sociais Aplicadas (172), (174):

- Centralidade de grau: refere-se ao número de conexões diretas que um nó (por exemplo, um usuário ou uma entidade) tem dentro da rede. Em termos práticos, pode representar, por exemplo, o número de amigos de uma pessoa no *Facebook* ou seguidores no *X* (Twitter). Quanto maior a centralidade de grau, maior é a influência imediata de um nó sobre seus vizinhos na rede. Em outro exemplo: ao analisar uma rede social como o *Twitter*, um usuário com elevada centralidade de grau terá mais seguidores ou seguirá mais pessoas. Em outra situação, o grau pode destacar uma celebridade ou um influenciador digital que possui milhões de seguidores. A métrica de grau pode ainda ser dividida em duas variações: grau de entrada (número de arestas direcionadas que chegam a um nó) e grau de saída (número de arestas direcionadas que partem de um nó). Em uma rede de transações financeiras, um nó com alto grau de entrada pode indicar um grande receptor de pagamentos, enquanto um nó com alto grau de saída pode indicar um grande emissor de pagamentos. Essas métricas podem ajudar a identificar padrões normais e anormais de comportamento na rede, sendo ferramentas importantes em áreas como detecção de fraudes, por exemplo. (173), (175).

- Centralidade de intermediação: diz respeito à quantidade de vezes que um nó age como uma "ponte" ao longo do caminho mais curto entre dois outros nós. Um nó com alta intermediação tem uma influência considerável sobre a transferência de informações na rede, pois controla a passagem dessa informação entre diferentes partes da rede. Em contextos sociais, indivíduos com alta centralidade de intermediação podem atuar como mediadores ou corretores de informação. Por exemplo, em uma rede organizacional, um gerente que atua como um ponto de conexão entre diferentes departamentos terá uma alta centralidade de intermediação, pois controla o fluxo de informações entre esses departamentos (167), (173).

- Centralidade de proximidade: mede a média do caminho mais curto entre um nó e todos os outros nós na rede. Em outras palavras, representa a "distância média" de um nó a todos os outros. Nós com alta centralidade de proximidade conseguem acessar informações de toda a rede de forma mais rápida e eficaz. Por exemplo, em uma rede de colaboração científica, um pesquisador com alta centralidade de proximidade pode alcançar outros pesquisadores por meio de um menor número de intermediários, facilitando a disseminação de informações ou colaborações (167), (175).

- Diâmetro: maior distância geodésica entre quaisquer dois nós na rede. Em outras palavras, é o caminho mais longo dentro da rede. Essa métrica é útil para entender o alcance e a dispersão de uma rede. Exemplo: em redes sociais como o *Instagram*, *Facebook* ou *Tiktok*, o diâmetro pode representar a maior série de conexões de "amigos" que alguém teria que percorrer para conectar duas pessoas quaisquer na plataforma, oferecendo *insights* sobre a interconectividade global da rede (174), (176).

- Densidade: refere-se à proporção de conexões existentes em relação ao número total de conexões possíveis em uma rede. Por exemplo, uma rede onde todos estão conectados a todos teria uma densidade de 1, enquanto uma rede sem conexões teria densidade 0. Redes densas indicam uma maior interconexão entre os membros, o que pode facilitar a difusão de informações, mas também pode indicar falta de diversidade ou redundância nas conexões (1), (177). Para exemplificar, em um grupo de *WhatsApp* de familiares próximos, pode existir uma alta densidade, pois a maioria dos membros se conhece e interage

regularmente, enquanto em um fórum on-line aberto a densidade pode ser muito mais baixa.

- Coeficiente de agrupamento: avalia a tendência dos nós estarem agrupados ou formarem grupos (clusters). Em termos práticos, se dois amigos seus são, por sua vez, amigos entre si, existe um "agrupamento". Um alto coeficiente de agrupamento em redes sociais indica a presença de comunidades ou grupos fortemente interligados (1), (177). Em uma rede de amizades no *Facebook*, se João é amigo de Maria e de Paulo, e Maria também é amiga de Paulo, isso forma um triângulo, indicando um alto coeficiente de agrupamento entre eles, por exemplo.

- *Eigenvector* de centralidade: mede a influência de um nó na rede, levando em consideração não apenas quantas conexões ele tem, mas também quão influentes são os nós aos quais está conectado. Em outras palavras, um nó é considerado mais importante se estiver conectado a muitos nós que, por sua vez, são importantes. O *eigenvector* de centralidade fornece uma medida mais rica e informativa da importância ou influência de um nó em uma rede, considerando não só a quantidade, mas também a qualidade de suas conexões (178), (179). Uma boa prática é utilizá-la em complemento à análise da estatística de grau. Exemplo: em uma rede de citações acadêmicas, um artigo que é frequentemente citado por outros artigos altamente citados terá um alto *eigenvector* de centralidade, indicando sua influência na comunidade acadêmica.

O que distingue as métricas utilizadas em análise de redes sociais das métricas de outros campos é sua capacidade inerente de representar e quantificar relações (180). Enquanto uma métrica tradicional poderia simplesmente contar o número de usuários em uma plataforma, uma métrica de rede vai além, investigando como esses usuários estão conectados, quão influentes eles são e em que comunidades eles tendem a se agrupar (176).

Duas mensurações um pouco mais avançadas são importantes de serem destacadas: coesão de grupos e modelos de difusão. A coesão de grupos avalia a força das conexões dentro de grupos em uma rede, identificando cliques (subgrupos densamente conectados) e comunidades (grupos mais amplos e coesos) (181). Por exemplo, em uma rede social *on-line*, um clique pode representar um grupo de amigos próximos que interagem

frequentemente entre si, enquanto uma comunidade pode incluir vários cliques e indivíduos interconectados com interesses ou características comuns. A análise dessa métrica é essencial, por exemplo, para o desenvolvimento de estratégias de *marketing* personalizadas, na qual entender a coesão de grupos pode auxiliar na identificação de públicos-alvo mais homogêneos e na criação de campanhas mais eficazes e direcionadas (182).

No mundo dos negócios, a métrica de coesão de grupos é fundamental em redes de colaboração empresarial. Suponha uma corporação internacional com várias equipes e departamentos diferentes. A análise da coesão pode revelar subgrupos de colaboradores que interagem e colaboram intensivamente entre si, indicando um alto grau de coesão interna, e possivelmente, eficácia operacional. Se, dentro dessa corporação, existir um departamento de pesquisa e desenvolvimento (P&D) onde certos membros de diferentes equipes frequentemente colaboram e compartilham informações, formando um grupo coeso, isso pode ser indicativo de uma possível inovação ou desenvolvimento de produto emergente. Identificar e entender tais grupos coesos permite às empresas otimizar a comunicação interna, alocar recursos de maneira mais eficiente, fomentar a inovação, estrategicamente gerenciar o capital humano, promovendo sinergias e melhorando a performance organizacional.

Já os modelos de difusão em ARS são ferramentas analíticas que simulam e estudam o modo como informações, ideias, inovações ou comportamentos se propagam por intervenção da rede. Esses modelos podem ajudar a prever a velocidade e o alcance da disseminação de um determinado elemento nela (183). Eles são fundamentais para entender fenômenos como a viralização de informações nas redes sociais, a adoção de novas tecnologias, ou a propagação de comportamentos e atitudes em comunidades (184). Também consideram fatores como a estrutura da rede, os atributos dos nós (indivíduos ou entidades), a probabilidade de adoção, os mecanismos de influência e resistência. Como exemplo de utilização na área de *marketing* e publicidade, as empresas podem otimizar estratégias de *marketing* viral, identificando indivíduos-chave (influenciadores) que podem acelerar a adoção de produtos ou a disseminação de mensagens publicitárias. Ou na área de inovação, esses modelos ajudam organizações e pesquisadores a entender como novas tecnologias e práticas são adotadas por comunidades e organizações, informando estratégias de implementação e adoção (185).

Assim, utilizar métricas de ARS oferece várias vantagens. Primeiramente, elas permitem uma compreensão mais profunda da estrutura e do comportamento de sistemas complexos. Ao focar nas relações, essas métricas revelam padrões ocultos que poderiam ser negligenciados por análises mais tradicionais. Além disso, ao desvendar tais padrões, pode-se identificar pontos de influência ou vulnerabilidade dentro de uma rede, direcionar estratégias de *marketing*, ou até prever a disseminação de informações.

6.5 O CALCANHAR DE AQUILES NO USO DE DADOS

A coleta, tratamento e preparação de dados representam etapas cruciais no processo de pesquisa científica, desempenhando um papel indispensável na construção do conhecimento. Esses procedimentos são o alicerce sobre o qual as inferências científicas são construídas e, portanto, qualquer incoerência ou negligência nesses processos pode comprometer irremediavelmente a validade e a confiabilidade dos resultados da pesquisa (186), (187). No campo das ARS, a coleta, tratamento e preparação de dados assumem uma relevância adicional dada a complexidade inerente das estruturas de rede e a diversidade de contextos em que são aplicadas. As pesquisas em ARS frequentemente lidam com grandes volumes de dados, provenientes de diferentes fontes, como redes sociais on-line, redes de colaboração científica ou redes de relações interpessoais, cada uma apresentando seus próprios desafios e peculiaridades (188), (189). Explora-se a seguir cada uma destas três etapas.

A *coleta de dados* é o processo inicial, onde informações pertinentes são adquiridas para análise posterior. Essa fase necessita de rigor metodológico e precisão, com a seleção adequada de variáveis, amostras representativas, métodos de coleta que minimizem vieses e erros sistemáticos (185). Uma coleta de dados mal executada pode resultar em informações imprecisas ou irrelevantes, limitando a aplicabilidade e a generalização dos resultados da pesquisa. Na coleta de dados em ARS, a definição clara de quais são os nós e as arestas é fundamental, pois qualquer imprecisão nessa fase pode afetar todo o estudo. As fontes de dados são variadas, refletindo a amplitude de aplicações dessa metodologia (187), sendo que as mais comuns para pesquisas em ARS são mostradas no Quadro 2.

Quadro 2 - Fontes de dados comuns em ARS.

Fonte	Descrição
Redes sociais	Plataformas como *Facebook*, *Twitter*, *Instagram*, *LinkedIn* são fontes riquíssimas de dados sobre interações sociais. Muitos estudos em ARS exploram dados de redes sociais para analisar padrões de comunicação, difusão de informações, formação de comunidades.
Publicações científicas	Bases de dados bibliográficas como *PubMed*, *Scopus*, *Web of Science* fornecem dados sobre coautoria, citações, colaborações acadêmicas, permitindo a análise de redes de conhecimento e inovação científica. Soma-se a essas fontes as bases de patentes extraídas diretamente dos escritórios de cada país ou por meio de plataformas como *Derwent*, *Orbit*, *Google Patents*, *Lens*.org, entre outras.
Comunicações organizacionais	E-mails, mensagens instantâneas, outros registros de comunicação interna são utilizados para analisar redes de colaboração e fluxos de informação dentro de organizações.
Fóruns e comunidades on-line	Sites como *Reddit* e *Stack Overflow* são fontes de dados sobre interações e trocas de conhecimento em comunidades on-line, possibilitando o estudo de dinâmicas de grupos e a formação de normas e culturas comunitárias.
Dados governamentais e institucionais	Dados públicos de governos e instituições, como registros de votações legislativas, podem ser usados para analisar redes de alianças políticas e colaborações institucionais.

Fonte	Descrição
Redes de telecomunicações	Registros de chamadas telefônicas e mensagens de texto são empregados em estudos sobre padrões de comunicação e relações sociais em redes de telecomunicações.
Pesquisas e questionários	Dados coletados por meio de entrevistas, *surveys* e questionários podem ser utilizados para construir redes de relações e preferências em diversos contextos, como redes de amizade em escolas ou redes de confiança em comunidades.
Redes de cooperação e comércio	Dados de transações comerciais, acordos de cooperação, parcerias entre empresas e países são utilizados para analisar redes de comércio e cooperação internacional.
Bancos de dados de relacionamento entre entidades	Bases de dados que mapeiam relações entre entidades diversas, como empresas e seus fornecedores, também são fontes comuns em ARS.

Fonte: Fu; Luo; Boos (2019) (190).

Cada uma dessas fontes de dados apresenta seus próprios desafios e oportunidades. A escolha criteriosa da fonte de dados e uma compreensão profunda de suas características e limitações são, portanto, essenciais para o sucesso de estudos em ARS.

Posteriormente, o *tratamento de dados* envolve a limpeza e transformação dos dados coletados. Esse passo é crucial para assegurar a qualidade dos dados, identificando e corrigindo inconsistências, valores ausentes, duplicatas e *outliers*. Tal procedimento tem como objetivo refinar o conjunto de dados, eliminando ruídos e distorções que possam comprometer as análises subsequentes. Um tratamento de dados rigoroso permite, assim, que os pesquisadores realizem análises mais robustas e extraíam inferências mais precisas sobre seus objetos de estudo (186). Como exemplo, observa-se um caso de coletas de dados de um microblog como

X (Twitter). Na limpeza dos dados, verifica-se se há *tweets* repetidos no conjunto de dados, eles devem ser identificados e removidos (remoção de duplicatas); *tweets* sem conteúdo textual, apenas com imagens, podem precisar de tratamento especial ou exclusão, dependendo dos objetivos do estudo (tratamento de valores ausentes ou irrelevantes); dados incorretos, como datas de postagem inválidas, devem ser corrigidos ou removidos (correção de erros). Na transformação dos dados, pode-se transformar o texto dos *tweets*, extraindo características relevantes como a frequência de palavras ou a presença de *hashtags* específicas (extração de características); os dados dos usuários e suas interações podem ser transformados em uma representação de rede, com usuários como nós e *retweets*, menções ou respostas como arestas (estruturação da rede). De forma análoga, em estudos que fazem uso de dados de patentes ou de publicações científicos, por mais que tratem de dados já previamente apurados e acurados conforme validação do próprio escritório patentário ou das editorias, também necessitam de atenção com relação à limpeza e tratamento de dados (191), (192).

A *preparação de dados*, por sua vez, engloba a organização e a formatação dos dados para análise. Isso inclui a categorização, a conversão de tipos de dados, a criação de subconjuntos de dados específicos para diferentes análises. Essa etapa é vital para estruturar os dados de maneira que facilitem a aplicação de métodos analíticos e estatísticos, permitindo que os pesquisadores explorem eficientemente as relações, padrões, tendências existentes nos dados (189). Especificamente em ARS, a preparação de dados implica a conversão de dados brutos em uma representação de rede adequada, como matrizes de adjacência ou listas de arestas. A escolha da representação influencia diretamente a aplicabilidade dos métodos de análise de rede e a interpretação dos resultados (1). A categorização e a classificação de nós e arestas, baseadas em atributos relevantes, são também passos essenciais para enriquecer a análise. Cada ferramenta de ARS pode eventualmente suportar tipos e formatos diferentes de dados para representação de uma rede. Essencialmente, o arquivo de dados deverá explicitar os nós e as relações entre eles, podendo ser incluídos no arquivo-base dados sobre o tipo da rede (direcionada ou não-direcionada), peso de cada relação entre nós, tipo dos nós (no caso de uma rede bi ou tripartida), datas, entre outros dados que poderão ser utilizados na análise da rede (3). Mais adiante serão mostrados exemplos de formatação de um arquivo para importação em um *software* de ARS.

No que tange os desafios na coleta, tratamento e preparação dos dados, o pesquisador deve se atentar aos seguintes pontos:

- O acesso a dados relevantes pode ser restringido por questões de privacidade ou por políticas das plataformas e das bases de dados (193).
- A quantidade avassaladora e a diversidade de dados disponíveis podem ser desafiadoras, exigindo soluções robustas de armazenamento e processamento (194).
- Dados brutos podem conter erros, omissões, duplicatas, inconsistências que precisam ser identificados e corrigidos, o que demanda tempo e expertise na limpeza e tratamento dos dados (186).
- Converter dados brutos em um formato adequado para análise de rede, como listas de arestas ou matrizes de adjacência, pode ser complexo e suscetível a erros (1).
- A definição do que constitui um nó e uma aresta na rede deve ser rigorosa e adaptada ao contexto específico da pesquisa ou negócio (3).
- A escolha de métodos de análise e a definição de parâmetros como valores de peso de aresta podem influenciar significativamente os resultados (1).
- A reprodutibilidade dos estudos em ARS requer documentação meticulosa de todos os passos do processo de pesquisa, desde a coleta até a análise de dados (195).

Por último, cabe ressaltar nesta seção o aspecto da ética na manipulação de dados. Os pesquisadores devem manter a integridade dos dados, evitando qualquer manipulação que possa distorcer os resultados, garantindo a confidencialidade das informações sensíveis. A transparência e a replicabilidade também são valores centrais na pesquisa científica, incluindo ARS, exigindo que os procedimentos de coleta, tratamento, preparação de dados sejam claramente documentados e passíveis de verificação por outros pesquisadores (196).

6.6 INTERPRETAÇÃO DE REDES E SUAS IMPLICAÇÕES ANALÍTICAS

A interpretação de grafos é um processo fundamental na ARS, que possibilita o entendimento das estruturas e dinâmicas intrínsecas às redes. Esse processo é essencial para extrair significado dos dados e para traduzir os *insights* obtidos em conhecimento explícito (197).

O grafo é um reflexo visual e matemático das conexões e interações dentro da rede, sua análise profunda pode revelar padrões, tendências e anomalias, desvendando, assim, as complexidades das relações sociais (198). Interpretar um grafo envolve tanto análises quantitativas quanto qualitativas. A análise quantitativa foca em métricas e algoritmos para quantificar propriedades estruturais, como densidade, centralidade e modularidade. Por outro lado, a análise qualitativa procura entender os contextos, atributos, significados das conexões, aprofundando a compreensão dos fenômenos sociais representados no grafo. A integração das abordagens quantitativa e qualitativa é muitas vezes necessária para obter uma compreensão holística de uma rede (199). A triangulação de métodos e a convergência de evidências podem enriquecer a interpretação dos dados e aumentar a robustez das conclusões. Ao combinar as forças de ambas as abordagens, os pesquisadores podem penetrar mais profundamente nas relações sociais, elucidando os mecanismos e os processos que moldam a estrutura e a dinâmica das redes (200).

As implicações analíticas da interpretação de grafos são vastas e podem ser estrategicamente valiosas em diversos campos. Por exemplo, em marketing, a identificação de nós com alta centralidade pode ajudar a localizar influenciadores chave, otimizando estratégias de marketing de influência. Em epidemiologia, a análise de grafos pode revelar padrões de transmissão de doenças, informando estratégias de intervenção e controle de surtos (1).

A interpretação visual de uma rede em ARS envolve a observação de nós (os participantes da rede) e arestas (as relações entre os participantes), bem como a forma como estão dispostos e conectados (posicionamento) (175). Como exemplo, imagine uma rede social de uma comunidade acadêmica, onde nós representam pesquisadores e arestas representam colaborações em publicações conjuntas. Ao visualizar o grafo da rede, avalia-se que: (i) os agrupamentos densos de nós podem representar departamentos

ou grupos de pesquisa onde a colaboração é intensa (201); (ii) nós centrais com muitas conexões podem indicar pesquisadores influentes ou colaborativos, possivelmente aqueles com uma variedade de coautores e projetos interdisciplinares (202); (iii) nós periféricos ou isolados podem representar pesquisadores que colaboram menos ou que estão focados em áreas de pesquisa muito específicas (203). Ao notar nós com cores ou tamanhos diferenciados, é vital entender o que cada atributo representa (um nó maior pode significar um pesquisador com mais publicações). Já nós mais próximos podem indicar relações mais fortes ou frequentes. Observar ainda se há padrões evidentes ou agrupamentos que possam indicar comunidades fortemente interligadas (176). Esse tipo de análise é complementar às análises quantitativas e qualitativas, ajudando os pesquisadores a desenvolver uma compreensão mais profunda e intuitiva das dinâmicas de rede.

Um desafio fundamental na interpretação de grafos é garantir a precisão e a validade das análises. Isso implica em considerar limitações e vieses nos dados, bem como a adequação dos métodos de análise empregados. A complexidade das redes e a diversidade de suas manifestações demandam uma abordagem cuidadosa e multifacetada para evitar conclusões precipitadas ou errôneas.

6.7 ALGUMAS FERRAMENTAS DE ARS

Várias ferramentas computacionais podem ser utilizadas em ARS, cada uma com suas próprias características, vantagens e limitações. O Quadro 3 traz algumas das ferramentas mais comumente usadas para ARS. Ao escolher a ferramenta de *software* para ARS, é crucial ponderar as necessidades específicas do projeto, o *background* do usuário em análise de redes e programação, ou mesmo preferências pessoais. Algumas ferramentas são mais adequadas para visualização interativa e exploração de dados, enquanto outras são mais robustas para análises quantitativas e modelagem. Para usuários sem experiência técnica ou em programação, ferramentas com interfaces intuitivas e recursos visuais, como *Gephi* e *NodeXL*, podem ser mais adequadas. Para pesquisadores e analistas com experiência em programação e análise de dados, bibliotecas como igraph

e pacotes *R* podem oferecer a flexibilidade e o poder necessários para análises avançadas e modelagem de redes.

Quadro 3 – Ferramentas computacionais para ARS

Software	Breve descrição	Vantagens	Limitações
Cytoscape	Inicialmente desenvolvido para análise de redes biológicas, o Cytoscape agora é amplamente utilizado em diversas disciplinas. Ele permite a visualização, análise e modelagem de redes complexas e é extensível por meio de plugins (204), (205).	• Extensível por meio de plugins. • Adequado para análise de redes em várias disciplinas, incluindo biologia e ciências sociais.	• Pode ser complexo para usuários sem experiência técnica. • Algumas funcionalidades podem ser excessivamente especializadas para usuários fora da biologia.
Gephi	É uma das ferramentas mais populares e visualmente intuitivas para análise de redes. É aberta e gratuita, é especialmente útil para visualizar e explorar redes complexas, permitindo a manipulação interativa e a detecção de comunidade (206), (207).	• Interface intuitiva e user-friendly. • Excelente para visualização e exploração de redes, análises qualitativas e quantitativas. • Possui robustas capacidades de detecção de comunidade e análise de modularidade. • Curva de aprendizagem acelerada. • Software gratuito.	• Pode ser desafiador para usuários sem experiência em análise de rede. • Limitado em termos de funcionalidades de modelagem e simulação. • Pode enfrentar dificuldades de desempenho com redes muito grandes, tornando-se lento ou instável. • Falta de Suporte e Desenvolvimento

Software	Breve descrição	Vantagens	Limitações
Igraph	É uma biblioteca de software de código aberto para a criação, análise e visualização de redes. Está disponível em várias linguagens de programação, como Python, R e C++, tornando-a uma opção flexível para desenvolvedores e pesquisadores (208), (209).	• Biblioteca versátil disponível em várias linguagens de programação. • Ideal para desenvolvedores e pesquisadores que preferem trabalhar com código.	• Requer conhecimento em programação. • Menos intuitivo para usuários sem experiência técnica.
NetLogo	É uma plataforma de modelagem e simulação multiagente que permite a modelagem de sistemas complexos e fenômenos emergentes. É útil para estudar a dinâmica e a evolução de redes sociais ao longo do tempo (210), (211).	• Poderoso para simulações multiagente e modelagem de sistemas complexos. • Ideal para estudo da evolução e dinâmica de redes ao longo do tempo.	• Curva de aprendizado acentuada para usuários novos em programação e modelagem. • Pode ser excessivamente complexo para análises simples e visualizações.

Software	Breve descrição	Vantagens	Limitações
NetworkX	Biblioteca Python para a criação, análise, visualização de redes e grafos complexos. É uma ferramenta muito versátil e extensível que suporta vários tipos de redes, incluindo redes simples, redes direcionadas e grafos multigrafo (212), (214).	• Criação e análise de vários tipos de redes e grafos, oferecendo uma ampla gama de algoritmos de análise de redes. • Permite implementar facilmente seus próprios algoritmos e funções. • Comunidade de desenvolvimento ativa e uma ampla base de usuários, facilitando o acesso a suporte e recursos. • Sendo uma biblioteca Python, NetworkX se beneficia da flexibilidade, facilidade de uso e riqueza de recursos da linguagem.	• Requer conhecimento de Python, sendo uma barreira para usuários sem experiência em programação. • Capacidades de visualização relativamente básicas comparadas a softwares como Gephi. • Para redes muito grandes, há desafios de desempenho, sendo necessário o uso de bibliotecas e ferramentas adicionais para grandes conjuntos de dados.
NodeXL	É um plugin de código aberto para o Excel que permite a criação, visualização e análise de redes sociais. É uma ferramenta útil para usuários que preferem uma interface familiar e é especialmente útil para análises exploratórias rápidas (215), (216).	• Integração com o Excel torna-o acessível para usuários com experiência em planilhas. • Adequado para análises exploratórias rápidas e visualizações simples.	• Limitado em termos de funcionalidades avançadas de análise. • Dependente do Microsoft Excel, não sendo uma solução multiplataforma.

Software	Breve descrição	Vantagens	Limitações
Pajek	É um software de análise e visualização de grandes redes. É eficiente e oferece uma variedade de algoritmos para análise de redes grandes e complexas (217), (218).	• Eficiente para redes de grande escala. • Variedade de algoritmos para análise de redes complexas. • Opções de visualização claras e customizáveis. • Software gratuito	• Interface gráfica um pouco desatualizada e menos intuitiva. • Menos user-friendly para novos usuários em comparação com opções mais modernas. • Curva de aprendizado acentuada. • Documentação insuficiente ou desatualizada.
SNA em R	O ambiente estatístico R possui pacotes especializados, como o "SNA" (Social Network Analysis), oferecem funcionalidades robustas para análise de redes sociais, sendo uma opção poderosa para análises estatísticas e modelagem de redes (219).	• Poderoso para análises estatísticas e modelagem de redes. • Beneficia-se da extensa comunidade e recursos do R.	• Necessita de conhecimento em R e em programação estatística. • Pode ser intimidante para quem é novo em análise de dados e programação.
UCINet	Ferramenta abrangente para análise de redes sociais, amplamente utilizada para pesquisas acadêmicas em ciências sociais. UCINet é versátil e potente, oferecendo uma variedade de métodos analíticos para estudar estruturas e padrões de rede (220), (221).	• Ferramenta abrangente com uma variedade de métodos analíticos. • Adequado para análises avançadas e pesquisas acadêmicas.	• Interface pode não ser muito intuitiva para novos usuários. • Curva de aprendizado mais acentuada comparada a outras ferramentas.

Software	Breve descrição	Vantagens	Limitações
VoSViewer	Usada para construir e visualizar redes de informação, como redes de coautoria, de citação, de palavras-chave, especialmente em análises bibliométricas e cientométricas. Usado para explorar e analisar padrões das redes de pesquisa científica, proporcionando insights sobre as relações entre autores, instituições, países, temas de pesquisa (222), (223).	• Visualizações claras e intuitivas de redes • Robustez para análises bibliométricas e cientométricas, ideal para explorar tendências de pesquisa e padrões de colaboração. • Interface amigável e relativamente fácil de usar.	• Não é tão versátil quanto outras ferramentas de ARS em funcionalidades e tipos de análise. • Não é tão flexível ou extensível quanto soluções baseadas em código. • A importação e integração de dados exige etapas adicionais de preparação e formatação de dados.

Fonte: elaborado pelo autor (2023) baseado (221), (224).

6.8 MÃO NA MASSA - CASO DE USO

A escolha entre *softwares* de ARS geralmente depende das necessidades específicas da pesquisa e do projeto, do conjunto de habilidades do pesquisador e de preferências pessoais. *Gephi* e *UCINET* são frequentemente favorecidos por suas interfaces amigáveis e funcionalidades abrangentes, enquanto a *SNA* em *R* ou *NetworkX* são preferidas por aqueles confortáveis com programação em *R* ou *Python*. Assim, cada *software* tem seu conjunto único de características e capacidades, mas há pesquisadores que optam por usar uma combinação deles para atender às diversas necessidades de suas pesquisas em ciências sociais aplicadas. Nesta seção, optou-se por apresentar um caso elucidativo de uso do *Gephi*. Tal escolha se deve especialmente ao abrangente arsenal de funções e potencial de análises oferecidos pela ferramenta, associado à rápida curva de aprendizagem.

O *Gephi* é frequentemente citado como um dos *softwares* mais utilizados em pesquisas acadêmicas, principalmente devido à sua interface gráfica amigável, robustas capacidades de visualização e sua natureza

open-source, tornando-o acessível para a maioria dos pesquisadores (207). Pode ser obtido pela *URL*[76], na qual sugere-se que seja utilizada a versão mais estável e recente. O *site* "gephi.org" ainda possui *links* para acesso a documentações oficiais para estudos, destacado-se o guia rápido, tutorial de visualização e tutorial de *layouts*. Ainda é possível consultar alguns vídeos tutoriais que auxiliam no aprendizado da ferramenta.

Recomenda-se a atenção para que, antes da instalação do *Gephi*, seja instalado previamente o *Java* (JRE) na última versão. Depois de instalado o *JRE* e o *Gephi*, é possível alterar o idioma do *Gephi* e fazer diferentes configurações no ambiente de análise. Não raro ocorre algum problema como "*JVM Creation failed*" ou também erro de memória. Para solucionar, basta seguir as instruções em *Installing the software*[77].

Antes de iniciar o uso do *Gephi* propriamente dito, uma vez feita a coleta e tratamento de dados conforme explicado na seção anterior, deve-se preparar os arquivos para serem importados na ferramenta. No *Gephi*, dois arquivos são necessários para construção de uma rede: um arquivo contendo dados referente aos *Nós* e outro com dados sobre as *Arestas*. O arquivo de *Nós* deve conter, no mínimo, duas colunas: *ID* e *Label*, podendo o primeiro ser um campo numérico e o segundo o nome a ser dado ao nó. O *ID* pode ter o mesmo valor que o *Label*, desde que tenha um tamanho curto.

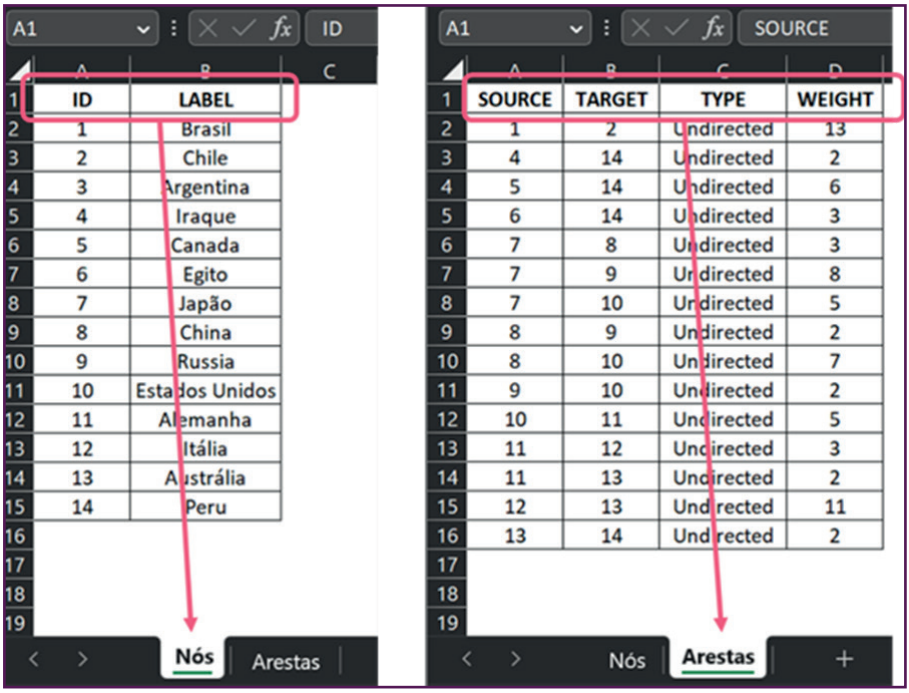
No arquivo de nós podem ainda ser incluídos campos adicionais que caracterizem cada nó e que o pesquisador julgue relevante para ser exibido na rede durante a análise dos resultados. Já o arquivo de *Arestas* deve ter minimamente os campos *Source, Target* e *Type*, sendo *Source* o nó de origem da relação, *Target* o nó de destino e *Type* o tipo da relação, podendo ter o valor "*Directed*" ou "*Undirected*". Assim, como no arquivo de nós, o arquivo de arestas pode conter campos adicionais, no qual destaca-se o campo "*Weight*" que dará a dimensão da intensidade da relação entre os nós. No *Gephi*, ambos os arquivos podem ser salvos separadamente no formato *.CSV* (comma separated value). Optando-se por preparar os dados utilizando um *software* de planilhas como *Excel*, *Google Planilhas* ou *LibreOffice*, deve-se salvar nós e arestas em abas separadas (Figura 1).

76 URL: https://gephi.org/users/download/

77 Instruções em: https://gephi.org/users/install/

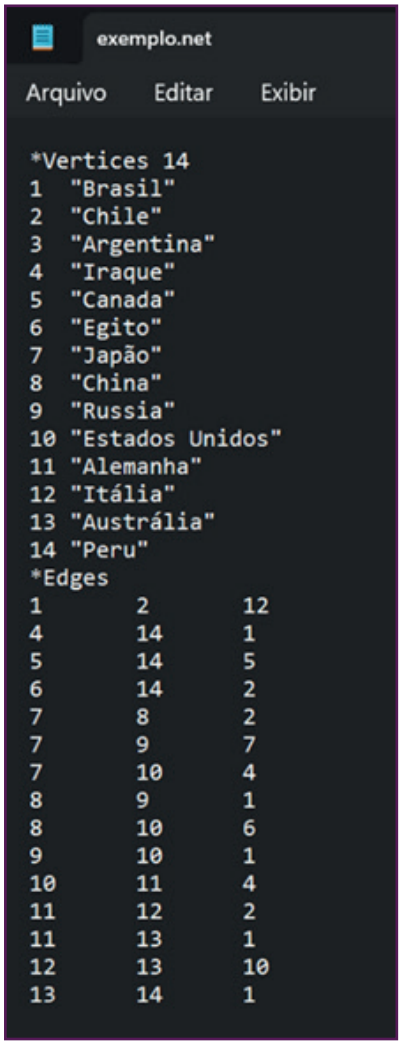
Há ainda o formato *.NET* muito comum para outros *softwares* e também suportado no *Gephi*. Nesse formato, os dados devem ser incluídos no arquivo de forma sequencial, mas explicitados quais são os nós e arestas (Figura 2). Esses não são os únicos tipos de arquivos suportados pelo *Gephi*, mas são os comumente utilizados.

Figura 1 – Formato dos dados dos Nós e Arestas em planilhas para importação no *Gephi*.



Fonte: elaborado pelo autor (2023).

Figura 2 - Formato do arquivo *.NET* para carga no *Gephi*.

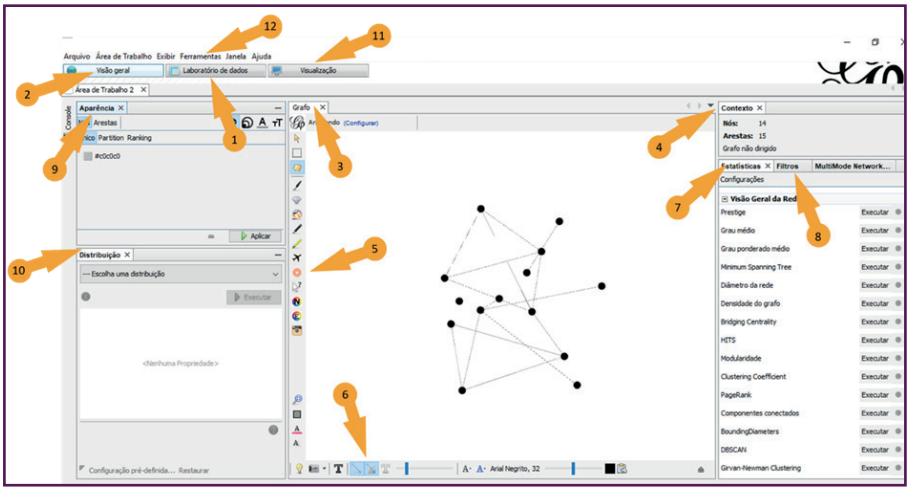


Fonte: elaborado pelo autor (2023).

Após o preparo dos arquivos de *Nós* e *Arestas*, eles devem ser importados para dentro da ferramenta. Para evitar erros ou eventuais duplicações de dados, primeiro deve-se importar para o *Gephi* o arquivo de *Nós* e, posteriormente, as *Arestas*. A lógica disso é simples pois, as relações não existem sem os nós. Logo, deve-se, primeiramente, "mostrar" para a ferramenta quais são os objetos ou indivíduos que constituem os nós para, depois, estabelecer as conexões entre eles por meio da importação das arestas.

Uma vez importados os dados no *Gephi*, está tudo pronto para a construção da rede. O *Gephi*, em sua essência, é uma ferramenta de análise e visualização de redes muito poderosa e pode parecer complicada no começo, mas ao se familiarizar com a interface e suas funcionalidades, se torna uma ferramenta muito útil. O ambiente de trabalho é composto por módulos com funções específicas que são apontadas na Figura 3.

Figura 3 – Exemplo da interface principal do *Gephi*



Fonte: elaborado pelo autor (2023).

Basicamente, ela consiste em três grandes áreas: Visão Geral, Laboratório de Dados e Visualização. Apesar da ferramenta mostrar inicialmente a área Visão Geral, quando vai se iniciar a construção de uma rede, é necessário primeiro ir para a área *Laboratório de Dados* (sinalizado como número 1 na Figura 3). Essa área, semelhante a tabelas, será onde os nós e as arestas serão inseridos, listando todas as entidades e suas relações, juntamente com seus atributos. Pode-se também, manualmente, adicionar, editar ou remover dados. No entanto, a opção mais comum é importar os dados de nós e arestas de um arquivo previamente preparado. Relevante ressaltar que esses dados mostrados no Laboratório de Dados também podem ser exportados para serem manuseados em outros *softwares*.

A área *Visão Geral* (item 2 na Figura 3) é a parte da interface que permite manipular a rede, aplicar *layouts* e filtrar os dados. É nessa área que se

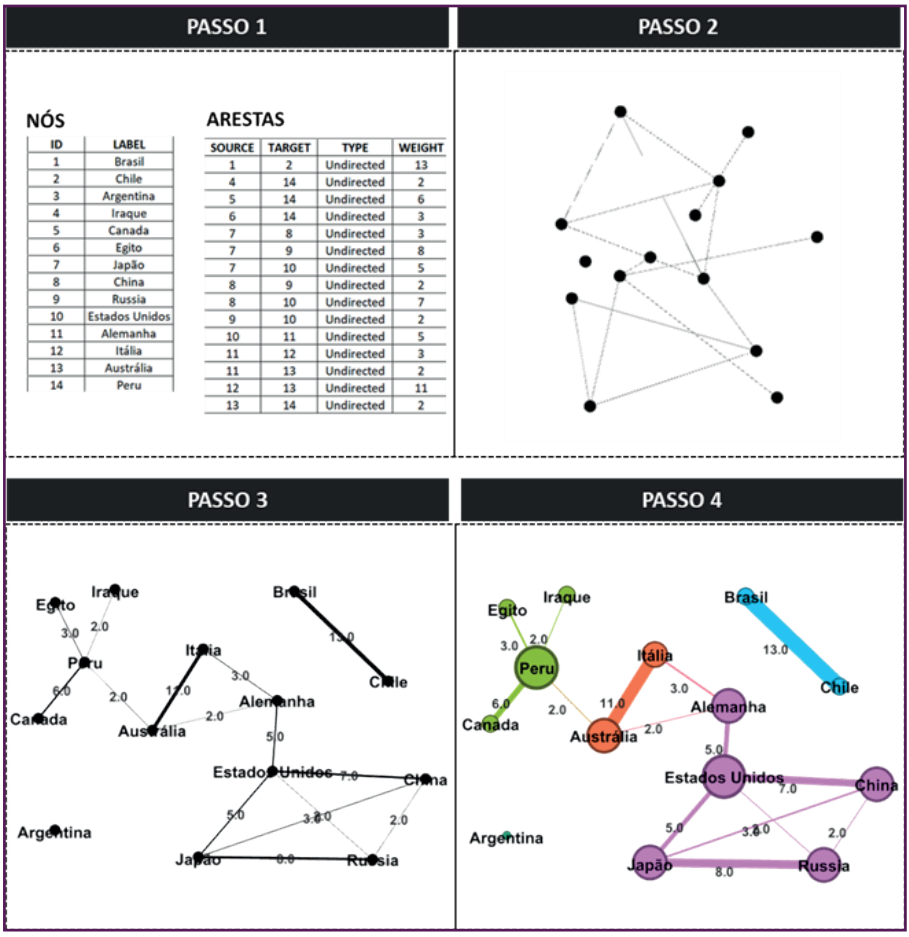
manuseia e se cria valor na rede por meio do *layout* e das estatísticas. Na aba Grafo (3), observa-se na parte central, uma rede ainda sem formato definitivo nem análise executada, caracterizando uma "rede crua" ou limpa. Na aba Contexto (4) é mostrado a quantidade de nós e arestas que constituem essa rede após a importação dos dados, bem como o apontamento se a rede é direcionada ou não. A barra de formatação vertical (5) mostra uma série de botões para personalização dos nós da rede. Já a barra de formatação horizontal (6) mostra opções para ajustes nas arestas, *print* da imagem do grafo, inclusão de rótulos nos nós, entre outras funcionalidades.

Na sequência, a aba Estatísticas (7) apresenta de forma nativa na ferramenta uma série de métricas, tais como grau, densidade, coeficiente de *clustering*, entre outros. Após executar as métricas, elas são adicionadas como atributos aos nós ou arestas e também podem ser usadas para ajustar a aparência da rede. A aba Filtros (8) permite focar em partes específicas da rede. Por exemplo, pode selecionar apenas nós com um certo valor de grau ou apenas parte da rede com base em outra métrica como proximidade. Na aba Aparência (9), define-se as cores, tamanhos e formas dos nós/arestas baseado em atributos e estatísticas geradas para a rede. Para exemplificar, pode-se colorir nós com base em uma métrica específica ou ajustar o tamanho de acordo com o grau. A aba chamada Distribuição (10) remete aos algoritmos de *layout* da rede. Os mais populares incluem "*Force Atlas* 2" (225), "*Fruchterman Reingold*" (226), mas há formatos bem específicos que podem ser explorados. Cada algoritmo tem seus próprios parâmetros que devem ser ajustados e que permitirão uma melhor análise visual da rede.

Após executado as métricas escolhidas, os ajustes nos nós, arestas e *layout*, é possível buscar uma formatação mais profunda por meio da área *Visualização* (11), na qual se pode exportar a figura da rede em formatos como *PNG* (Portable Network Graphics) ou *SVG* (Scalable Vector Graphics), que são úteis para apresentações ou publicações. Relevante apontar ainda que no menu suspenso *Ferramentas* (12) há uma opção para instalação de *Plugins* com implementações de funcionalidades, como novas métricas, *layouts*, importação/exportação especiais. Tais *plugins* são resultantes de colaborações disponibilizadas de forma gratuita pela comunidade de usuários do *Gephi*.

O exemplo implementado a seguir simula um caso hipotético referente a transações comerciais entre um grupo de países (Figura 4).

Figura 4 - Passo a passo para implementação da rede de transações comerciais entre países.



ID	LABEL
1	Brasil
2	Chile
3	Argentina
4	Iraque
5	Canada
6	Egito
7	Japão
8	China
9	Russia
10	Estados Unidos
11	Alemanha
12	Itália
13	Austrália
14	Peru

SOURCE	TARGET	TYPE	WEIGHT
1	2	Undirected	13
4	14	Undirected	2
5	14	Undirected	6
6	14	Undirected	3
7	8	Undirected	3
7	9	Undirected	8
7	10	Undirected	5
8	9	Undirected	2
8	10	Undirected	7
9	10	Undirected	2
10	11	Undirected	5
11	12	Undirected	3
11	13	Undirected	2
12	13	Undirected	11
13	14	Undirected	2

Fonte: elaborado pelo autor (2023).

O passo 1 mostra os arquivos de nós e arestas previamente preparados e que deve ser importado no *Gephi* como uma rede não direcionada. O passo 2 mostra o visual da rede "crua" entre os países ainda sem qualquer formatação ou manuseio da rede. O passo 3 já mostra uma distribuição visual entre os nós, sendo utilizado o *layout Force Atlas* 2. Também nesse

passo, por meio da barra horizontal na aba Grafo, foi marcada a opção de mostrar nomes dos nós (*label*) e os pesos das arestas (*weight*).

No passo 4, executou-se algumas estatísticas de nós tais como grau e densidade. Utilizou-se a estatística de centralidade de intermediação para formatar o tamanho de cada nó na rede. Ou seja, os tamanhos dos nós foram definidos proporcionalmente ao valor da métrica de centralidade de intermediação de cada um deles. Essa ação permite destacar os nós que são maiores controladores do fluxo de informações (nesse caso, os maiores intermediadores de transações comerciais entre os países da rede). Depois, ajustou-se a espessura das arestas para sobressaltar aquelas com maior peso, destacando as conexões mais intensas entre países, isso é, aquelas relações que expressam maior volume de transações comerciais. Por último, executou-se a estatística de Modularidade que utiliza o algoritmo de clusterização de *Blondel* (227) e, por meio da aba Aparência, fez-se a partição (aplicação de cores distintas) conforme os nós de cada *cluster*. Com isso, obteve-se 5 *clusters* de comércio entre países, em que se ressalta o cluster desconectado do restante da rede formado por Brasil e Chile, bem como a Argentina de forma isolada sem qualquer conexão com os demais.

Dado o contexto proposto no exemplo referente às relações comerciais entre os países, a ARS permite destacar os países mais relevantes com base na estatística de intermediação, mostrando os nós que têm maior poder de controle do fluxo de negócios entre países, no caso Alemanha, Austrália e Peru (Figura 5). Nota-se, ainda, a posição desses atores na rede, tornando-se verdadeiros *hubs* de negócios. Como observado no grafo da rede, o Brasil e o Chile são parceiros comerciais intensos, mas isolados do restante da rede, o que pode lhes trazer complicações devido à dependência total da parceria. A identificação dos *clusters* também possibilita analisar em maior profundidade as relações entre os países que compõem cada agrupamento.

Figura 5 – Resultados das estatísticas de rede no Laboratório de Dados

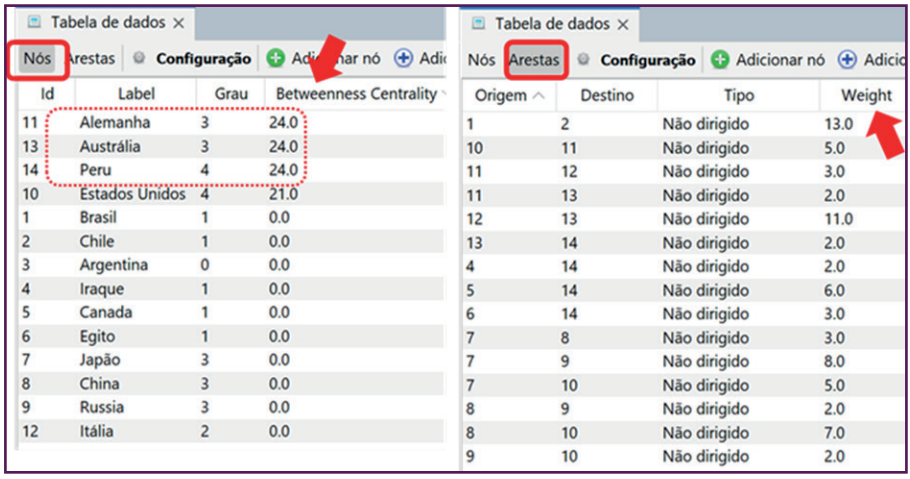


Tabela de dados

Nós | Arestas | Configuração | Adicionar nó

Id	Label	Grau	Betweenness Centrality
11	Alemanha	3	24.0
13	Austrália	3	24.0
14	Peru	4	24.0
10	Estados Unidos	4	21.0
1	Brasil	1	0.0
2	Chile	1	0.0
3	Argentina	0	0.0
4	Iraque	1	0.0
5	Canada	1	0.0
6	Egito	1	0.0
7	Japão	3	0.0
8	China	3	0.0
9	Russia	3	0.0
12	Itália	2	0.0

Tabela de dados

Nós | Arestas | Configuração | Adicionar nó | Adicic

Origem	Destino	Tipo	Weight
1	2	Não dirigido	13.0
10	11	Não dirigido	5.0
11	12	Não dirigido	3.0
11	13	Não dirigido	2.0
12	13	Não dirigido	11.0
13	14	Não dirigido	2.0
4	14	Não dirigido	2.0
5	14	Não dirigido	6.0
6	14	Não dirigido	3.0
7	8	Não dirigido	3.0
7	9	Não dirigido	8.0
7	10	Não dirigido	5.0
8	9	Não dirigido	2.0
8	10	Não dirigido	7.0
9	10	Não dirigido	2.0

Fonte: elaborado pelo autor (2023).

Apesar de ser um exemplo hipotético para fins de compreensão sobre a construção e interpretação de uma rede, nota-se o potencial de exploração de conhecimento que a ARS disponibiliza aos profissionais, permitindo ampliar as perspectivas analíticas e, ao mesmo tempo, conciliar com uma representação visual de fácil entendimento.

6.9 CONSIDERAÇÕES FINAIS

A Análise de Redes Sociais é uma abordagem interdisciplinar que surgiu de raízes sociológicas e matemáticas, evoluindo ao longo do tempo para se tornar uma ferramenta poderosa para compreender as complexas interações sociais em várias áreas do conhecimento. Seu papel na pesquisa continua a crescer à medida que novas tecnologias e plataformas de comunicação transformam os comportamentos sociais.

A ARS desempenha um papel fundamental nas Ciências Sociais Aplicadas, oferecendo uma ferramenta poderosa para compreender as interações sociais, identificar estruturas e hierarquias, fornecer *insights* valiosos para a tomada de decisões, formulação de políticas e resolução de problemas em

uma ampla variedade de contextos sociais e organizacionais. Ela continua a ser uma abordagem dinâmica e vital na pesquisa e na prática.

O uso consciente e criterioso das métricas de ARS, como grau, centralidade de proximidade, entre outras, é crucial para explorar resultados significativos sobre as estruturas de redes, contribuindo para uma variedade de campos, desde a pesquisa acadêmica até aplicações práticas como *marketing*, gestão da inovação e detecção de fraudes. Durante a exploração das potencialidades da ARS, foram discutidas diversas ferramentas e técnicas, cada uma com suas peculiaridades, vantagens e desafios. Fica evidente que a escolha da ferramenta adequada deve ser feita com base nas necessidades específicas do projeto e no nível de conhecimento do usuário em programação e sobre a teoria acerca de ARS. Ferramentas como *Gephi* e *NodeXL* podem ser mais acessíveis para aqueles menos experientes, enquanto bibliotecas como *iGraph*, *NetworkX* e pacotes *R* podem ser mais apropriados para análises avançadas.

O futuro da ARS é promissor, com potencial para novos desenvolvimentos e descobertas que beneficiarão a academia e a prática profissional. Especialmente para a área de Ciências Sociais Aplicadas, a ARS se destaca em meio a outras técnicas de pesquisa pela sua capacidade única de investigar padrões de relações, fluxos de informações e de influência, aspectos que outras metodologias não conseguem alcançar com o mesmo grau de profundidade e precisão. Assim, optar pela ARS em pesquisas na área de Ciências Sociais Aplicadas não é apenas uma escolha metodológica estratégica: é um compromisso com a busca pela compreensão autêntica e pela geração de conhecimento significativo, que tem o poder de informar, transformar e inovar. É imperativo que pesquisadores e profissionais considerem seriamente a incorporação da ARS como uma técnica central em seus estudos, dada a sua relevância e o valor inigualável que ela adiciona ao campo das Ciências Sociais Aplicadas.

REFERÊNCIAS

1 BORGATTI, S. P.; EVERETT, M. G. ; JOHNSON, J. C. **Analyzing Social Networks**. 2nd. ed. New York: Sage, 2018. 384 p.

2 O'MALLEY A. J.; MARSDEN, P. V. The Analysis of Social Networks. **Health Serv Outcomes Res Methodol**, [*s. l.*], v. 8, n. 4, p. 222, Dec. 2008. DOI: 10.1007/S10742-008-0041-Z. Disponível em: https://link.springer.com/article/10.1007/s10742-008-0041-z. Acesso em: 8 out. 2023.

3 WINSHIP, C.; WASSERMAN, S.; FAUST, K. Social Network Analysis: Methods and Applications. **J Am Stat Assoc**, [*s. l.*], v. 91, n. 435, p. 1373, Sept. 1996. DOI: 10.2307/2291756. Disponível em: https://www.jstor.org/stable/i314319. Acesso em: 8 out. 2023.

4 MORENO, J. L. **Who shall survive?** A new approach to the problem of human interrelations. Washington: Nervous and Mental Disease Publishing Co, 1934. DOI: 10.1037/10648-000. Disponível em: https://psycnet.apa.org/record/2005-00641-000. Acesso em: 8 out. 2023.

5 LEVY MORENO, J.; GIESSMANN, S. **Drawing the Social**: Jacob Levy Moreno, Sociometry, and the Rise of Network Diagrammatics. [*S. l.*], v. 2, 2017. DOI: 10.25969/MEDIAREP/3794. Disponível em: https://www.semanticscholar.org/paper/Drawing-the-Social%3A-Jacob-Levy--Moreno%2C-Sociometry%2C-Giessmann/1fe4f335171215cb57ffdfa95fd0cbc4d37ef104. Acesso em: 8 out. 2023.

6 GRANOVETTER, M. S. The Strength of Weak Ties. **American Journal of Sociology**, [*s. l.*], v. 78, n. 6, p. 1360–1380, May 1973. DOI: 10.1086/225469. Disponível: https://www.journals.uchicago.edu/doi/10.1086/225469. Acesso em: 8 out. 2023.

7 BORGATTI, S. P.; MEHRA, A.; BRASS, D. J.; LABIANCA, G. Network analysis in the social sciences. **Science**, [*s. l.*], v. 323, n. 5916, p. 892–895, Feb. 2009. DOI: 10.1126/SCIENCE.1165821. Disponível em: https://www.science.org/doi/10.1126/science.1165821. Acesso em: 8 out. 2023.

8 VAN DER HULST, R. C. Introduction to Social Network Analysis (SNA) as an investigative tool. **Trends Organ Crime**, [*s. l.*], v. 12, n. 2, p. 101–121, Dec.

2009. DOI: 10.1007/S12117-008-9057-6. Disponível em: https://link.springer.com/article/10.1007/s12117-008-9057-6. Acesso em: 8 out. 2023.

9 CUCE, G. Social network analysis in organization development studies. *In*: IEEE/ACM INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING. **Anais** [...]. [*S. l.*]: Institute of Electrical and Electronics Engineers (IEEE), Feb. 2013, p. 943–947. DOI: 10.1109/ASONAM.2012.240. Disponível em: https://ieeexplore.ieee.org/document/6425638. Acesso em: 8 out. 2023.

10 TTE, E.; ROUSSEAU, R. Social network analysis: a powerful strategy, also for the information sciences. **J Inf Sci**, [*s. l.*], v. 28, n. 6, p. 441–453, Dec. 2002. DOI: 10.1177/016555150202800601. Disponível em: https://journals.sagepub.com/doi/10.1177/016555150202800601. Acesso em: 8 out. 2023.

11 DU, N.; WU, B.; PEI, X.; WANG, B.; XU, L. Community detection in large-scale social networks. *In*: JOINT NINTH WEBKDD AND FIRST SNA-KDD 2007 WORKSHOP ON WEB MINING AND SOCIAL NETWORK ANALYSIS. **Anais** [...]. [*S. l.*: *s. n.*], 2007, p. 16–25. DOI: 10.1145/1348549.1348552. Disponível em: https://dl.acm.org/doi/10.1145/1348549.1348552. Acesso em: 8 out. 2023.

12 ZAEFARIAN, G.; MISRA, S.; KOVAL, M.; IURKOV, V. Editorial: Social network analysis in marketing: A step-by-step guide for researchers. **Industrial Marketing Management**, [*s. l.*], v. 107, p. A11–A24, Nov. 2022. DOI: 10.1016/J.INDMARMAN.2022.10.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0019850122002358. Acesso em: 8 out. 2023.

13 TSUGAWA, S. A Survey of Social Network Analysis Techniques and their Applications to Socially Aware Networking. **IEICE Transactions on Communications**, [*s. l.*], v. E102.B, n. 1, p. 17–39, Jan. 2019. DOI: 10.1587/TRANSCOM.2017EBI0003. Disponível em: https://www.jstage.jst.go.jp/article/transcom/advpub/0/advpub_2017EBI0003/_pdf. Acesso em: 8 out. 2023.

14 MAJEED, S.; UZAIR, M.; QAMAR, U.; FAROOQ, A. Social Network Analysis Visualization Tools: A Comparative Review. *In*: PROCEEDINGS - 2020 23RD IEEE INTERNATIONAL MULTI-TOPIC

CONFERENCE, INMIC 2020. **Anais** [...]. [*S. l.*: *s. n.*], 2020. DOI: 10.1109/INMIC50486.2020.9318162. Disponível em: https://ieeexplore.ieee.org/document/9318162. Acesso em: 8 out. 2023.

15 MOLITERNO, T. P.; MAHONY, D. M. Network theory of organization: A multilevel approach. **J Manage**, v. 37, n. 2, p. 443–467, Mar. 2011. DOI: 10.1177/0149206310371692. Disponível em: https://journals.sagepub.com/doi/10.1177/0149206310371692. Acesso em: 8 out. 2023.

16 BRASS, D. J.; BUTTERFIELD, K. D.; SKAGGS, B. C. Relationships and Unethical Behavior: A Social Network Perspective. **The Academy of Management Review**, [*s. l.*], v. 23, n. 1, p. 14, Jan. 1998. DOI: 10.2307/259097. Disponível em: https://www.jstor.org/stable/259097. Acesso em: 8 out. 2023.

17 FU, X.; LUO, J.-D.; BOOS, M. Methods for Interdisciplinary Social Network Studies. **Social Network Analysis**, [*s. l.*], Mar. 2017, p. 3–20. DOI: 10.1201/9781315369594-2. Disponível em: https://www.researchgate.net/publication/316315121_Chapter_1_Methods_for_Interdisciplinary_Social_Network_Studies_Interdisciplinary_Approaches_and_Case_Studies. Acesso em: 8 out. 2023.

18 ERFANMANESH, M. A.; MOROVATI ARDAKANI, M. A Scientometrics and Collaboration Network Analysis of the Quarterly Journal of Interdisciplinary Studies in the Humanities. **Interdisciplinary Studies in the Humanities**, [*s. l.*], v. 8, n. 4, p. 55–77, Sept. 2016. DOI: 10.22035/ISIH.2016.230. Disponível em: http://www.isih.ir/?_action=article&au=3708&_au=M.+A.++Erfanmanesh&lang=en. Acesso em: 8 out. 2023.

19 HAYASHI, M. C. P. I.; HAYASHI, C. R. M.; LIMA, M. Y. Análise de redes de co-autoria na produção científica em educação especial. **Liinc em Revista**, [*s. l.*], v. 4, n. 1, p. 84–103, 2008. DOI: 10.18617/liinc.v4i1.255. Disponível em: https://revista.ibict.br/liinc/article/view/3151. Acesso em: 8 out. 2023.

20 TELMO, F. A.; FEITOZA, R. A. B.; SILVA, A. K. A. Análise de redes sociais da produção científica em memória organizacional na Ciência da Informação. **Conhecimento em Ação**, [*s. l.*], v. 4, n. 1, p.

102–127, 2013. Disponível em: https://revistas.ufrj.br/index.php/rca/article/view/26126. Acesso em: 7 out. 2023.

21 SILVA, A. K. A. A dinâmica das redes sociais e as redes de coautoria. **Perspectivas em Gestão & Conhecimento**, [*s. l.*], v. 4, p. 27–47, 2014. Edição Especial. Disponível em: http://periodicos.ufpb.br/ojs2/index.php/pgc. Acesso em: 7 out. 2023.

22 LI, E. Y.; LIAO, C. H.; YEN, H. R. Co-authorship networks and research impact: A social capital perspective. **Res Policy**, [*s. l.*], v. 42, n. 9, p. 1515–1530, 2013. DOI: 10.1016/j.respol.2013.06.012. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0048733313001169. Acesso em: 9 out. 2023.

23 RACHERLA, P.; HU, C. A social network perspective of tourism research collaborations. **Ann Tour Res**, [*s. l.*], v. 37, n. 4, p. 1012–1034, Oct. 2010. DOI: 10.1016/j.annals.2010.03.008. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0160738310000411. Acesso em: 9 out. 2023.

24 ABBASI, A.; CHUNG, K. S. K.; HOSSAIN, L. Egocentric analysis of co-authorship network structure, position and performance. **Inf Process Manag**, [*s. l.*], v. 48, n. 4, p. 671–679, 2012. DOI: 10.1016/j.ipm.2011.09.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0306457311000975. Acesso em: 9 out. 2023.

25 LIU, X.; BOLLEN, J.; NELSON, M. L.; VAN DE SOMPEL, H. Co-authorship networks in the digital library research community. **Inf Process Manag**, [*s. l.*], v. 41, n. 6, p. 1462–1480, dez. 2005. DOI: 10.1016/j.ipm.2005.03.012. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0306457305000336. Acesso em: 9 out. 2023.

26 WAGNER, C. S.; LEYDESDORFF, L. Network structure, self-organization, and the growth of international collaboration in science. **Res Policy**, [*s. l.*], v. 34, p. 1608–1618, 2005. DOI: 10.1016/j.respol.2005.08.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0048733305001745. Acesso em: 9 out. 2023.

27 MARTIN, R.; MOODYSSON, J. Comparing knowledge bases: on the geography and organization of knowledge sourcing in the regional innovation system of Scania, Sweden. **European Urban and Regional Studies**, [*s. l.*], v. 20, n. 2, p. 170–187, Dec. 2011. DOI: 10.1177/0969776411427326. Disponível em: https://journals.sagepub.com/doi/10.1177/0969776411427326. Acesso em: 9 out. 2023.

28 HERMANS, F.; STUIVER, M.; BEERS, P. J.; KOK, K. The distribution of roles and functions for upscaling and outscaling innovations in agricultural innovation systems. **Agric Syst**, [*s. l.*], v. 115, p. 117–128, 2013. DOI: 10.1016/J.AGSY.2012.09.006. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0308521X12001436. Acesso em: 9 out. 2023.

29 SALAVISA, I.; SOUSA, C.; FONTES, M. Topologies of innovation networks in knowledge-intensive sectors: Sectoral differences in the access to knowledge and complementary assets through formal and informal ties. **Technovation**, [*s. l.*], v. 32, n. 6, p. 380–399, June 2012. DOI: 10.1016/J.TECHNOVATION.2012.02.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0166497212000296. Acesso em: 9 out. 2023.

30 HERMANS, F.; SARTAS, M.; VAN SCHAGEN, B.; VAN ASTEN, P.; SCHUT, M. Social network analysis of multi-stakeholder platforms in agricultural research for development: Opportunities and constraints for innovation and scaling. **PLoS One**, [*s. l.*], v. 12, n. 2, p. e0169634, Feb. 2017. DOI: 10.1371/JOURNAL.PONE.0169634. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0169634. Acesso em: 9 out. 2023.

31 HERMANS, F.; VAN APELDOORN, D.; STUIVER, M.; KOK, K. Niches and networks: Explaining network evolution through niche formation processes. **Res Policy**, [*s. l.*], v. 42, n. 3, p. 613–623, Apr. 2013. DOI: 10.1016/J.RESPOL.2012.10.004. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0048733312002314. Acesso em: 9 out. 2023.

32 VAN DER VALK, T.; CHAPPIN, M. M. H.; GIJSBERS, G. W. Evaluating innovation networks in emerging technologies. **Technol Forecast Soc Change**, [*s. l.*], v. 78, n. 1, p. 25–39, Jan. 2011. DOI: 10.1016/J.TECHFORE.2010.07.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0040162510001496. Acesso em: 9 out. 2023.

33 BARRIE, J.; ZAWDIE, G.; JOÃO, E. Assessing the role of triple helix system intermediaries in nurturing an industrial biotechnology innovation network. **J Clean Prod**, [*s. l.*], v. 214, p. 209–223, Mar. 2019. DOI: 10.1016/J.JCLEPRO.2018.12.287. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652618340186. Acesso em: 9 out. 2023.

34 SAMMARRA, A.; BIGGIERO, L. Heterogeneity and Specificity of Inter-Firm Knowledge Flows in Innovation Networks. **Journal of Management Studies**, [*s. l.*], v. 45, n. 4, p. 800–829, June 2008. DOI: 10.1111/J.1467-6486.2008.00770.X. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6486.2008.00770.x. Acesso em: 9 out. 2023.

35 WANG, J.; CAO, X.; ZHU, J.; MA, H. Evolution mechanism and empirical analysis of innovation network in advanced manufacturing industry. **Transinformação**, [*s. l.*], v. 34, p. e210038, July 2022. DOI: 10.1590/2318-0889202234E210038. Disponível em: https://www.scielo.br/j/tinf/a/H3zQxbY68C6zpc8nGLR9mZp/. Acesso em: 9 out. 2023.

36 DURÃO, I. L.; MEIRIÑO, M. J.; MÉXAS, M. P.; BRASIL. Inovação em serviços de saúde a partir do Teste Myers-Briggs Type Indicator (MBTI®) associado à análise de redes sociais (ARS). **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**, [*s. l.*], v. 12, n. 3, Sept. 2018. DOI: 10.29397/RECIIS.V12I3.1368. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1368. Acesso em: 9 out. 2023.

37 FRITSCH, M.; KAUFFELD-MONZ, M. The impact of network structure on knowledge transfer: An application of social network analysis in the context of regional innovation networks. **Annals of Regional Science**, [*s. l.*], v. 44, n. 1, p. 21–38, Jan. 2009. DOI: 10.1007/S00168-008-0245-8. Disponível em: https://link.springer.com/article/10.1007/s00168-008-0245-8. Acesso em: 9 out. 2023.

38 BRANDES, U.; KENIS, P.; RAAB, J.; SCHNEIDER, V.; WAGNER, D. Explorations into the Visualization of Policy Networks. **Journal of Theoretical Politics**, [*s. l.*], v. 11, n. 1, p. 75–106, Jan. 1999. DOI: 10.1177/0951692899011001004. Disponível em: https://journals.sagepub.com/doi/10.1177/0951692899011001004. Acesso em: 9 out. 2023.

39 RAMIA, G.; PATULNY, R.; MARSTON, G.; CASSELLS, K. The Relationship between Governance Networks and Social Networks: Progress, Problems and Prospects. **Political Studies Review**, [*s. l.*], v. 16, n. 4, p. 331–341, Oct. 2017. DOI: 10.1177/1478929917713952. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/1478929917713952?-journalCode=pswa. Aesso em: 9 out. 2023.

40 ZHANG, W.; ZHANG, M.; YUAN, L.; FAN, F. Social network analysis and public policy: what's new? **J Asian Public Policy**, [*s. l.*], v. 16, n. 2, p. 115–145, May. 2023. DOI: 10.1080/17516234.2021.1996869. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/17516234.2021.1996869. Acesso em: 9 out. 2023.

41 YIN, C.; HUANG, Z. The evolving policy network in sustainable transitions: The case of new energy vehicle niche in China. **J Clean Prod**, [*s. l.*], v. 411, p. 137299, July 2023. DOI: 10.1016/J.JCLEPRO.2023.137299. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652623014579. Acesso em: 9 out. 2023.

42 RESENDE, C. A. da S. Redes de interesses organizados no sistema comissional da Câmara dos Deputados. **Revista de Sociologia e Política**, [*s. l.*], v. 30, p. e011, jul. 2022. DOI: 10.1590/1678-98732230E011. Disponível em: https://scielo.br/j/rsocp/a/YkqZzZFgkrqF8M3yBDg66yr/?lang=pt. Acesso em: 9 out. 2023.

43 LEIFELD, P.; HAUNSS, S. Political discourse networks and the conflict over software patents in Europe. **Eur J Polit Res**, [*s. l.*], v. 51, n. 3, p. 382–409, May. 2012. DOI: 10.1111/J.1475-6765.2011.02003.X. Disponível em: https://ejpr.onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6765.2011.02003.x. Acesso em: 9 out. 2023.

44 LECY, J. D.; MERGEL, I. A.; SCHMITZ, H. P. Networks in Public Administration: Current scholarship in review. **Public Management Review**, [*s. l.*], v. 16, n. 5, p. 643–665, 2014. DOI: 10.1080/14719037.2012.743577. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/14719037.2012.743577. Acesso em: 9 out. 2023.

45 LEIFELD, P. Reconceptualizing Major Policy Change in the Advocacy Coalition Framework: A Discourse Network Analysis of German

Pension Politics. **Policy Studies Journal**, [*s. l.*], v. 41, n. 1, p. 169–198, Feb. 2013. DOI: 10.1111/PSJ.12007. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/psj.12007. Acesso em: 9 out. 2023.

46 SANDSTRÖM, A.; CARLSSON, L. The Performance of Policy Networks: The Relation between Network Structure and Network Performance. **Policy Studies Journal**, [*s. l.*], v. 36, n. 4, p. 497–524, Nov. 2008. DOI: 10.1111/J.1541-0072.2008.00281.X. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0072.2008.00281.x. Acesso em: 9 out. 2023.

47 LAU, A. K. W.; KAJIKAWA, Y.; SHARIF, N. The roles of supply network centralities in firm performance and the moderating effects of reputation and export-orientation. **Production Planning & Control**, [*s. l.*], v. 31, n. 13, p. 1110–1127, Oct. 2020. DOI: 10.1080/09537287.2019.1700569. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/09537287.2019.1700569. Acesso em: 9 out. 2023.

48 BURT, R. S.; OPPER, S.; ZOU, N. Social network and family business: Uncovering hybrid family firms. **Soc Networks**, [*s. l.*], v. 65, p. 141–156, May 2021. DOI: 10.1016/J.SOCNET.2020.12.005. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0378873320301040. Acesso em: 9 out. 2023.

49 FALCONE, E. C.; FUGATE, B. S.; DOBRZYKOWSKI, D. D. Supply chain plasticity during a global disruption: Effects of CEO and supply chain networks on operational repurposing. **Journal of Business Logistics**, [*s. l.*], v. 43, n. 1, p. 116–139, Mar. 2022. DOI: 10.1111/JBL.12291. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/jbl.12291. Acesso em: 9 out. 2023.

50 MARQUES, L.; MANZANARES, M. D. Towards social network metrics for supply network circularity. **International Journal of Operations and Production Management**, [*s. l.*], v. 43, n. 4, p. 595–618, Mar. 2023. DOI: 10.1108/IJOPM-02-2022-0139. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/IJOPM-02-2022-0139/full/html. Acesso em: 9 out. 2023.

51 ROMEIROA, P.; COSTAB, C. The potential of management networks in the innovation and competitiveness of rural tourism: a case study on the Valle del Jerte (Spain). **Current Issues in Tourism**, [*s. l.*], v. 13, n. 1, p. 75–91, 2010. DOI: 10.1080/13683500902730452. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/13683500902730452. Acesso em: 9 out. 2023.

52 MONAGHAN, S.; LAVELLE, J.; GUNNIGLE, P. Mapping networks: Exploring the utility of social network analysis in management research and practice. **J Bus Res**, [*s. l.*], v. 76, p. 136–144, July 2017. DOI: 10.1016/J.JBUSRES.2017.03.020. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S014829631730108X. Acesso em: 9 out. 2023.

53 GALASKIEWICZ, J. Studying supply chains from a social network perspective. **Journal of Supply Chain Management**, [*s. l.*], v. 47, n. 1, p. 4–8, Jan. 2011. DOI: 10.1111/J.1745-493X.2010.03209.X. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-493X.2010.03209.x. Acesso em: 9 out. 2023.

54 CHOI, T. Y.; KIM, Y. Structural embeddedness and supplier management: a network perspective. **Journal of Supply Chain Management**, [*s. l.*], v. 44, n. 4, p. 5–13, Sept. 2008. DOI: 10.1111/J.1745-493X.2008.00069.X. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/j.1745-493X.2008.00069.x. Acesso em: 9 out. 2023.

55 BELLAMY, M. A.; GHOSH, S.; HORA, M. The influence of supply network structure on firm innovation. **Journal of Operations Management**, [*s. l.*], v. 32, n. 6, p. 357–373, Sept. 2014. DOI: 10.1016/J.JOM.2014.06.004. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0272696314000485. Acesso em: 9 out. 2023.

56 KIM, Y.; CHOI, T. Y.; YAN, T.; DOOLEY, K. Structural investigation of supply networks: A social network analysis approach. **Journal of Operations Management**, [*s. l.*], v. 29, n. 3, p. 194–211, Mar. 2011. DOI: 10.1016/J.JOM.2010.11.001. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1016/j.jom.2010.11.001. Acesso em: 9 out. 2023.

57 GIULIANI, E. The selective nature of knowledge networks in clusters: evidence from the wine industry. **J Econ Geogr**, [*s. l.*], v. 7, n. 2, p.

139–168, Mar. 2007. DOI: 10.1093/JEG/LBL014. Disponível em: https://academic.oup.com/joeg/article-abstract/7/2/139/886855?redirectedFrom=fulltext. Acesso em: 9 out. 2023.

58 EDWARDS, G. Infectious Innovations? The Diffusion of Tactical Innovation in Social Movement Networks, the Case of Suffragette Militancy. **Soc Mov Stud**, [*s. l.*], v. 13, n. 1, p. 48–69, 2014. DOI: 10.1080/14742837.2013.834251. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/14742837.2013.834251. Acesso em: 9 out. 2023.

59 KRINSKY, J.; CROSSLEY, N. Social Movements and Social Networks: Introduction. **Soc Mov Stud**, [*s. l.*], v. 13, n. 1, p. 1–21, 2014. DOI: 10.1080/14742837.2013.862787. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/14742837.2013.862787. Acesso em: 9 out. 2023.

60 UYSAL, N.; YANG, A. The power of activist networks in the mass self-communication era: A triangulation study of the impact of WikiLeaks on the stock value of Bank of America. **Public Relat Rev**, [*s. l.*], v. 39, n. 5, p. 459-469, Dec. 2013. DOI: 10.1016/J.PUBREV.2013.09.007. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0363811113001318. Acesso em: 9 out. 2023.

61 GRUZD, A.; TSYGANOVA, K. Information Wars and Online Activism During the 2013/2014 Crisis in Ukraine: Examining the Social Structures of Pro- and Anti-Maidan Groups. **Policy Internet**, [*s. l.*], v. 7, n. 2, p. 121–158, June 2015. DOI: 10.1002/POI3.91. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.91. Acesso em: 9 out. 2023.

62 DI GREGORIO, M. Networking in environmental movement organisation coalitions: interest, values or discourse? **Env Polit**, [*s. l.*], v. 21, n. 1, p. 1–25, Feb. 2012. DOI: 10.1080/09644016.2011.643366. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/09644016.2011.643366. Acesso em: 9 out. 2023.

63 TROTT, V. Networked feminism: counterpublics and the intersectional issues of #MeToo. **Fem Media Stud**, [*s. l.*], v. 21, n. 7, p. 1125–1142, Oct. 2021. DOI: 10.1080/14680777.2020.1718176. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/14680777.2020.1718176. Acesso em: 9 out. 2023.

64 YANG, A.; UYSAL, N.; TAYLOR, M. Unleashing the Power of Networks: Shareholder Activism, Sustainable Development and Corporate Environmental Policy. **Bus Strategy Environ**, [*s. l.*], v. 27, n. 6, p. 712–727, Sept. 2018. DOI: 10.1002/BSE.2026. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/bse.2026. Acesso em: 9 out. 2023.

65 THEOCHARIS, Y. The Wealth of (Occupation) Networks? Communication Patterns and Information Distribution in a Twitter Protest Network. **Journal of Information Technology & Politics**, [*s. l.*], v. 10, n. 1, p. 35–56, Jan. 2013. DOI: 10.1080/19331681.2012.701106. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/19331681.2012.701106. Acesso em: 9 out. 2023.

66 BIDDIX, J. P.; PARK, H. W. Online networks of student protest: the case of the living wage campaign. **New Media & Society**, [*s. l.*], v. 10, n. 6, p. 871-891, Dec. 2008. DOI: 10.1177/1461444808096249. Disponível em: https://journals.sagepub.com/doi/10.1177/1461444808096249. Acesso em: 9 out. 2023.

67 KENNEY, M. *et al*. Organisational adaptation in an activist network: Social networks, leadership, and change in al-Muhajiroun. **Appl Ergon**, [*s. l.*], v. 44, n. 5, p. 739–747, Sept. 2013. DOI: 10.1016/J.APERGO.2012.05.005. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0003687012000683. Acesso em: 9 out. 2023.

68 KENNEY, M.; COULTHART, S.; WRIGHT, D. Structure and Performance in a Violent Extremist Network: The Small-world Solution. **Journal of Conflict Resolution**, [*s. l.*], v. 61, n. 10, p. 2208–2234, Nov. 2017. DOI: 10.1177/0022002716631104. Disponível em: https://journals.sagepub.com/doi/10.1177/0022002716631104. Acesso em: 9 out. 2023.

69 ROMEIRO, N. L.; DA SILVA, F. C. G.; DE ANDRADA SOBRAL BRISOLA, A. C. C. A página arrumando letras como um espaço para a desconstrução da dominação do patriarcado. **RDBCI**: Revista Digital de Biblioteconomia e Ciência da Informação, [*s. l.*], v. 16, n. 3, p. 317–337, jun. 2018. DOI: 10.20396/RDBCI.V16I3.8651276. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8651276. Acesso em: 9 out. 2023.

70 PEÑA-FERNÁNDEZ, S.; LARRONDO-URETA, A.; MORALES-I-GRAS, J. Feminism, identidad de género y polarización en TikTok y Twitter. **Oxbridge Publishing House**, [*s. l.*], v. 31, n. 75, p. 49–60, 2023. DOI: 10.3916/C75-2023-04. Disponível em: https://www.revistacomunicar.com/index.php?contenido=detalles&numero=75&articulo=75-2023-04&idioma=en. Acesso em: 9 out. 2023.

71 CHAN, R. Y. Y. *et al*. Facebook and information security education: What can we know from social network analyses on Hong Kong engineering students? *In*: PROCEEDINGS OF 2016 IEEE INTERNATIONAL CONFERENCE ON TEACHING, ASSESSMENT AND LEARNING FOR ENGINEERING, TALE 2016. **Proceedings** [...]. [*S. l.*]: IEEE, 2017. p. 303–307. DOI: 10.1109/TALE.2016.7851811. Disponível em: https://ieeexplore.ieee.org/document/7851811. Acesso em: 9 out. 2023.

72 LIN, H.; LI, S. Analysis of User Social Support Network in Online Tumor Community. **Data Inf Manag**, [*s. l.*], v. 5, n. 1, p. 184–194, Jan. 2021. DOI: 10.2478/DIM-2020-0040. Disponível em: https://www.semanticscholar.org/paper/Analysis-of-User-Social-Support-Network-in-Online-Lin-Li/b95139c89fa2b726894c8a0e1f60e2b8a861fd2d. Acesso em: 9 out. 2023.

73 RAMPONI, G. *et al*. Content-based characterization of online social communities. **Inf Process Manag**, [*s. l.*], v. 57, n. 6, p. 102133, Nov. 2020. DOI: 10.1016/J.IPM.2019.102133. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0306457319303516. Acesso em: 9 out. 2023.

74 JOKSIMOVI, S. *et al*. Exploring development of social capital in a CMOOC through language and discourse. **Internet High Educ**, [*s. l.*], v. 36, p. 54–64, Jan. 2018. DOI: 10.1016/J.IHEDUC.2017.09.004. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1096751617304554. Acesso em: 9 out. 2023.

75 GLOOR, P. *et al*. Put your money where your mouth is: Using deep learning to identify consumer tribes from word usage. **Int J Inf Manage**, [*s. l.*], v. 51, p. 101924, Apr. 2020. DOI: 10.1016/J.IJINFOMGT.2019.03.011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0268401218313057. Acesso em: 9 out. 2023.

76 NORMAN, H. *et al*. Exploring the Roles of Social Participation in Mobile Social Media Learning: A Social Network Analysis. **The International Review of Research in Open and Distributed Learning**, [*s. l.*], v. 16, n. 4, p. 205–224, Nov. 2015. DOI: 10.19173/IRRODL.V16I4.2124. Disponível em: https://www.irrodl.org/index.php/irrodl/article/view/2124. Acesso em: 9 out. 2023.

77 DUNN, A. G.; WESTBROOK, J. I. Interpreting social network metrics in healthcare organizations: A review and guide to validating small networks. **Soc Sci Med**, [*s. l.*], v. 72, n. 7, p. 1064–1068, Apr. 2011. DOI: 10.1016/J.SOCSCIMED.2011.01.029. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0277953611000724. Acesso em: 9 out. 2023.

78 BLIUC, A. M. *et al*. Building addiction recovery capital through online participation in a recovery community. **Soc Sci Med**, [*s. l.*], v. 193, p. 110–117, Nov. 2017. DOI: 10.1016/J.SOCSCIMED.2017.09.050. Disponível em: https://pubmed.ncbi.nlm.nih.gov/29032310/. Acesso em: 9 out. 2023.

79 MASCIA, D.; CICCHETTI, A. Physician social capital and the reported adoption of evidence-based medicine: Exploring the role of structural holes. **Soc Sci Med**, [*s. l.*], v. 72, n. 5, p. 798–805, Mar. 2011. DOI: 10.1016/J.SOCSCIMED.2010.12.011. Disponível em: https://pubmed.ncbi.nlm.nih.gov/21306807/. Acesso em: 9 out. 2023.

80 YUSUF, A. M.; SAPUTRO, M. R. G.; MAHARANI, W. Identifying Influencers on Twitter for Covid-19 Education and Vaccination Using Social Network Analysis. *In*: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND COMPUTER SYSTEMS AND 4TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND INFORMATION MANAGEMENT, ICSECS-ICOCSIM 2021. **Proceedings** [...]. [*S. l.*]: IEEE, 2021. p. 488–492. DOI: 10.1109/ICSECS52883.2021.00095. Disponível em: https://ieeexplore.ieee.org/document/9537011. Acesso em: 9 out. 2023.

81 GARCIA, E.; ELBELTAGI, I. M.; DUNGAY, K.; HARDAKER, G. Student use of Facebook for informal learning and peer support. **International Journal of Information and Learning Technology**, [*s. l.*], v. 32, n. 5, p. 286–299, Nov. 2015. DOI: 10.1108/IJILT-09-2015-0024.

Disponível em: https://www.emerald.com/insight/content/doi/10.1108/IJILT-09-2015-0024/full/html. Acesso em: 9 out. 2023.

82 DAWSON, S. "Seeing" the learning community: An exploration of the development of a resource for monitoring online student networking. **British Journal of Educational Technology**, [*s. l.*], v. 41, n. 5, p. 736–752, Sept. 2010. DOI: 10.1111/J.1467-8535.2009.00970.X. Disponível em: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8535.2009.00970.x. Acesso em: 9 out. 2023.

83 BRITO, S. R. *et al*. Employing online social networks to monitor and evaluate training of digital inclusion agents. **Social Network Analysis and Mining**, [*s. l.*], v. 3, n. 3, p. 497–519, Jan. 2013. DOI: 10.1007/S13278-012-0093-5/FIGURES/11. Disponível em: https://www.researchgate.net/publication/257801189_Employing_online_social_networks_to_monitor_and_evaluate_training_of_digital_inclusion_agents. Acesso em: 9 out. 2023.

84 SUN, Z.; THEUSSEN, A. Assessing negotiation skill and its development in an online collaborative simulation game: A social network analysis study. **British Journal of Educational Technology**, [*s. l.*], v. 54, n. 1, p. 222–246, Jan. 2023. DOI: 10.1111/BJET.13263. Disponível em: https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.13263. Acesso em: 9 out. 2023.

85 DU, H.; XING, W.; ZHU, G. Mining Teacher Informal Online Learning Networks: *Insights* From Massive Educational Chat Tweets. **Journal of Educational Computing Research**, [*s. l.*], v. 61, n. 1, p. 127–150, Mar. 2023. DOI: 10.1177/07356331221103764. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/07356331221103764. Acesso em: 9 out. 2023.

86 DADO, M.; BODEMER, D. A review of methodological applications of social network analysis in computer-supported collaborative learning. **Educational Research Review**, [*s. l.*], v. 22, p. 159–180, Nov. 2017. DOI: 10.1016/J.EDUREV.2017.08.005. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1747938X17300325. Acesso em: 9 out. 2023.

87 CHEN, H.; PARK, H. W.; Breazeal, C. Teaching and learning with children: Impact of reciprocal peer learning with a social robot

on children's learning and emotive engagement. **Computers & Education**, [*s. l.*], v. 150, p. 103836, Jun. 2020. DOI: 10.1016/J.COMPEDU.2020.103836. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0360131520300373. Acesso em: 9 out. 2023.

88 GALIKYAN, I.; Admiraal, W. Students' engagement in asynchronous online discussion: The relationship between cognitive presence, learner prominence, and academic performance. **Internet and Higher Education**, [*s. l.*], v. 43, p. 100692, Oct. 2019. DOI: 10.1016/J.IHEDUC.2019.100692. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1096751619304105. Acesso em: 9 out. 2023.

89 BARÃO, A.; VASCONCELOS, J. B.; ROCHA, Á.; PEREIRA, R. A knowledge management approach to capture organizational learning networks. **International Journal of Information Management**, [*s. l.*], v. 37, n. 6, p. 735–740, Dec. 2017. DOI: 10.1016/J.IJINFOMGT.2017.07.013. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0268401217306035. Acesso em: 9 out. 2023.

90 RIENTIES, B.; NOLAN, E. M. Understanding friendship and learning networks of international and host students using longitudinal Social Network Analysis. **International Journal of Intercultural Relations**, [*s. l.*], v. 41, p. 165–180, July 2014. DOI: 10.1016/J.IJINTREL.2013.12.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0147176713001417. Acesso em: 9 out. 2023.

91 MARTÍNEZ, A. *et al*. Combining qualitative evaluation and social network analysis for the study of classroom social interactions. **Comput Educ**, [*s. l.*], v. 41, n. 4, p. 353–368, Dec. 2003. DOI: 10.1016/J.COMPEDU.2003.06.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0360131503000824. Acesso em: 9 out. 2023.

92 DE LAAT, M. *et al*. Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. **Int J Comput Support Collab Learn**, [*s. l.*], v. 2, n. 1, p. 87–103, Mar. 2007. DOI: 10.1007/S11412-007-9006-4. Disponível em: https://www.semanticscholar.org/paper/Investigating-patterns-of-interaction-in-networked-Laat-Lally/5cdf6bdb216c4dabbc3f56823240a3685e6f57a3. Acesso em: 9 out. 2023.

93 CHO, H. *et al*. Social networks, communication styles, and learning performance in a CSCL community. **Comput Educ**, [*s. l.*], v. 49, n. 2, p. 309–329, Sept. 2007. DOI: 10.1016/J.COMPEDU.2005.07.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0360131505001272. Acesso em: 9 out. 2023.

94 MOORE, S. *et al*. International NGOs and the Role of Network Centrality in Humanitarian Aid Operations: A Case Study of Coordination During the 2000 Mozambique Floods. **Disasters**, [*s. l.*], v. 27, n. 4, p. 305–318, Dec. 2003. DOI: 10.1111/J.0361-3666.2003.00235.X. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0361-3666.2003.00235.x. Acesso em: 9 out. 2023.

95 MIZRUCHI, M. S. Social Network Analysis: Recent Achievements and Current Controversies. **Acta Sociologica**, [*s. l.*], v. 37, n. 4, p. 329–343, Oct. 1994. DOI: 10.1177/000169939403700403. Disponível em: https://journals.sagepub.com/doi/10.1177/000169939403700403. Acesso em: 9 out. 2023.

96 ALEXANDER, S. M.; ARMITAGE, D.; CHARLES, A. Social networks and transitions to co-management in Jamaican marine reserves and small-scale fisheries. **Global Environmental Change**, [*s. l.*], v. 35, p. 213–225, Nov. 2015. DOI: 10.1016/J.GLOENVCHA.2015.09.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959378015300376. Acesso em: 9 out. 2023.

97 HU, Q.; KNOX, C. C.; KAPUCU, N. What Have We Learned since September 11, 2001? A Network Study of the Boston Marathon Bombings Response. **Public Adm Rev**, [*s. l.*], v. 74, n. 6, p. 698–712, Nov. 2014. DOI: 10.1111/PUAR.12284. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/puar.12284. Acesso em: 9 out. 2023.

98 BARBER, M. J.; FISCHER, M. M.; SCHERNGELL, T. The Community Structure of Research and Development Cooperation in Europe: Evidence from a Social Network Perspective. **Geogr Anal**, [*s. l.*], v. 43, n. 4, p. 415–432, Oct. 2011. DOI: 10.1111/J.1538-4632.2011.00830.X. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.2011.00830.x. Acesso em: 9 out. 2023.

99 O'CONNOR, A.; SHUMATE, M. Differences Among NGOs in the Business-NGO Cooperative Network. **Bus Soc**, [*s. l.*], v. 53, n. 1, p. 105–133, Jan. 2014. DOI: 10.1177/0007650311418195. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/0007650311418195?journalCode=basa. Acesso em: 9 out. 2023.

100 CASTRO, I.; GALÁN, J. L.; CASANUEVA, C. Antecedents of construction project coalitions: a study of the Spanish construction industry. **Construction Management and Economics**, [*s. l.*], v. 27, n. 9, p. 809–822, 2009. DOI: 10.1080/01446190903117751. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/01446190903117751. Acesso em: 9 out. 2023.

101 PACHOUD, C. *et al*. A Relational Approach to Studying Collective Action in Dairy Cooperatives Producing Mountain Cheeses in the Alps: The Case of the Primiero Cooperative in the Eastern Italians Alps. **Sustainability 2020**, [*s. l.*], v. 12, p. 4596, v. 12, n. 11, p. 4596, June 2020. DOI: 10.3390/SU12114596. Disponível em: https://www.mdpi.com/2071-1050/12/11/4596. Acesso em: 9 out. 2023.

102 AHMED, T. K.; CHAN, K. L. G.; MUTALIB, M. H. A. Social Networks as Social Capital for Volunteering With Syrian Refugees in Slemani City, Kurdistan Region, Iraq. **Sage Open**, [*s. l.*], v. 13, n. 1, Jan. 2023. DOI: 10.1177/21582440231155850. Disponível em: https://journals.sagepub.com/doi/full/10.1177/21582440231155850. Acesso em: 9 out. 2023.

103 KASPER, E. Seeing change in urban informal settlements with social network analysis. **Environ Urban**, [*s. l.*], v. 33, n. 1, p. 151–172, Apr. 2021. DOI: 10.1177/0956247820953757. Disponível em: https://journals.sagepub.com/doi/full/10.1177/0956247820953757/. Acesso em: 9 out. 2023.

104 KOLLECK, N. The power of third sector organizations in public education. **Journal of Educational Administration**, [*s. l.*], v. 57, n. 4, p. 411–425, June 2019. DOI: 10.1108/JEA-08-2018-0142. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JEA-08-2018-0142/full/html. Acesso em: 9 out. 2023.

105 ENGBERG, J.; MAIER, C. D. The dynamics of knowledge and expertise in social media interactions: Knowledge types, processes of

co-constructing knowledge and discursive reactions. *In*: ENGBERG, Jan; BUTLER-FAGE, Antoinette; KASTBERG, Peter. **Perspectives on Knowledge Communication**: Concepts and Settings, [*s. l.*], p. 57–76, Jan. 2023. DOI: 10.4324/9781003285120-4. Disponível em: https://books.google.com.br/books?id=sf_KEAAAQBAJ&pg=PT21&lpg=PT21&dq=10.4324/9781003285120-4.&source=bl&ots=REImx62FHY&sig=ACfU3U3Y7cdXY96D7If38Y2cq_U2t3SSKA&hl=pt-BR&sa=X&ved=2ahUKEwi-4NuCn-qBAxXOjZUCHW5VCJUQ6AF6BAgJEAM#v=onepage&q=10.4324%2F9781003285120-4.&f=false. Acesso em: 9 out. 2023.

106 TORQUATI, B. *et al*. Participatory Guarantee System and Social Capital for Sustainable Development in Brazil: The Case Study of OPAC Orgânicos Sul de Minas. **Sustainability 2021**, [*s. l.*], v. 13, n. 20, p. 11555, Oct. 2021. DOI: 10.3390/SU132011555. Disponível em: https://www.mdpi.com/2071-1050/13/20/11555. Acesso em: 9 out. 2023.

107 SAMPAIO, G. C.; MARINI, M. J.; SANTOS, G. D. Capital Social e Ações Conjuntas: um estudo de caso no Arranjo Produtivo de vinhos de altitude catarinense. **Revista de Economia e Sociologia Rural**, [*s. l.*], v. 56, n. 4, p. 605–622, out. 2018. DOI: 10.1590/1234-56781806-94790560404. Disponível em: https://www.scielo.br/j/resr/a/8RSTHLcz-3395ZpRpczcwJMq/?lang=pt. Acesso em: 9 out. 2023.

108 BASTOS, A. V. B.; SANTOS, M. V. Redes sociais informais e compartilhamento de significados sobre mudança organizacional. Revista de Administração de Empresas, [*s. l.*], v. 47, n. 3, p. 27–39, 2007. DOI: 10.1590/S0034-75902007000300003. Disponível em: https://www.scielo.br/j/rae/a/Z47wmzKygNMbCy9FV8ZYWcL/. Acesso em: 9 out. 2023.

109 ROSSONI, L.; ARANHA, C. E.; MENDES-DA-SILVA, W. Does the capital of social capital matter? Relational resources of the board and the performance of Brazilian companies. **Journal of Management and Governance**, [*s. l.*], v. 22, n. 1, p. 153–185, Mar. 2018. DOI: 10.1007/S10997-017-9382-8/. Disponível em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/25508/2-s2.0-85019687527.pdf. Acesso em: 9 out. 2023.

110 ZHANG, X. *et al*. Does regional cooperation constrain urban sprawl? Evidence from the Guangdong-Hong Kong-Macao Greater Bay Area. **Landsc Urban Plan**, [*s. l.*], v. 235, p. 104742, July 2023. DOI:

10.1016/J.LANDURBPLAN.2023.104742. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0169204623000610. Acesso em: 9 out. 2023.

111 BERITELLI, P. Cooperation among prominent actors in a tourist destination. **Ann Tour Res**, [*s. l.*], v. 38, n. 2, p. 607–629, Apr. 2011. DOI: 10.1016/J.ANNALS.2010.11.015. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S016073831000157X. Acesso em: 9 out. 2023.

112 LI, E. Y.; LIAO, C. H.; YEN, H. R. Co-authorship networks and research impact: A social capital perspective. **Res Policy**, [*s. l.*], v. 42, n. 9, p. 1515–1530, Nov. 2013. DOI: 10.1016/J.RESPOL.2013.06.012. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0048733313001169. Acesso em: 9 out. 2023.

113 ZAHEER, A.; GÖZÜBÜYÜK, R.; MILANOV, H. It's the connections: The network perspective in interorganizational research. **Academy of Management Perspectives**, [*s. l.*], v. 24, n. 1, p. 62–77, 2010. DOI: 10.5465/AMP.2010.50304417. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2233981. Acesso em: 9 out. 2023.

114 ROBERT, L. P.; DENNIS, A. R.; AHUJA, M. K. Social Capital and Knowledge Integration in Digitally Enabled Teams. **Information Systems Research**, [*s. l.*], v. 19, n. 3, p. 314–334, Sept. 2008. DOI: 10.1287/ISRE.1080.0177. Disponível em: https://pubsonline.informs.org/doi/abs/10.1287/isre.1080.0177. Acesso em: 9 out. 2023.

115 PENG, Y. Kinship Networks and Entrepreneurs in China's Transitional Economy. **American Journal of Sociology**, [*s. l.*], v. 109, n. 5, p. 1045–1074, Mar. 2004. DOI: 10.1086/382347. Disponível em: https://www.journals.uchicago.edu/doi/abs/10.1086/382347. Acesso em: 9 out. 2023.

116 WAGNER, C. S.; LEYDESDORFF, L. Network structure, self-organization, and the growth of international collaboration in science. **Res Policy**, [*s. l.*], v. 34, n. 10, p. 1608–1618, Dec. 2005. DOI: 10.1016/J.RESPOL.2005.08.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0048733305001745. Acesso em: 9 out. 2023.

117 REAGANS, R.; ZUCKERMAN, E. W. Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams. **Organization Science**, [*s. l.*], v. 12, n. 4, p. 502–517, Aug. 2001. DOI: 10.1287/ORSC.12.4.502.10637. Disponível em: https://pubsonline.informs.org/doi/10.1287/orsc.12.4.502.10637. Acesso em: 9 out. 2023.

118 MARTINUS, K. Labor Networks Connecting Peripheral Economies to the National Innovation System. **Ann Am Assoc Geogr,** [*s. l.*], v. 108, n. 3, p. 845–863, May 2018. DOI: 10.1080/24694452.2017.1374163. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/24694452.2017.1374163. Acesso em: 9 out. 2023.

119 DE PAULO, A. F. *et al*. Emerging green technologies for vehicle propulsion systems. **Technol Forecast Soc Change**, [*s. l.*], v. 159, p. 120054, Oct. 2020. DOI: 10.1016/J.TECHFORE.2020.120054. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0040162519302288. Acesso em: 9 out. 2023.

120 LINARES, I. M. P.; DE PAULO, A. F.; PORTO, G. S. Patent-based network analysis to understand technological innovation pathways and trends. **Technol Soc**, [*s. l.*], v. 59, p. 101134, Nov. 2019. DOI: 10.1016/J.TECHSOC.2019.04.010. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0160791X18301891. Acesso em: 9 out. 2023.

121 PEREIRA, C. G. *et al*. Patent mining and landscaping of emerging recombinant factor VIII through network analysis. **Nature Publishing Group**, [*s. l.*], v. 36, n. 7, 2018. DOI: 10.1038/nbt.4178. Disponível em: https://www.nature.com/articles/nbt.4178. Acesso em: 9 out. 2023.

122 LAI, K. K. *et al*. Mapping technological trajectories and exploring knowledge sources: A case study of E-payment technologies. **Technol Forecast Soc Change**, [*s. l.*], v. 186, p. 122173, Jan. 2023. DOI: 10.1016/J.TECHFORE.2022.122173. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0040162522006941. Acesso em: 9 out. 2023.

123 JIANG, J.; ZHAO, Y. Technology Trend Analysis of Japanese Green Vehicle Powertrains Technology Using Patent Citation Data. **Energies (Basel)**, [*s. l.*], v. 16, n. 5, p. 2221, Mar. 2023. DOI: 10.3390/

EN16052221/S1. Disponível em: https://www.mdpi.com/1996-1073/16/5/2221. Acesso em: 9 out. 2023.

124 HE, J.; WANG, Y. Patent-Based Analysis of China's Emergency Logistics Industry Convergence. **Sustainability 2023**, [*s. l.*], v. 15, n. 5, p. 4419, Mar. 2023. DOI: 10.3390/SU15054419. Disponível em: https://www.mdpi.com/2071-1050/15/5/4419. Acesso em: 9 out. 2023.

125 ESHOV, M. *et al*. Effective use of block chain technology in business process (in case of Uzbekistan). **E3S Web of Conferences**, [*s. l.*], v. 401, p. 05069, July 2023. DOI: 10.1051/E3SCONF/202340105069. Disponível em: https://www.e3s-conferences.org/articles/e3sconf/pdf/2023/38/e3sconf_conmechydro23_05069.pdf. Acesso em: 9 out. 2023.

126 HSU, C. W.; LIN, C. Y. Using social network analysis to examine the technological evolution of fermentative hydrogen production from biomass. **Int J Hydrogen Energy**, [*s. l.*], v. 41, n. 46, p. 21573–21582, Dec. 2016. DOI: 10.1016/J.IJHYDENE.2016.07.157. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0360319915319777. Acesso em: 9 out. 2023.

127 MA, S. C.; XU, J. H.; FAN, Y. Characteristics and key trends of global electric vehicle technology development: A multi-method patent analysis. **J Clean Prod**, [*s. l.*], v. 338, p. 130502, Mar. 2022. DOI: 10.1016/J.JCLEPRO.2022.130502. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652622001457. Acesso em: 9 out. 2023.

128 DE PAULO, A. F.; PORTO, G. S. Evolution of collaborative networks of solar energy applied technologies. **J Clean Prod**, [*s. l.*], v. 204, p. 310–320, Dec. 2018. DOI: 10.1016/J.JCLEPRO.2018.08.344. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652618327057. Acesso em: 9 out. 2023.

129 FAN, J.; XIAO, Z. Analysis of spatial correlation network of China's green innovation. **J Clean Prod**, [*s. l.*], v. 299, p. 126815, May 2021. DOI: 10.1016/J.JCLEPRO.2021.126815. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652621010349. Acesso em: 9 out. 2023.

130 YOON, J.; CHOI, S.; KIM, K. Invention property-function network analysis of patents: A case of silicon-based thin film solar cells. **Scientometrics**, [*s. l.*], v. 86, n. 3, p. 687–703, Mar. 2011. DOI: 10.1007/S11192-010-0303-8. Disponível em: https://link.springer.com/article/10.1007/s11192-010-0303-8. Acesso em: 9 out. 2023.

131 FISCHER, B. B.; SCHAEFFER, P. R.; VONORTAS, N. S. Evolution of university-industry collaboration in Brazil from a technology upgrading perspective. **Technol Forecast Soc Change**, [*s. l.*], v. 145, p. 330–340, Aug. 2019. DOI: 10.1016/J.TECHFORE.2018.05.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0040162517312751. Acesso em: 9 out. 2023.

132 JUN, S.; PARK, S. S. Examining technological innovation of Apple using patent analysis. **Industrial Management and Data Systems**, [*s. l.*], v. 113, n. 6, p. 890–907, 2013. DOI: 10.1108/IMDS-01-2013-0032. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/IMDS-01-2013-0032/full/html. Acesso em: 9 out. 2023.

133 CIFFOLILLI, A.; MUSCIO, A. Industry 4.0: national and regional comparative advantages in key enabling technologies. **European Planning Studies**, [*s. l.*], v. 26, n. 12, p. 2323–2343, Dec. 2018. DOI: 10.1080/09654313.2018.1529145. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/09654313.2018.1529145. Acesso em: 9 out. 2023.

134 PAULO, A. F.; GRAEFF, C. F. O.; PORTO, G. S. Uncovering emerging photovoltaic technologies based on patent analysis. **World Patent Information**, [*s. l.*], v. 73, p. 102181, June 2023. DOI: 10.1016/J.WPI.2023.102181. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S017221902300011X. Acesso em: 9 out. 2023.

135 FISCHER, B. B.; SCHAEFFER, P. R.; VONORTAS, N. S. Evolution of university-industry collaboration in Brazil from a technology upgrading perspective. **Technol Forecast Soc Change**, [*s. l.*], v. 145, p. 330–340, Aug. 2019. DOI: 10.1016/J.TECHFORE.2018.05.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0040162517312751. Acesso em: 9 out. 2023.

136 SCHOCAIR, M. M.; DIAS, A. A.; GALINA, S. V. R.; AMARAL, M. The Evolution of the Triple Helix Thematic: a Social Networks Analysis. **Triple Helix**, [*s. l.*], v. 9, n. 3, p. 325–368, Apr. 2023. DOI: 10.1163/21971927-BJA10037. Disponível em: https://brill.com/view/journals/thj/9/3/article-p325_6.xml?language=en. Acesso em: 9 out. 2023.

137 ZHANG, R. *et al*. Collaborative relationship discovery in green building technology innovation: Evidence from patents in China's construction industry. **J Clean Prod**, [*s. l.*], v. 391, p. 136041, Mar. 2023. DOI: 10.1016/J.JCLEPRO.2023.136041. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652623001993. Acesso em: 9 out. 2023.

138 PAULO, A. F.; RIBEIRO, E. M. S.; PORTO, G. S. Mapping countries cooperation networks in photovoltaic technology development based on patent analysis. **Scientometrics**, [*s. l.*], v. 117, n. 2, p. 667–686, Nov. 2018. DOI: 10.1007/S11192-018-2892-6. Disponível em: https://link.springer.com/article/10.1007/s11192-018-2892-6. Acesso em: 9 out. 2023.

139 ZHANG, R. *et al*. Collaborative relationship discovery in green building technology innovation: Evidence from patents in China's construction industry. **J Clean Prod**, [*s. l.*], v. 391, p. 136041, Mar. 2023. DOI: 10.1016/J.JCLEPRO.2023.136041. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0959652623001993. Acesso em: 9 out. 2023.

140 WANG, X. *et al*. Collaboration network and pattern analysis: Case study of dye-sensitized solar cells. **Scientometrics**, [*s. l.*], v. 98, n. 3, p. 1745–1762, Nov. 2014. DOI: 10.1007/S11192-013-1180-8. Disponível em: https://link.springer.com/article/10.1007/s11192-013-1180-8. Acesso em: 9 out. 2023.

141 HUANG, M. H.; DONG, H. R.; CHEN, D. Z. Globalization of collaborative creativity through cross-border patent activities. **J Informetr**, [*s. l.*], v. 6, n. 2, p. 226–236, Apr. 2012. DOI: 10.1016/J.JOI.2011.10.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S175115771100099X. Acesso em: 9 out. 2023.

142 LIU, W. *et al*. Exploring and Visualizing the Patent Collaboration Network: A Case Study of Smart Grid Field in China. **Sustainability**, [*s. l.*],

v. 11, n. 2, p. 465, Jan. 2019. DOI: 10.3390/SU11020465. Disponível em: https://www.mdpi.com/2071-1050/11/2/465. Acesso em: 9 out. 2023.

143 PENG, F. *et al.* Evolution Characteristics of Government-Industry-University Cooperative Innovation Network of Electronic Information Industry in Liaoning Province, China. **Chin Geogr Sci**, [*s. l.*], v. 29, n. 3, p. 528–540, June 2019. DOI: 10.1007/S11769-019-1047-X. Disponível em: https://link.springer.com/article/10.1007/s11769-019-1047-x. Acesso em: 9 out. 2023.

144 BASSO, F. G.; PEREIRA, C. G.; PORTO, G. S. Cooperation and technological areas in the state universities of São Paulo: An analysis from the perspective of the triple helix model. **Technol Soc**, [*s. l.*], v. 65, p. 101566, May 2021. DOI: 10.1016/J.TECHSOC.2021.101566. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0160791X21000415. Acesso em: 9 out. 2023.

145 FIORI, G. M. L.; BASSO, F. G.; PORTO, G. S. Cooperation in R&D in the pharmaceutical industry: Technological and clinical trial networks in oncology. **Technol Forecast Soc Change**, [*s. l.*], v. 176, p. 121426, Mar. 2022. DOI: 10.1016/J.TECHFORE.2021.121426. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S004016252100857X. Acesso em: 9 out. 2023.

146 XIANG, J. Q.; MA, F.; WANG, H. The effect of intellectual property treaties on international innovation collaboration: a study based on USPTO patents during 1976–2017. **Library Hi Tech**, [*s. l.*], v. 41, n. 2, p. 666–682, June 2023. DOI: 10.1108/LHT-08-2020-0202. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/LHT-08-2020-0202/full/html. Acesso em: 9 out. 2023.

147 ISFANDYARI-MOGHADDAM, A. *et al.* Global scientific collaboration: A social network analysis and data mining of the co-authorship networks. **J Inf Sci**, [*s. l.*], v. 49, n. 4, p. 1126-1141, Aug. 2021. DOI: 10.1177/01655515211040655. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/01655515211040655. Acesso em: 9 out. 2023.

148 PAULO, A. F.; PORTO, G. S. Unveiling the cooperation dynamics in the photovoltaic technologies' development. **Renewable and Sustainable Energy Reviews**, [*s. l.*], v. 187, p. 113694, Nov. 2023. DOI: 10.1016/J.

RSER.2023.113694. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1364032123005518. Acesso em: 9 out. 2023.

149 LEYDESDORFF, L.; WAGNER, C. S. International collaboration in science and the formation of a core group. **J Informetr**, [*s. l.*], v. 2, n. 4, p. 317–325, Oct. 2008. DOI: 10.1016/J.JOI.2008.07.003. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1751157708000448. Acesso em: 9 out. 2023.

150 ABBASI, A.; HOSSAIN, L.; LEYDESDORFF, L. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. **J Informetr**, [*s. l.*], v. 6, n. 3, p. 403–412, July 2012. DOI: 10.1016/J.JOI.2012.01.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S175115771200003X. Acesso em: 9 out. 2023.

151 AHUJA, M. K.; GALLETTA, D. F.; CARLEY, K. M. Individual Centrality and Performance in Virtual R&D Groups: An Empirical Study. **Management Science**, [*s. l.*], v. 49, n. 1, p. 21–38, Jan. 2003. DOI: 10.1287/MNSC.49.1.21.12756. Disponível em: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.49.1.21.12756. Acesso em: 9 out. 2023.

152 WAGNER, C. S.; LEYDESDORFF, L. Network structure, self-organization, and the growth of international collaboration in science. **Res Policy**, [*s. l.*], v. 34, n. 10, p. 1608–1618, Dec. 2005. DOI: 10.1016/J.RESPOL.2005.08.002. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0048733305001745. Acesso em: 9 out. 2023.

153 OTTE, E.; ROUSSEAU, R. Social network analysis: a powerful strategy, also for the information sciences. **Journal of Documentation**, [*s. l.*], v. 28, n. 6, p. 441–453, Dec. 2002. DOI: 10.1177/016555150202800601. Disponível em: https://journals.sagepub.com/doi/10.1177/016555150202800601. Acesso em: 9 out. 2023.

154 PAULO, A. F.; PORTO, G. S. Solar energy technologies and open innovation: A study based on bibliometric and social network analysis. Energy Policy, [*s. l.*], v. 108, p. 228–238, Sept. 2017. DOI: 10.1016/J.ENPOL.2017.06.007. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0301421517303610. Acesso em: 9 out. 2023.

155 PILKINGTON, A.; MEREDITH, J. The evolution of the intellectual structure of operations management—1980–2006: A citation/co-citation analysis. **Journal of Operations Management**, [*s. l.*], v. 27, n. 3, p. 185–202, June 2009. DOI: 10.1016/J.JOM.2008.08.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0272696308000582. Acesso em: 9 out. 2023.

156 NOBRE, G. C.; TAVARES, E. Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study. **Scientometrics**, [*s. l.*], v. 111, n. 1, p. 463–492, Apr. 2017. DOI: 10.1007/S11192-017-2281-6. Disponível em: https://link.springer.com/article/10.1007/s11192-017-2281-6. Acesso em: 9 out. 2023.

157 HOTA, P. K.; SUBRAMANIAN, B.; NARAYANAMURTHY, G. Mapping the Intellectual Structure of Social Entrepreneurship Research: A Citation/Co-citation Analysis. **Journal of Business Ethics**, [*s. l.*], v. 166, n. 1, p. 89–114, Sept. 2020. DOI: 10.1007/S10551-019-04129-4. Disponível em: https://link.springer.com/article/10.1007/s10551-019-04129-4. Acesso em: 9 out. 2023.

158 MARIANI, M.; BORGHI, M. Industry 4.0: A bibliometric review of its managerial intellectual structure and potential evolution in the service industries. **Technol Forecast Soc Change**, [*s. l.*], v. 149, p. 119752, Dec. 2019. DOI: 10.1016/J.TECHFORE.2019.119752. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0040162519311345. Acesso em: 9 out. 2023.

159 MARTÍ-PARREÑO, J.; MÉNDEZ-I BÁÑEZ, E.; ALONSO-ARROYO, A. The use of gamification in education: a bibliometric and text mining analysis. **J Comput Assist Learn**, [*s. l.*], v. 32, n. 6, p. 663–676, Dec. 2016. DOI: 10.1111/jcal.12161. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/jcal.12161. Acesso em: 9 out. 2023.

160 KOSEOGLU, M. A. Mapping the institutional collaboration network of strategic management research: 1980–2014. **Scientometrics**, [*s. l.*], v. 109, n. 1, p. 203–226, Oct. 2016. DOI: 10.1007/S11192-016-1894-5. Disponível em: https://link.springer.com/article/10.1007/s11192-016-1894-5. Acesso em: 9 out. 2023.

161 SREENIVASAN, A. *et al*. Mapping analytical hierarchy process research to sustainable development goals: Bibliometric and social network analysis. **Heliyon**, [*s. l.*], v. 9, n. 8, p. e19077, Aug. 2023. DOI: 10.1016/J.HELIYON.2023.E19077. Disponível em: https://pubmed.ncbi.nlm.nih.gov/37636443/. Acesso em: 9 out. 2023.

162 TRABSKAIA, I. *et al*. A Bibliometric Analysis of Social Entrepreneurship and Entrepreneurial Ecosystems. **Administrative Sciences 2023**, [*s. l.*], v. 13, n. 3, p. 75, Mar. 2023. DOI: 10.3390/ADMSCI13030075. Disponível em: https://www.mdpi.com/2076-3387/13/3/75. Acesso em: 9 out. 2023.

163 ZHANG, G.; WEI, F.; WANG, P. Opening the black box of Library Hi Tech: a social network and bibliometric analysis. **Library Hi Tech**, [*s. l.*], 2023. DOI: 10.1108/LHT-12-2022-0556. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/LHT-12-2022-0556/full/html. Acesso em: 9 out. 2023.

164 CRUMPTON, C. D. *et al*. Evaluation of public policies in Brazil and the United States: a research analysis in the last 10 years. **Revista de Administração Pública**, [*s. l.*], v. 50, n. 6, p. 981–1001, Nov. 2016. DOI: 10.1590/0034-7612156363. Disponível em: https://www.scielo.br/j/rap/a/ptZ4nqddFYXYsL3ZqCSKgRz/?lang=en. Acesso em: 9 out. 2023.

165 RIBEIRO, H. C. M. Estratégia em destaque: duas décadas de produção científica do evento 3Es à luz da análise de redes sociais. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 26, n. 4, p. 113–150, Jan. 2022. DOI: 10.1590/1981-5344/25199.

166 GOMES, V. D. S.; SILVA, M. R. Produção em Análise de Redes Sociais: estudo bibliométrico na BRAPCI. **AtoZ**: novas práticas em informação e conhecimento, [*s. l.*], v. 11, p. 1, mar. 2022. DOI: 10.5380/atoz.v11i0.80813. Disponível em: https://brapci.inf.br/index.php/res/v/193829. Acesso em: 9 out. 2023.

167 JACKSON, M. O. An Overview of Social Networks and Economic. *In*: BENHABIB, J.; BISIN, A.; JACKSON, M. **Handbook of Social Economics**. Palo Alto: Elsevier Press, 2010. 96 p. Disponível em: https://web.stanford.edu/~jacksonm/socialnetecon-chapter.pdf/ Acesso em: 9 out. 2023.

168 BISGIN, H.; AGARWAL, N.; XU, X. Investigating homophily in online social networks. *In*: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, WI 2010, 1st, 2010. **Proceedings** [...]. [*S. l.*]: IEEE, 2010. DOI: 10.1109/WI-IAT.2010.61. Disponível em: https://ieeexplore.ieee.org/document/5616320. Acesso em: 9 out. 2023.

169 KHANAM, K. Z.; SRIVASTAVA, G.; MAGO, V. The Homophily Principle in Social Network Analysis. **Proc. ACM Meas. Anal. Comput. Syst**, [*s. l.*], v. 37, n. 111, p. 28, Aug. 2020. DOI: 10.1145/1122445.1122456. Disponível em: https://arxiv.org/abs/2008.10383. Acesso em: 9 out. 2023.

170 NEWMAN, M. Measures and metrics. *In*: NEWMAN, M. **Networks**, 2. ed. Oxford: Oxford University Press, 2018. p. 158–217. DOI: 10.1093/OSO/9780198805090.003.0007. Disponível em: https://academic.oup.com/book/27884/chapter=-abstract203815468/?redirectedFrom-fulltext. Acesso em: 9 out. 2023.

171 ANDRAS, P. Research: metrics, quality, and management implications. **Res Eval**, [*s. l.*], v. 20, n. 2, p. 90–106, June 2011. DOI: 10.3152/095820211X12941371876265. Disponível em: https://academic.oup.com/rev/article-abstract/20/2/90/1577626. Acesso em: 9 out. 2023.

172 JACKSON, M. O. **Social and Economic Networks**. [*S. l.: s. n.*], 2010. 520 p. Disponível em: https://academic.oup.com/book/12738/chapter-abstract/162849942?redirectedFrom=fulltext. Acesso em: 9 out. 2023.

173 NEWMAN, M. E. J. Mathematics of networks. *In*: NEWMAN, M. E. J. **Networks**: An Introduction, 2010. p. 109–167. DOI: 10.1093/acprof:oso/9780199206650.003.0006. Disponível em: https://academic.oup.com/book/27303/chapter-abstract/196961844?redirectedFrom=fulltext. Acesso em: 9 out. 2023.

174 EASLEY, D.; KLEINBERG, J. **Networks, Crowds, and Markets**: Reasoning about a Highly Connected World. Cambridge: University Press, 2010.

175 ZHANG, M. **Handbook of Social Network Technologies and Applications**. [*S. l.: s. n.*], 2010. DOI: 10.1007/978-1-4419-7142-5. Disponível em: https://link.springer.com/book/10.1007/978-1-4419-7142-5. Acesso em: 9 out. 2023.

176 SCOTT, J.; CARRINGTON, P. J. **The SAGE Handbook of Social Network Analysis**. New York: Sage Publishing, 2011. 640 p.

177 WATERS, N. Social network analysis**.** *In*: **Handbook of Regional Science**. [*S. l.*: *s. n.*], 2014. p. 725–740. DOI: 10.1007/978-3-642-23430-9_49. Disponível em: https://www.researchgate.net/publication/304182204_Social_Network_Analysis. Acesso em: 9 out. 2023.

178 IZQUIERDO, L. R.; HANNEMAN, R. A. Introduction to the Formal Analysis of Social Networks. **Published Electronically**, [*s. l.*], 2006. Disponível em: https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55.))/reference/referencespapers.aspx?referenceid=2391862. Acesso em: 9 out. 2023.

179 JACKSON, M. O. **Social and Economic Networks**. [*S. l.: s. n.*]. 2010. 520 p. Disponível em: https://academic.oup.com/book/12738/chapter-abstract/162849942?redirectedFrom=fulltext. Acesso em: 9 out. 2023.

180 LEEM, B. H.; CHUN, H. Measuring the influence of efficient ports using social Network Metrics. **International Journal of Engineering Business Management**, [*s. l.*], v. 7, n. 1, p. 1–8, Jan. 2015. DOI: 10.5772/60040. Disponível em: https://journals.sagepub.com/doi/10.5772/60040. Acesso em: 9 out. 2023.

181 NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. **Phys Rev E**, [*s. l.*], v. 69, n. 2, p. 026113, Feb. 2004. DOI: 10.1103/PhysRevE.69.026113. Disponível em: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.69.026113. Acesso em: 9 out. 2023.

182 CHUNAEV, P. Community detection in node-attributed social networks: A survey. **Comput Sci Rev**, [*s. l.*], v. 37, p. 100286, Aug. 2020. DOI: 10.1016/J.COSREV.2020.100286. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1574013720303865. Acesso em: 9 out. 2023.

183 CENTOLA, D. The spread of behavior in an online social network experiment. **Science (1979)**, [*s. l.*], v. 329, n. 5996, p. 1194–1197, Sept. 2010. DOI: 10.1126/SCIENCE.1185231. Disponível em: https://www.science.org/doi/10.1126/science.1185231. Acesso em: 9 out. 2023.

184 LESKOVEC, J.; ADAMIC, L. A.; HUBERMAN, B. A. The dynamics of viral marketing. **ACM Transactions on the Web (TWEB)**, [*s. l.*], v. 1, n. 1, May. 2007. DOI: 10.1145/1232722.1232727. Disponível em: https://dl.acm.org/doi/10.1145/1232722.1232727. Acesso em: 9 out. 2023.

185 ROGERS, E. M.; SINGHAL, A.; QUINLAN, M. M. Diffusion of Innovations. *In:* STACKS, D.; SALWEN, M. **An Integrated Approach to Communication Theory and Research**. New York: Routledge, 2014. p. 432–448. DOI: 10.4324/9780203887011-36. Disponível em: https://www.taylorfrancis.com/chapters/edit/10.4324/9780203887011-36/diffusion-innovations-everett-rogers-arvind-singhal-margaret-quinlan. Acesso em: 9 out. 2023.

186 PYLE, D.; CERRA, D. D.; KAUFMANN, M. **Data Preparation for Data Mining**. [*S. l.*: *s. n.*], 1999.

187 RANKOVI, N. Contribution to methods and techniques of scientific research: Structure of the scientific research report - 2. Methods and data collection and processing. **Glasnik Sumarskog fakulteta**, [*s. l.*], n. 124, p. 137–142, 2021. DOI: 10.2298/GSF2124137R. Disponível em: https://doiserbia.nb.rs/Article.aspx?id=0353-45372124137R. Acesso em: 9 out. 2023.

188 SNIJDERS, T. A. B. **Models for Longitudinal Network Data**. New York: Cambridge University Press, (in press), 2006.

189 GOLDER, S. A.; MACY, M. W. Digital Footprints: Opportunities and Challenges for Online Social Research. *In*: **Annual Review of Sociology**, [*s. l.*], v. 40, p. 129–152, July 2014. DOI: 10.1146/ANNUREV-SOC-071913-043145. Disponível em: https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-071913-043145. Acesso em: 9 out. 2023.

190 FU, X.; LUO, J. D.; BOOS, M. **Social network analysis**: interdisciplinary approaches and case studies. Flórida: CRC Press, 2017.

191 MICHEL, M. C. How significant are your data? the need for a culture shift. **Naunyn Schmiedebergs Arch Pharmacol**, [*s. l.*], v. 387, n. 11, p. 1015–1016, Aug. 2014. DOI: 10.1007/S00210-014-1044-7. Disponível em: https://link.springer.com/article/10.1007/s00210-014-1044-7. Acesso em: 9 out. 2023.

192 VANNAN, S. *et al*. Data Sets Are Foundational to Research. Why Don't We Cite Them? **Eos, Transactions American Geophysical Union**, [*s. l.*], v. 101, Nov. 2020. DOI: 10.1029/2020EO151665. Disponível em: https://eos.org/opinions/data-sets-are-foundational-to-research-why-dont-we-cite-them. Acesso em: 9 out. 2023.

193 WINKLER, S.; ZEADALLY, S. Privacy Policy Analysis of Popular Web Platforms. **IEEE Technology and Society Magazine**, [*s. l.*], v. 35, n. 2, p. 75–85, June 2016. DOI: 10.1109/MTS.2016.2554419. Disponível em: https://ieeexplore.ieee.org/document/7484849. Acesso em: 9 out. 2023.

194 ZOU, D. *et al*. Biological Databases for Human Research. **Genomics Proteomics Bioinformatics**, [*s. l.*], v. 13, n. 1, p. 55–63, Feb. 2015. DOI: 10.1016/J.GPB.2015.01.006. Disponível em: https://pubmed.ncbi.nlm.nih.gov/25712261/. Acesso em: 9 out. 2023.

195 HUTTON, L.; HENDERSON, T. Toward Reproducibility in Online Social Network Research. **IEEE Trans Emerg Top Comput**, [*s. l.*], v. 6, n. 1, p. 156–167, Jan. 2018. DOI: 10.1109/TETC.2015.2458574. Disponível em: https://ieeexplore.ieee.org/document/7163309. Acesso em: 9 out. 2023.

196 HOFMAN, J. M. *et al*. Expanding the Scope of Reproducibility Research Through Data Analysis Replications. *In*: THE WEB CONFERENCE 2020 - COMPANION OF THE WORLD WIDE WEB CONFERENCE, WWW 2020. [*S. l.*]: Association for Computing Machinery, 2020. p. 567–571. DOI: 10.1145/3366424.3383417. Disponível em: https://dl.acm.org/doi/10.1145/3366424.3383417. Acesso em: 9 out. 2023.

197 FRIEMEL, T. N. Social Network Analysis. *In*: WILEY, J. *et al*. **The International Encyclopedia of Communication Research Methods**. [*S. l.: s. n.*], 2017. p. 1-14. DOI: 10.1002/9781118901731.IECRM0235. Disponível em: https://www.researchgate.net/publication/315893553_Social_Network_Analysis. Acesso em: 9 out. 2023.

198 CAMPBELL, W. M.; DAGLI, C. K.; WEINSTEIN, C. J. Social network analysis with content and graphs. **Lincoln Laboratory Journal**, [*s. l.*], v. 20, 2013. Disponível em: https://www.ll.mit.edu/sites/default/files/publication/doc/social-network-analysis-content-graphs-campbell-ja-22727.pdf. Acesso em: 9 out. 2023.

199 MACINDOE, O.; RICHARDS, W. Graph comparison using fine structure analysis. *In*: SOCIALCOM 2010: 2ND IEEE INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING, PASSAT 2010: 2ND IEEE INTERNATIONAL CONFERENCE ON PRIVACY, SECURITY, RISK AND TRUST. **Proceedings** [...]. Minneapolis: IEEE, p. 193–200, 2010. DOI: 10.1109/SOCIALCOM.2010.35. Disponível em: https://ieeexplore.ieee.org/document/5590440. Acesso em: 9 out. 2023.

200 DEHMER, M.; EMMERT-STREIB, F.; SHI, Y. Quantitative Graph Theory: A new branch of graph theory and network science. **Inf Sci**, [*s. l.*], v. 418/419, p. 575–580, Dec. 2017. DOI: 10.1016/J.INS.2017.08.009. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0020025517308666. Acesso em: 9 out. 2023.

201 SERGEANT, S. *et al*. On the Road Together: Issues Observed in the Process of a Research Duo Working Together in a Long-Term and Intense Collaboration in an Inclusive Research Project. **Social Sciences**, [*s. l.*], v. 11, n. 5, p. 185, Apr. 2022. DOI: 10.3390/SOCSCI11050185. Disponível em: https://www.mdpi.com/2076-0760/11/5/185. Acesso em: 9 out. 2023.

202 DAS, K. *et al*. Influential nodes in social networks: Centrality measures. *In*: PAL, Madhumangal; SAMANTA, Sovan; PAL, Anita. **Handbook of Research on Advanced Applications of Graph Theory in Modern Society**, [*s. l.*], p. 371–385, Aug. 2019. DOI: 10.4018/978-1-5225-9380-5.CH015. Disponível em: https://www.igi-global.com/chapter/influential-nodes-in-social-networks/235544. Acesso em: 9 out. 2023.

203 NOOSRIKONG, C.; NGAMSURIYAROJ, S.; AYUDHYA, S. P. N. Identifying focus research areas of computer science researchers from publications. *In*: IEEE REGION 10 ANNUAL INTERNATIONAL CONFERENCE, PROCEEDINGS/TENCON. **Proceedings** [...]. Penang: IEEE, 2017. DOI: 10.1109/TENCON.2017.8227970. Disponível em: https://ieeexplore.ieee.org/document/8227970. Acesso em: 9 out. 2023.

204 KOHL, M.; WIESE, S.; WARSCHAID, B. Cytoscape: Software for Visualization and Analysis of Biological Networks. **Methods in Molecular Biology**, [*s. l.*], v. 696, p. 291–303, 2011. DOI: 10.1007/978-1-60761-987-1_18. Disponível em: https://link.springer.com/protocol/10.1007/978-1-60761-987-1_18. Acesso em: 9 out. 2023.

205 SHANNON, P. *et al*. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. **Genome Res**, [*s. l.*], v. 13, n. 11, p. 2498–2504, Nov. 2003. DOI: 10.1101/GR.1239303. Disponível em: https://pubmed.ncbi.nlm.nih.gov/14597658/. Acesso em: 9 out. 2023.

206 AMAT, C. B. Análisis y visualización de redes con Gephi. Redes. **Revista hispana para el análisis de redes sociales**, [*s. l.*], v. 25, n. 1, p. 201–209, mayo 2014. DOI: 10.5565/rev/redes.499. Disponível em: https://revistes.uab.cat/redes/article/view/v25-n1-benito/. Acesso em: 9 out. 2023.

207 BASTIAN, M.; HEYMANN, S.; JACOMY, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *In*: INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA, 3rd. **Proceedings** [...]. Palo Alto: AAAI Press, 2009. DOI: 10.1609/ICWSM.V3I1.13937. Disponível em: https://ojs.aaai.org/index.php/ICWSM/article/view/13937. Acesso em: 9 out. 2023.

208 CSARDI, G. **The Igraph Software Package for Complex Network Research**. [*S. l.*: *s. n.*], 2005. Disponível em: https://www.researchgate.net/publication/221995787. Acesso em: 7 out. 2023.

209 JU, W. *et al*. iGraph: an incremental data processing system for dynamic graph. **Front Comput Sci**, [*s. l.*], v. 10, n. 3, p. 462–476, June 2016. DOI: 10.1007/S11704-016-5485-7. Disponível em: https://link.springer.com/article/10.1007/s11704-016-5485-7. Acesso em: 9 out. 2023.

210 SKLAR, E. NetLogo, a Multi-agent Simulation Environment. **Artif Life**, [*s. l.*], v. 13, n. 3, p. 303–311, July 2007. DOI: 10.1162/ARTL.2007.13.3.303. Disponível em: https://direct.mit.edu/artl/article-abstract/13/3/303/2563/NetLogo-a-Multi-agent-Simulation-Environment?redirectedFrom=fulltext. Acesso em: 9 out. 2023.

211 TISUE, S.; WILENSKY, U. NetLogo: A simple environment for modeling complexity. *In*: INTERNATIONAL CONFERENCE ON COMPLEX SYSTEMS. **Proceedings** [...]. [*S. l.: s. n.*], 2004. Disponível em: https://www.researchgate.net/publication/230818221. Acesso em: 7 out. 2023.

212 HAGBERG, A.; SWART, P. J.; SCHULT, D. A. Exploring network structure, dynamics, and function using NetworkX. *In*: CONFERENCE: SCIPY, 8th, 2008, Pasadena. **Proceedings** [...]. Pasadena, 2008. Disponível em: https://www.osti.gov/biblio/960616. Acesso em: 9 out. 2023.

213 PLATT, E. L. **Network Science with Python and NetworkX Quick Start Guide**. [*S. l.*]: Packt Publishing Ltd, 2019. 190 p.

214 AL-TAIE, M. Z.; KADRY, S. **Python for Graph and Network Analysis**. [*S. l.: s. n.*], 2017. DOI: 10.1007/978-3-319-53004-8. Disponível em: https://link.springer.com/book/10.1007/978-3-319-53004-8. Acesso em: 9 out. 2023.

215 HIMELBOIM, I.; SMITH, M. A. NodeXL. **The International Encyclopedia of Communication Research Methods.** [*S. l.*: *s. n.*], 2017. p. 1–3. DOI: 10.1002/9781118901731.IECRM0167. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118901731.iecrm0167. Acesso em: 9 out. 2023.

216 SMITH, M. A. *et al*. Analyzing (social media) networks with NodeXL. *In*: INTERNATIONAL CONFERENCE ON COMMUNITIES AND TECHNOLOGIES, 4th, 2009. **Proceedings** [...]. New York: Association for Computing Machinery, 2009. DOI: 10.1145/1556460.1556497. Disponível em: https://dl.acm.org/doi/10.1145/1556460.1556497. Acesso em: 9 out. 2023.

217 MRVAR, A.; BATAGELJ, V. Analysis and visualization of large networks with program package Pajek. **Complex Adaptive Systems Modeling**, [*s. l.*], v. 4, n. 1, p. 1–8, Dec. 2016. DOI: 10.1186/S40294-016-0017-8. Disponível em: https://casmodeling.springeropen.com/articles/10.1186/s40294-016-0017-8. Acesso em: 9 out. 2023.

218 BATAGELJ, V.; Ljubljana, A. M. **Pajek Program for Analysis and Visualization of Large Networks**. [*S. l.*: *s. n.*], 2011. Disponível em: http://vlado.fmf.uni-lj.si/. Acesso em: 7 out. 2023.

219 LUKE, D. **A User's Guide to Network Analysis in R**. [*S. l.*: *s. n.*], 2015. DOI: 10.1007/978-3-319-23883-8. Disponível em: https://link.springer.com/book/10.1007/978-3-319-23883-8. Acesso em: 9 out. 2023.

220 JOHNSON, J. D. UCINET: A software tool for network analysis. **Commun Educ**, [*s. l.*], v. 36, n. 1, p. 92–94, Jan. 1987. DOI:

10.1080/03634528709378647. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/03634528709378647. Acesso em: 9 out. 2023.

221 APOSTOLATO, I.-A. An overview of Software Applications for Social Network Analysis. **Int Rev Soc Res**, [*s. l.*], v. 3, n. 3, p. 71–77, Oct. 2013. DOI: 10.1515/IRSR-2013-0023. Disponível em: http://archive.sciendo.com/IRSR/irsr.2013.3.issue-3/irsr-2013-0023/irsr-2013-0023.pdf. Acesso em: 9 out. 2023.

222 ARRUDA, H. *et al*. VOSviewer and Bibliometrix. **J Med Libr Assoc**, [*s. l.*], v. 110, n. 3, p. 392, July 2022. DOI: 10.5195/JMLA.2022.1434. Disponível em: https://jmla.pitt.edu/ojs/jmla/article/view/1434. Acesso em: 9 out. 2023.

223 VAN ECK, N. J.; WALTMAN, L. **VOSviewer Manual**. [*S. l.: s. n.*], 2012. Disponível em: www.vosviewer.com. Acesso em: 7 out. 2023.

224 AKHTAR, N. Social network analysis tools. *In*: INTERNATIONAL CONFERENCE ON COMMUNICATION SYSTEMS AND NETWORK TECHNOLOGIES, CSNT 2014, 4th. **Proceedings [...]**. Bhopal: IEEE, 2014. DOI: 10.1109/CSNT.2014.83. Disponível em: https://ieeexplore.ieee.org/document/6821424. Acesso em: 9 out. 2023.

225 JACOMY, M. *et al*. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. **PLoS One**, [*s. l.*], v. 9, n. 6, p. 1–12, 2014. DOI: 10.1371/journal.pone.0098679. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679. Acesso em: 9 out. 2023.

226 FRUCHTERMAN, T. M. J.; REINGOLD, E. M. Graph drawing by force-directed placement. **Softw Pract Exp**, [*s. l.*], v. 21, n. 11, p. 1129–1164, Nov. 1991. DOI: 10.1002/SPE.4380211102. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/spe.4380211102. Acesso em: 9 out. 2023.

227 BLONDEL, V. D. *et al*. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics**: Theory and Experiment, [*s. l.*], v. 2008, n. 10, 2008. DOI: 10.1088/1742-5468/2008/10/P10008. Disponível em: https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008. Acesso em: 9 out. 2023.

DADOS DO AUTOR:

Alex Fabianne de Paulo

Alex Fabianne de Paulo é doutor em Administração de Organizações pela Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto na Universidade de São Paulo (FEA-RP/USP), mestre e graduado em Ciência da Computação pela Universidade Federal de Uberlândia (UFU). É Professor Adjunto do curso de Gestão da Informação e do Programa de Pós-Graduação em Administração na Universidade Federal de Goiás (UFG). Foi pesquisador visitante na Aston Business School (Reino Unido). É pesquisador colaborador do Instituto de Estudos Avançados (IEA/USP) e do Núcleo de Gestão em Pesquisa em Tecnologia da Informação (NGPTI/UFG). Seus interesses de pesquisa incluem temáticas relacionadas à administração da informação, gestão da inovação, tendências tecnológicas, inteligência de negócios, tomada de decisão e competitividade.

https://orcid.org/0000-0003-3610-2255
alex.fabianne@gmail.com

Como referenciar o capítulo 6:

DE PAULO, Alex Fabianne. Potencialidades investigativas utilizando análise de redes sociais. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 6. p. 139-208. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap6.

7. FACEPAGER: UMA FERRAMENTA DE EXTRAÇÃO E RASPAGEM DE DADOS DE CÓDIGO ABERTO

Fábio Castro Gouveia

7.1 INTRODUÇÃO

Obter dados disponíveis na internet por intermédio de *APIs* (*Application Programming Interface*, ou Interface de Programação de Aplicações) nem sempre é uma tarefa fácil. Há que se conhecer os processos de autenticação para acesso, os campos que podem ser buscados e os tipos de consultas permitidas. Além disso, pode ser necessário seguir padrões para a paginação dos conteúdos desejados e a automatização das consultas. Ao final, os dados obtidos precisam ser armazenados em algum formato de base de dados. Para obtermos sucesso neste objetivo, podemos recorrer a diversos *scripts* disponíveis no *GitHub* ou repositórios similares, ou mesmo pedir para algum sistema de inteligência artificial generativa, como o *ChatGPT* ou *Bard*, criar um código inicial que, após alguns ajustes, pode já estar efetuando o processo de coleta desejado. Entretanto, este processo pode ficar mais fácil se tivermos uma ferramenta que já possua exemplos de coletas possíveis, e que facilite todos os processos que acabamos de listar. É neste lugar que se situa o *Facepager*, uma ferramenta de extração e raspagem de dados de código aberto. Ela foi projetada por Jünger e Keyling (2019) com o objetivo de facilitar a coleta e gerenciamento de dados de várias plataformas de mídia social e serviços da *web*. O processo é simplificado pela oferta de *presets* e, também, com a possibilidade de uso de uma aba genérica para coletas em *APIs* diversas. Assim toda execução e gerenciamento de tarefas de extração e raspagem de dados é favorecida por uma interface gráfica amigável que independe da fonte de dados e que ao final armazena os dados em uma base de dados *SQLite*[78]

78 Uma base de dados em *SQL* desenvolvida e operada sob uma *biblioteca de linguagem C* de tamanho reduzido, autocontida, de alta confiabilidade e com todos os recursos *SQL*.

e permite que se exporte o conteúdo em arquivos *CSV*[79] (*Comma-Separated Values*, ou valores separados por vírgulas).

7.2 PRINCIPAIS CARACTERÍSTICAS DO FACEPAGER

O *Facepager*[80] é um *software* desenvolvido pelo grupo de pesquisa *Digital Media & Computational Methods* do Departamento de Comunicação da *University of Münster* - Alemanha, liderado pelo Dr. Jakob Jünger. Uma série de tutoriais são disponibilizados em inglês no canal do *YouTube*[81].

Dentre as principais características do *Facepager* podemos citar sua versatilidade ao coletar dados de diversas plataformas, efetuar consultas parametrizadas de forma sequencial, lidar com a paginação das consultas, seleção de dados a serem apresentados e exportados em diferentes formatos, uso de uma *GUI* (*Graphical User Interface*, ou Interface Gráfica do Usuário) amigável, ampliação de suas capacidades por *scripts* adicionais e o fato de ser um software aberto.

Pode ser executado nas versões para *Windows* e *MacOS*[82], e tem consultas já definidas para o *Facebook*, *Twitter*, *YouTube*, *Amazon*, dentre outros sites e fontes on-line listados nos *presets*, podendo inclusive ser configurado para tarefas de raspagem de dados (Figura 1).

Todo o processo de coleta de dados pode ser configurado com conjuntos de parâmetros necessários aos objetivos da pesquisa, assim como tem a

79 *CSVs* são arquivos cujos campos são separados por algum caractere definido. Em geral se utiliza a vírgula (,) ou o ponto e vírgula (;). Quando um ponto e vírgula é utilizado o arquivo também segue descrito como um *CSV*, porém tecnicamente ele seria um arquivo "Semicolon-Separated Values". Há também casos em que a tabulação é utilizada como separador. Nestes casos, eles podem ser chamados de *CSV* ou *TSVs* (*Tab-Separated Values*, ou valores separados por tabulação).

80 Disponível em: https://github.com/strohne/Facepager/. Acesso em: 1 out. 2023.

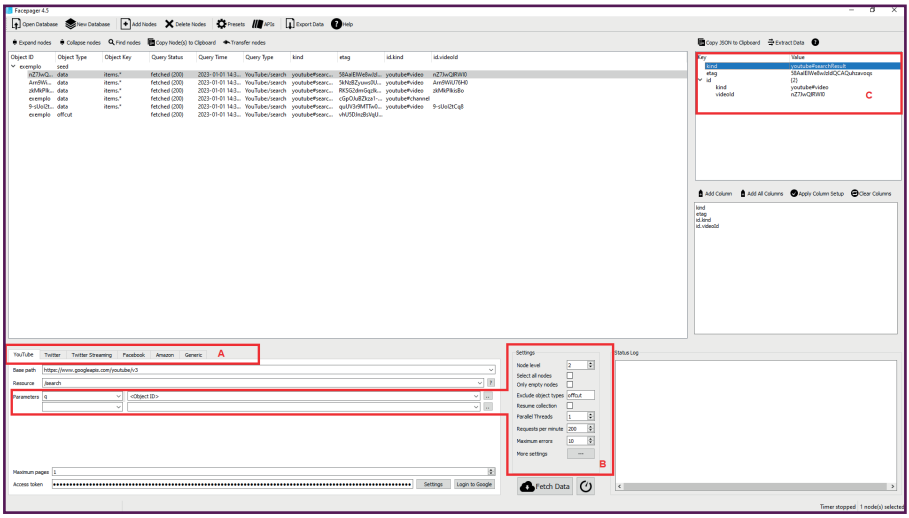
81 Disponível em: https://www.youtube.com/@facepager1740. Acesso em: 1 out. 2023.

82 Não há binários para *Linux*, mas é possível executar o *Facepager* seguindo passos descritos no *GitHub* do *software*.

capacidade de passar parâmetros necessários à paginação das consultas para executar um levantamento de grandes *datasets* (Figura 1). O *software* sofre atualizações constantes para se adaptar às instabilidades e mudanças nos acessos a dados, podendo se manter compatível com diferentes versões de *APIs* de um mesmo sistema. O usuário pode assim construir suas próprias consultas mantendo os resultados obtidos guardados com todos os parâmetros solicitados armazenados na base criada.

Os campos que serão apresentados e exportados podem ser selecionados a partir de consultas individuais onde os dados apresentados em uma árvore hierárquica em *JSON* (JavaScript Object Notation)[83] podem ser elencados como colunas que formarão um *CSV* final (Figura 1).

Figura 1 - Tela principal do *Facepager* com destaque para as diferentes fontes de dados (A), configurações possíveis dos parâmetros de uma *API* (B) e campos obtidos pela consulta (C).



Fonte: Elaborado pelo autor (2023).

Sua interface gráfica torna o processo de consultas as *APIs* mais intuitiva e sua capacidade de expansão por *scripts* permite que ele seja utilizado

83 *JSON*, ou *Notação de Objetos JavaScript*, é um formato de intercâmbio de dados leve e relativamente fácil de ser lido por humanos e gerado, interpretado e analisado por máquinas.

pelo pesquisador em diversos projetos que necessitem de acesso à dados disponibilizados *on-line* servindo até mesmo para raspagem de dados que não se encontram estruturados e acessíveis diretamente por uma interface de consulta.

Sendo um *software livre*, conta com apoio da comunidade que o utiliza, podendo sugerir melhorias no seu desenvolvimento de versões futuras. Seu uso por pesquisadores e analistas de dados interessados na coleta via *API* de redes sociais é o mais recorrente.

7.3 AS APIS E SEUS POTENCIAIS E CONSIDERAÇÕES

Uma *API* é um conjunto de regras e protocolos estabelecidos para permitir que um *software* interaja com outro. Num ambiente onde diversas fontes de dados estão disponíveis, definir métodos e formatos de dados que uma determinada aplicação pode usar para requerer e trocar informações com outra aplicação se tornou fundamental com o advento do *Big Data*. Ao mesmo tempo, o próprio processo de criação e documentação das *APIs* deve em tese dialogar com os *princípios FAIR* (*Findable, Accessible, Interoperable, and Reusable*, ou Localizável, Acessível, Interoperável e Reutilizável) (Wilkinson *et al.*, 2016), mas esta questão nem sempre têm êxito completo.

A documentação da *API* auxilia os usuários no entendimento de suas funcionalidades, processos e parâmetros, facilitando que os dados sejam localizáveis. A Acessibilidade se dá por intermédio de processos de controle de acesso por autenticação e pela disponibilização *online* dos dados. A interoperabilidade se configura pelo uso frequente de formatos de consumo[84] padronizados como *JSON* (JavaScript Object Notation) e *XML* (eXtensible

84 É importante compreender que os dados podem ser armazenados de diversas maneiras sendo os formatos *JSON* ou *XML* formatos de consumo e interface e não necessariamente os de armazenamento.

Markup Language)[85] permitindo que a troca de informações possa ocorrer entre dois sistemas sem problemas. Complementarmente, o uso de arquiteturas *REST* (Representational State Transfer) para *APIs* permite que sejam utilizados padrões de métodos *HTML* tornando as comunicações compatíveis com diversos sistemas e plataformas. Por último, a reutilização se promove tanto pelo aspecto modular do desenvolvimento das *APIs*, onde o próprio processo de codificação pode ser reutilizado, quanto pelo controle de versões que permite que sistemas sigam usando versões anteriores da *API* sem precisar de imediata atualização.

Lonborg e Bechmann (2014) já debatiam os benefícios para os estudos sobre redes sociais da abertura pelas empresas de seus dados via *APIs*. Uma das questões centrais era a possibilidade de se poder efetuar levantamentos tecnicamente neutros, servindo a estudos empíricos e ao desenvolvimento de metodologias e debates críticos sobre os processos e trocas dentro destas plataformas, assim como a exploração e superação de desafios associados aos estudos quantitativos e qualitativos no âmbito dos métodos digitais.

Com as restrições que o *Facebook* passou a empregar no acesso às postagens públicas, mesmo por intermédio de *APIs*, passamos a viver o que Bruns (2019) chamou de *APIcalypse*. Os estudos que se baseavam, por exemplo, no *Netvizz*[86] deixaram de ser possíveis, o que levou na prática à redução na frequência de um tipo de estudo de redes de relações entre páginas. Se por um lado o *Facebook* foi pioneiro no desenvolvimento de uma plataforma para desenvolvedores com acesso à perfis de usuários e diversos outros dados (Fetterman, 2006), após o escândalo da *Cambridge Analytica* os acessos aos seus conteúdos foram se tornando mais e mais

85 *XML*, ou Linguagem de Marcação Extensível, é uma linguagem de marcação como o *HTML*, porém utilizada para armazenamento e transporte de dados estruturados. Por ser extensível ela permite que sejam definidos conjuntos de regras e codificações. O formato é relativamente fácil de ser lido por humanos e facilmente interpretado por máquinas.

86 *Netvizz* era um aplicativo executado dentro do *Facebook* que permitia um conjunto de estudos de relações entre páginas com a exportação de dados em tabelas e arquivos de grafos com as relações entre os entes do *Facebook*. Foi desenvolvido por Bernhard Rieder e teve decretado o fim do seu acesso aos dados do *Facebook* pela *Meta* em 2019.

restritos, sendo hoje os estudos acadêmicos possíveis basicamente por intermédio do *CrowdTangle*[87].

Mais recentemente com o fechamento do acesso acadêmico pelo *X* (ex-Twitter), e a oferta de planos de acesso com custos exorbitantes, novamente o cenário dos estudos com métodos digitais se modificou. No caso do *X*, estudos acadêmicos se tornaram muito mais difíceis de serem efetuados, levantando uma questão sobre o necessário acompanhamento social dos ambientes *online* por pesquisadores (Mozelli, 2023).

De certa forma, o *APIcalypse* de Bruns (2019) se assemelha ao momento de crise ainda sem superação que os estudos *webométricos* passaram a vivenciar a partir do fechamento do acesso aos dados de *links* entre *sites* descrito por Gouveia (2012). Em um cenário de crise, novas metodologias podem surgir e os estudos podem ter de mudar seus objetivos alcançáveis ante o que passa a ser disponibilizado pelas plataformas. Estes são desafios que, como descreve Omena (2019, p. 5), fazem parte das

[...] limitações impostas pelas plataformas *web* (instabilidade, mudanças contínuas ou restrições ao acesso a dados públicos via interfaces técnicas) são alvo de crítica e reflexão para os investigadores dos media digitais.

Porém, estas questões têm relação com as ações daqueles que gerenciam as plataformas. As *APIs* seguem sendo um modo padronizado para que sistemas se comuniquem entre si, o que as torna uma tecnologia chave para a implementação de *princípios FAIR*, desde que sejam corretamente desenvolvidas e documentadas.

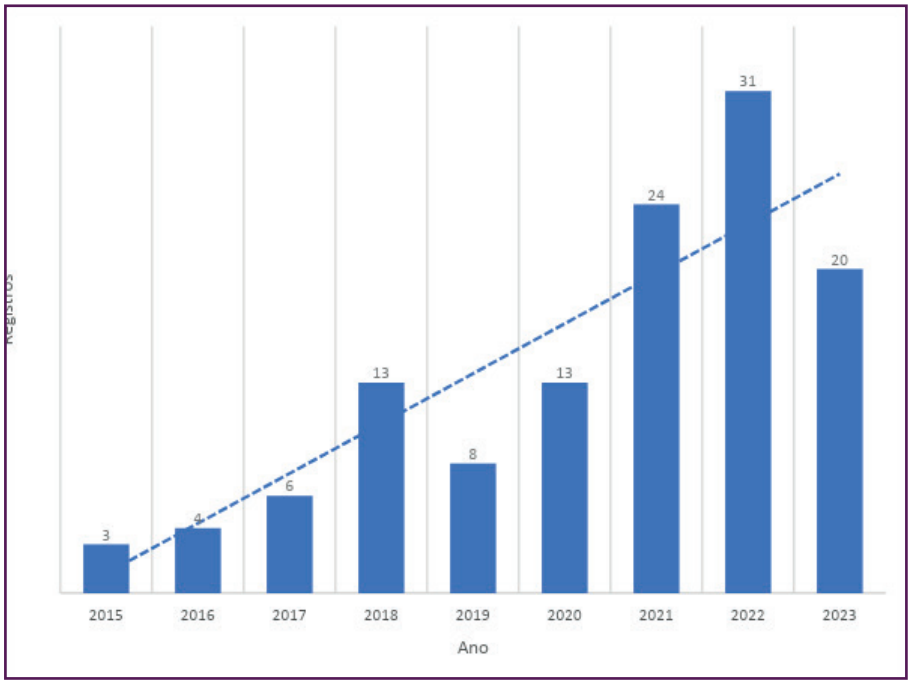
7.4 ESTUDOS UTILIZADO O FACEPAGER

Para melhor entender os usos do software em pesquisas acadêmicas, efetuamos uma consulta à base de dados *Scopus*, em setembro de 2023, em busca de artigos que citassem o *Facepager*. Por se tratar de um termo

87 Disponível em: https://www.crowdtangle.com/. Acesso em: 1 out. 2023.

bastante específico, optamos por fazer a consulta simples "REF(facepager)" com a qual obtivemos 122 registros (artigo, conferência, capítulo de livro etc.) na base. Apenas em 2019 os autores passaram também a sugerir a citação de um artigo que descreve a ferramenta. Com isso, uma busca pela citação do nome do software nos pareceu mais adequado para este levantamento. A Figura 2 apresenta o número de registros por ano encontrados na amostra.

Figura 2 - Número de registros por ano para consulta *REF(facepager)* na *Base Scopus.*



Fonte: Elaborado pelo autor (2023).

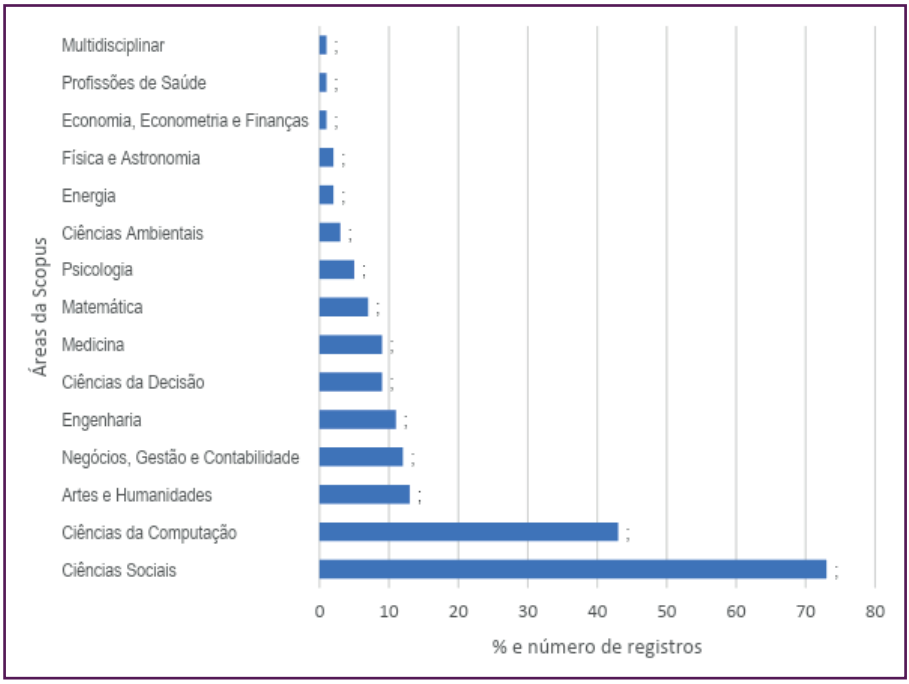
Podemos perceber que o uso do software tem sido crescente com um ponto acima do esperado em 2018. Não surpreende também que o ano de 2023, por se tratar de uma coleta incompleta, esteja abaixo do ano anterior.

A *Scopus* atribui uma ou mais áreas de estudo aos seus registros baseada em onde ele é publicado. As duas principais áreas para a amostra são a de Ciências Sociais (59,8% ou 73) e a de Ciências da Computação (35,2% ou

43). A Figura 3 apresenta os diferentes percentuais e quantitativos para as 15 áreas atribuídas pela *Scopus* aos 122 registros encontrados.

Esses resultados indicam a forte vocação do *software* para os estudos de métodos digitais que focam principalmente no campo das Ciências Sociais com alguma interface com a computação. Outrossim, "concebidas em uma era de escassez, as ciências sociais entram em uma época de abundância" com os dados e métodos digitais (Venturini; Latour, 2019, p. 43). E esta abundância se reflete na sua centralidade no uso de *softwares* de levantamento de dados de mídias sociais para estudos em métodos digitais.

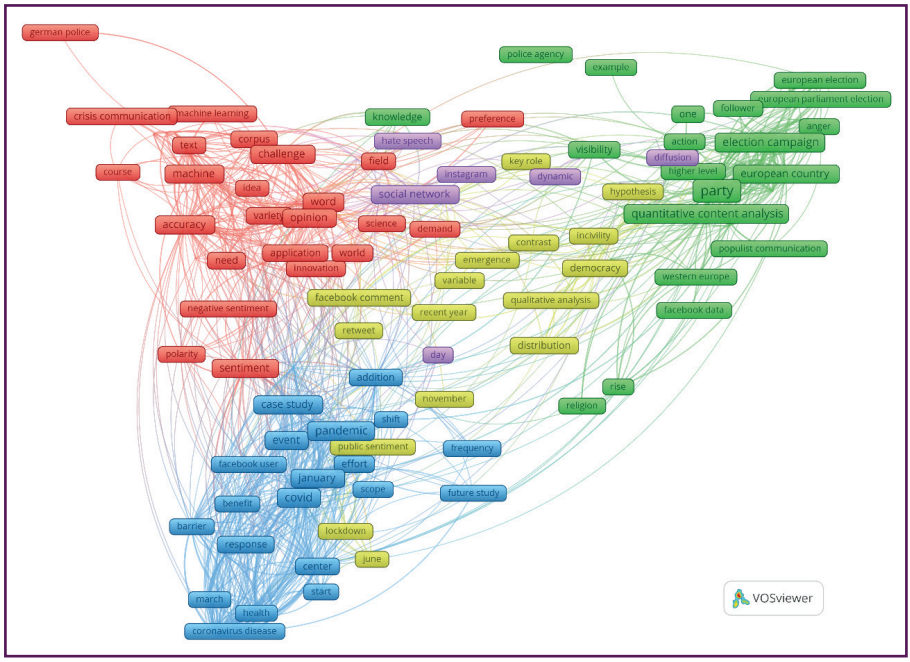
Figura 3 - Áreas atribuídas pela *Base Scopus* aos registros para consulta *REF(facepager).*



Fonte: Elaborado pelo autor (2023).

De posse destes registros, geramos um grafo com o *software VOSviewer* (Van Eck; Waltman, 2010), versão 1.6.19[88] para a co-ocorrência de termos nos títulos e resumos dos registros. A contagem dos termos foi binária, ou seja, considerou-se apenas a presença ou ausência do termo no par formado por título e resumo. Optamos também, para uma melhor visualização, por um corte inicial de termos com ao menos três ocorrências e apresentação dos 150 com maior *score* de relevância segundo o método de pontuação do *software*. O grafo resultante com a rede de relações entre os termos está apresentado na Figura 4 onde as cores representam os agrupamentos formados a partir do uso do método de modularidade aplicado pelo *software* sem alterações por parte do autor.

Figura 4 - Grafo de co-ocorrência de termos nos títulos e resumos de registros encontrados na *Base Scopus* para a busca por *REF(facepager).*

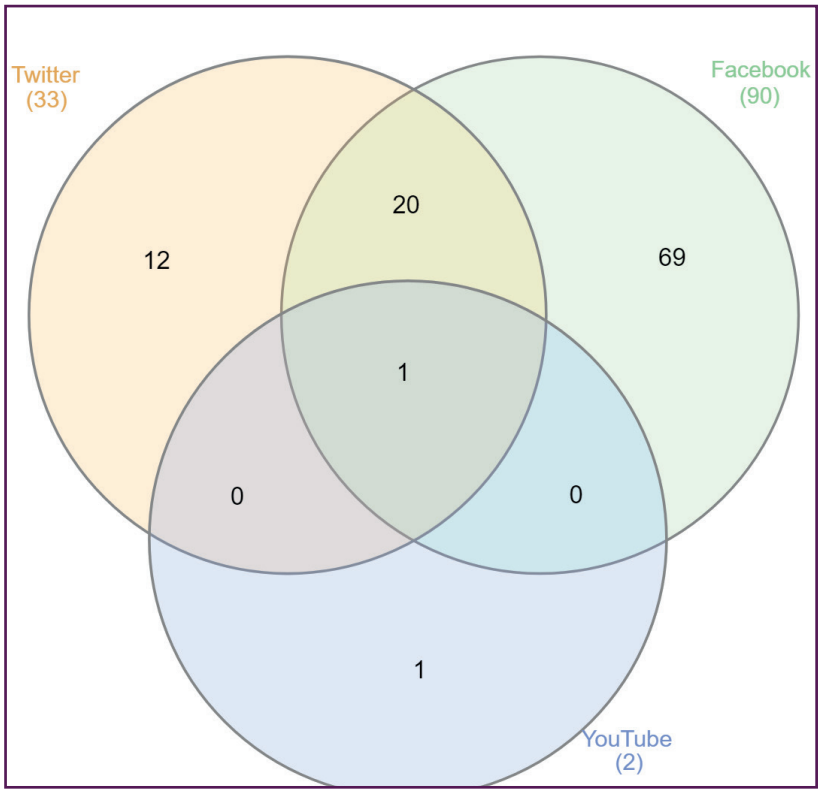


Fonte: Elaborado pelo autor (2023).

88 Disponível em: http://www.vosviewer.com. Acesso em: 1 out. 2023.

Pode-se perceber que os grupos se distribuem entre temas como pandemia e covid (agrupamento azul), eleições na Europa (agrupamento verde), aplicações de estudos de *corpus textuais* (agrupamento vermelho) e os agrupamentos em amarelo e lilás situados mais no meio do grafo onde se encontram termos mais gerais ligados às *affordances* analisadas (comentários do *Facebook* e *retweets*) e seus métodos ou características (análises qualitativas e distribuição). Uma versão interativa deste grafo foi gerada e está disponível para acesso[89]. Em relação às fontes de dados, a grande maioria dos registros fazem menção ao *Facebook* (90 ou 73,8%) e próximo a um quarto deles ao *Twitter*, hoje renomeado para *X*, (33 ou 27,0%) (Figura 5).

Figura 5 - Diagrama de Venn das fontes de redes sociais citadas nos resumos dos registros para busca *REF(facepager)* na *Base Scopus*



Fonte: Elaborado pelo autor (2023).

89 Disponível em: VOSviewer Online. Acesso em: 1 out. 2023.

Podemos observar as sobreposições quanto à fonte de dados no *diagrama de Venn* da Figura 5 feito no *Interactivenn*[90] (Heberle *et al.*, 2015), onde também elencamos os estudos que citaram o *YouTube*. Dos 122 registros, 20 (16,4%) fazem referência ao *Twitter* e ao *Facebook*, e um às duas redes e, também, ao *YouTube*. O *Facebook* foi fonte de pesquisa única em mais da metade dos registros encontrados (69 ou 56,6%), contra 12 (9,8%) estudos focados apenas no *Twitter* e apenas um (0,8%) no *YouTube*.

Além dessas menções objetivas às mídias sociais representadas no menu[91] do *software*, 18 (14,8%) registros fizeram menção em seu resumo à dados de redes sociais, sem especificar qual foi objeto de sua coleta, com casos pontuais elencando *sites* específicos da *web*.

7.5 CONSIDERAÇÕES FINAIS

Apesar de todo o potencial que a ferramenta apresenta, seu uso é ainda relativamente restrito, sendo destaque o fato de o software ser majoritariamente utilizado por autores da Alemanha, país onde foi desenvolvida a aplicação. A autoria de pesquisadores afiliados a instituições alemãs representa 27,0% (ou 33 registros) de todas as produções encontradas na *Scopus*, sendo mais de três vezes superior ao segundo país de afiliação colocado, a Áustria. A atenção a demandas dos usuários do software pelo desenvolvedor é constante, o que justificaria a sua escolha como plataforma de coleta de dados, porém, surpreendentemente vemos que os pesquisadores optam ou por soluções mais customizadas como *scripts* em *Python* ou *R*, ou por serviços *online* na sua modalidade de teste ou com pagamento para altos volumes de dados.

90 Disponível em: http://www.interactivenn.net. Acesso em: 1 out. 2023.

91 Não foram encontradas menções para "Amazon", a quarta e última fonte especificada no menu do *Facepager*.

Um fator que pode estar sendo determinante é que o *Facepager* não integra as iniciativas do *DMI*[92] (Digital Methods Initiative). Entretanto, ele já foi apresentado em um evento de *Sprint de Dados* (#SMARTDataSprint2021) em Lisboa, Portugal, na Universidade Nova de Lisboa, evento irmão dos *Sprints de Dados* da Universidade de Amsterdam e que conta com a presença de diversos pesquisadores do campo dos métodos digitais da Europa.

O fato de grande parte dos estudos utilizando o *Facepager* serem focados no *Facebook* pode ser uma consequência de seu nome que o associa mais facilmente à rede social da *Meta*. Por outro lado, a mudança na política de coleta de dados pelo *X* pode também levar ao desinteresse por se utilizar esta fonte de dados, independentemente da manutenção do acesso via autenticação pelo *Facepager*. Quanto ao *YouTube*, a existência do *YouTube Data Tools*[93] da iniciativa *DMI* acaba por se estabelecer como a principal opção para os estudos na plataforma, apesar das limitações que um levantamento executado dentro de um navegador pode ter. Em casos de necessidade de buscas mais robustas, o *Facepager* pode ser uma opção melhor, sem, no entanto, já oferecer um tratamento de dados como o *YouTube Data Tools* fornece.

Por fim, o *Facepager* tem características que atendem melhor os pesquisadores que desejam a geração de uma base de coleta robusta, organizada e sem necessitar de um conhecimento dos processos de autenticação via *API* (Application Programming Interface), parâmetros de consulta e paginação dos resultados. Acreditamos que esta é uma oportunidade de divulgação de suas características e potencialidades para pesquisadores brasileiros, quem sabe levando a uma nova geração de estudos com métodos digitais fazendo uso deste *software*.

92 Nome de um dos mais influentes grupos de pesquisa sobre uso de métodos digitais da Europa com ferramentas listadas em: https://wiki.digitalmethods.net/. Acesso em: 1 out. 2023.

93 Disponível em: https://labs.polsys.net/tools/youtube/. Acesso em: 1 out. 2023.

REFERÊNCIAS

BRUNS, A. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. **Information, Communication & Society**, London, v. 22, n. 11, p. 1544-1566, Sept. 2019.

FETTERMAN, D. **Facebook development platform launches**. August 14th, 2006. Disponível em: https://www.facebook.com/notes/2207512130. Acesso em: 3 out. 2023.

GOUVEIA, F. C. Novos caminhos e alternativas para a Webometria. **Em Questão**, Porto Alegre, v. 18, n. 3, p. 249-261, 2012.

HEBERLE, H.; MEIRELLES, G. V.; SILVA, F. R.; TELLES, G. P.; MINGHIM, R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. **BMC Bioinformatics**, London v. 16, n. 1, p. 169, May 2015.

JÜNGER, J.; KEYLING, T. **Facepager**: an application for automated data retrieval on the web. 2019. Disponível em: https://github.com/strohne/Facepager/. Acesso em: 3 out. 2023.

LOMBORG, S.; BECHMANN, A. Using APIs for Data Collection on Social Media. **The Information Society**, New York, v. 30, n. 4, p. 256-265, Aug. 2014.

MOZELLI, R. Twitter: API cara impede uso para pesquisas acadêmicas. **Olhar Digital**, 1 jun. 2023. [*Online*]. Disponível em: https://olhardigital.com.br/2023/06/01/internet-e-redes-sociais/twitter-api-cara-impede-uso-para-pesquisas-academicas/. Acesso em: 3 out. 2023.

OMENA, J. J. O que são métodos digitais? *In*: OMENA, J. J. **Métodos Digitais:** Teoria-Prática-Crítica. Lisboa: ICNOVA, 2019. p. 5-15.

VAN ECK, N. J.; WALTMAN, L. Software survey: VOSviewer, a computer program for bibliometric mapping. **Scientometrics**, Dordrecht, v. 84, n. 2, p. 523-538, Aug. 2010.

VENTURINI, T.; LATOUR, B. O tecido social: rastros digitais e métodos quali-quantitativos *In*: OMENA, J. J. **Métodos Digitais:** Teoria-Prática-Crítica. Lisboa: ICNOVA, 2019. p. 37-46.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J. W.; SANTOS, L. B. da S.; BOURNE, P. E.; BOUWMAN, J.; BROOKES, A. J.; CLARK, T.; CROSAS, M.; DILLO, I.; DUMON, O.; EDMUNDS, S.; EVELO, C. T.; FINKERS, R.; GONZALEZ-BELTRAN, A.; GRAY, A. J. G.; GROTH, P.; GOBLE, C.; GRETHE, J. S.; HERINGA, J.; HOEN, P. A. C.; HOOFT, R.; KUHN, T.; KOK, R.; KOK, J.; LUSHER, S. J.; MARTONE, M. E.; MONS, A.; PACKER, A. L.; PERSSON, B.; ROCCA-SERRA, P.; ROOS, M.; VAN SCHAIK, R.; SANSONE, S. A.; SCHULTES, E.; SENGSTAG, T.; SLATER, T.; STRAWN, G.; SWERTZ, M. A.; THOMPSON, M.; VAN DER LEI, J.; VAN MULLIGEN, E.; VELTEROP, J.; WAAGMEESTER, A.; WITTENBURG, P.; WOLSTENCROFT, K.; ZHAO, J.; MONS, B. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, London, v. 3, n. 1, p. 160018, Mar. 2016.

DADOS DO AUTOR:

Fábio Castro Gouveia

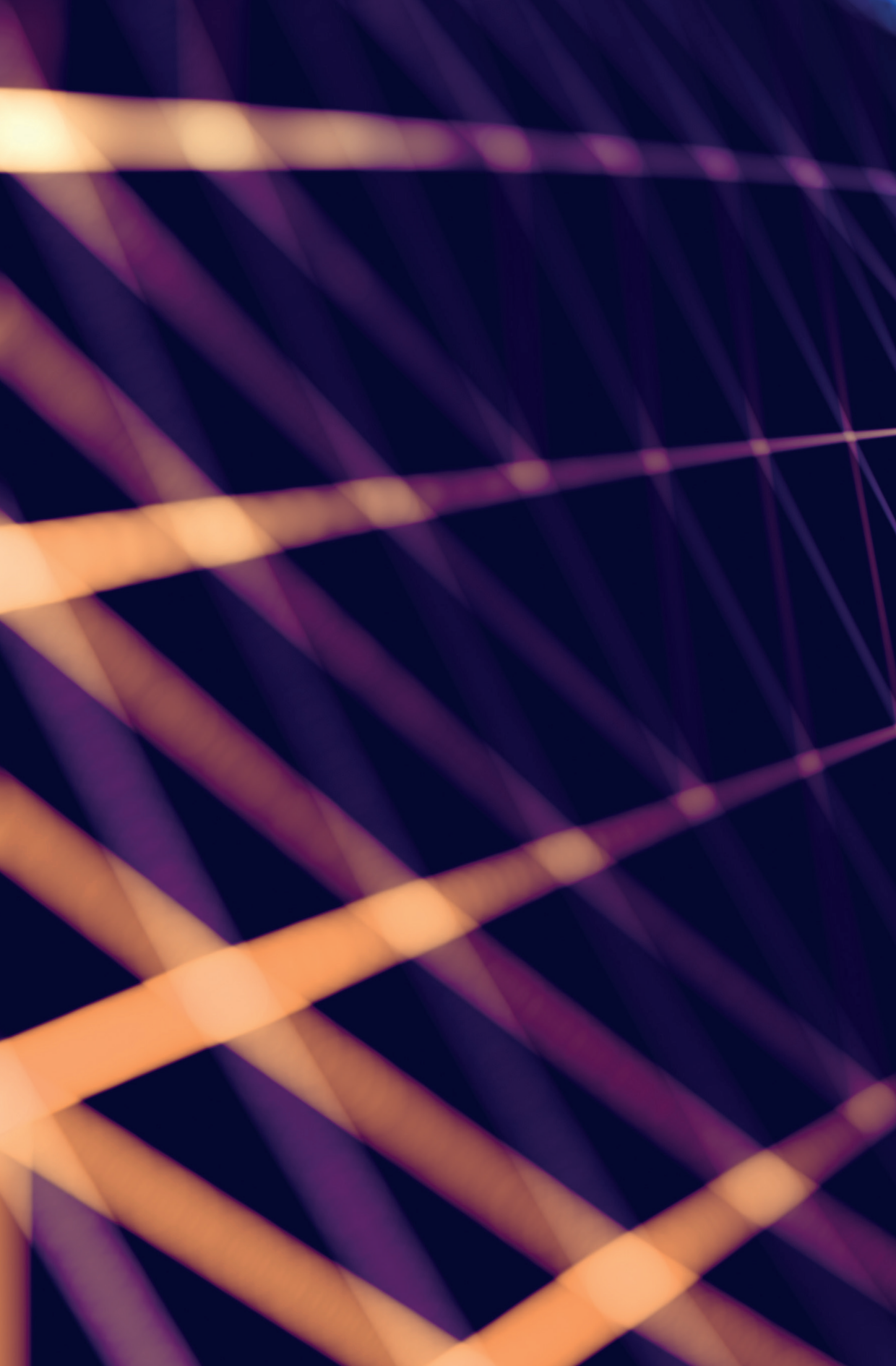
Fábio Castro Gouveia é Tecnologista em Saúde Pública da Fundação Oswaldo Cruz no Brasil cedido para o Instituto Brasileiro de Informação em Ciência e Tecnologia IBICT. Gouveia é Biólogo, mestre em Microbiologia e Imunologia e doutor em Química Biológica (Educação, Gestão e Difusão de Biociências). Ele fez um pós-doutorado curto como Visiting Fellow da Katolieke Universiteit Leuven (Bélgica). É docente permanente do Programa de Pós-Graduação em Ciência da Informação do IBICT/Eco-UFRJ e do Mestrado em Divulgação da Ciência, Tecnologia e Saúde da Fiocruz. Gouveia desenvolve pesquisas na área

de Ciência da Informação, com ênfase em Estudos Métricos da Informação (Cientometria, Webometria, Altmetria e Indicadores de Ciência, Tecnologia e Inovação), Métodos Digitais, STS, Data Science e Tecnologia Blockchain, e na área de Divulgação Científica e Comunicação em Saúde, com ênfase em estudos sobre internet e mídias sociais.

https://orcid.org/0000-0002-0082-2392
fgouveia@gmail.com

Como referenciar o capítulo 7:

GOUVEIA, Fábio Castro. Facepager: uma ferramenta de extração e raspagem de dados de código aberto. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 7. p. 209-223. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap7.



8. OPENREFINE COMO FERRAMENTA PARA TRATAMENTO DE REGISTROS BIBLIOGRÁFICOS

Tiago Rodrigo Marçal Murakami
Ingrid Torres Schiessl
Diego José Macêdo
Milton Shintaku

8.1 INTRODUÇÃO

Bibliotecas são cenários férteis para pesquisas científicas, pois são consideradas unidades de informação milenares que atuam de forma transversal e estão presentes nos mais diversos contextos. Nesse sentido, podem ser fonte de pesquisas em várias disciplinas, como a arqueologia com as bibliotecas do mundo antigo, a ciência da computação com o uso de novas tecnologias para atendimento às necessidades informacionais, entre outras. Entretanto, é na biblioteconomia e na ciência da informação que se fomenta a maior parte dos estudos, com pesquisas sobre os catálogos, usuários, formação de coleções, evolução etc.

Nesse ponto, métricas foram criadas para estudar a informação gerenciada pela biblioteca, como a métrica basal *bibliometria* e suas derivações como a *cientometria*, *infometria* e as mais novas *webometria* e *altimetria*. Com as métricas, com o uso forte dos preceitos da estatística, é possível obter indicadores sobre o uso da informação em canais tradicionais como os livros, periódicos e anais de eventos e dos canais mais novos como os *sites* e *mídias sociais*.

A biblioteca como objeto de pesquisa tem ampla possibilidade de estudos, desde o seu papel social na formação dos cidadãos, em si só ampla e complexa, até estudos pontuais como os estudos de caso. Em muito, pela grande possibilidade da tipologia e transversalidade da biblioteca, podendo ser pública ou restrita, educacional em todos os níveis, voltada ao atendimento ao público adulto ou infantil, ser especializada ou de

acervo amplo, ou seja, as múltiplas possibilidades revelam terreno fértil para investigações.

Entre as possibilidades de pesquisas científicas em bibliotecas, repousa sobre o seu catálogo, composto por registros bibliográficos, não apenas obras físicas como também digitais. Hancock-Beaulieu (1990), por exemplo, estudando sobre o comportamento de busca dos usuários, revelava desafios e problemas, principalmente quanto ao assunto. Nesse ponto, são criadas oportunidades de estudos em várias áreas, como para a indexação e criação de vocabulários controlados, criação de novas tecnologias para a recuperação da informação, em letramento informacional para uso de ferramentas informatizadas e outros. Todos esses estudos são baseados no catálogo, na sua formação, representação e uso.

Nesse contexto, as bibliotecas, com seu catálogo formado por registros bibliográficos, podem ser alvos de pesquisas, mas requerem ferramentas que possam atuar com as suas peculiaridades. Grande parte dos catálogos utiliza o padrão *Machine Readable Cataloging* (MARC) para registros das informações. Por isso, exige ferramentas que consigam processar essa tipologia de informação em toda a sua complexidade. Entre as tecnologias existentes está o *OpenRefine*, opção pela simplicidade, flexibilidade e robustez.

8.2 REGISTROS BIBLIOGRÁFICOS

Na composição dos catálogos de bibliotecas, os registros bibliográficos desempenham um papel central. Conforme definido por Cunha e Cavalcanti (2008, p. 313), esses registros têm três aspectos fundamentais: primeiro, são armazenados em formatos informatizados e contêm informações bibliográficas que descrevem um ou mais segmentos de registro; segundo, constituem coleções de itens relacionados, tratados como uma unidade e fixados em suporte automatizado; terceiro, nas bases de dados bibliográficos, esses registros substituem ou representam artigos, livros ou outras formas documentais.

A origem da palavra "registros" remonta ao latim "*registrum*", que significa uma lista ou catálogo de coisas registradas. Com o tempo, essa palavra foi adotada em várias línguas para denotar a ação de registrar ou documentar

informações. Por outro lado, a palavra "bibliográfico" tem suas raízes no grego antigo, com "*biblion*" (livro) e "*grapho*" (escrever), referindo-se, assim, a tudo relacionado à escrita e ao estudo de livros e documentos. Quando combinadas, essas duas palavras formam o termo "Registro Bibliográfico", que reflete a essência da atividade de documentação e catalogação de materiais, sejam eles impressos ou digitais, em bibliotecas e sistemas de informação.

Compreender o significado dos registros bibliográficos exige uma análise da evolução da catalogação ao longo da história. A criação de registros bibliográficos está intrinsecamente vinculada à prática da catalogação em bibliotecas. Ao longo do tempo, foram desenvolvidas diretrizes e regras para orientar esse processo, como evidenciado em estudos de Ferraz (1991), Fiuza (1987), Machado e Zafalon (2020), Mey (2003) e Selbach *et al.* (2020). Diante do avanço tecnológico, novas regras e diretrizes continuam a ser elaboradas para a catalogação. Isso resultou na evolução dos registros bibliográficos, que se adaptaram às mudanças na forma como as informações são registradas e acessadas nas bibliotecas e sistemas de informação.

Em cenários de migração de sistemas, o tratamento dos registros bibliográficos desempenha um papel crucial. Isso se deve à capacidade de adequação às novas regras de catalogação e aos novos modelos conceituais e recursos tecnológicos que permeiam o ambiente da biblioteca. Isso se alinha com a quinta lei de Ranganathan, que postula que "A biblioteca é um organismo em crescimento" (Ranganathan, 2009, p. 241). De acordo com essa perspectiva, a biblioteca como instituição deve acompanhar as tendências da sociedade em aspectos sociais, educacionais, econômicos, tecnológicos e políticos para evoluir de maneira satisfatória e continuar a atender às necessidades de seus usuários.

8.3 SOBRE O OPENREFINE

O *OpenRefine* foi criado pela *Metaweb Technologies, Inc.*, com base no produto originalmente escrito e concebido por David Huynh, voltado para limpeza e transformação de registros. Posteriormente, a *Metaweb Technologies, Inc.* foi adquirida pelo *Google, Inc*. Criado originalmente como *Freebase Gridworks*, em Julho 2010, foi renomeado para *Google Refine*. Em outubro de 2012, foi

novamente renomeado para *OpenRefine*, em sua transição para um produto suportado pela comunidade (OpenRefine, 2023) .

O *OpenRefine* é uma ferramenta de código aberto projetada para ajudar na *limpeza*, *transformação e reconciliação de dados*, oferecendo *interface amigável* e uma ampla gama de recursos que podem ser especialmente úteis para bibliotecários e profissionais de informação ao lidar com a migração de dados bibliográficos.

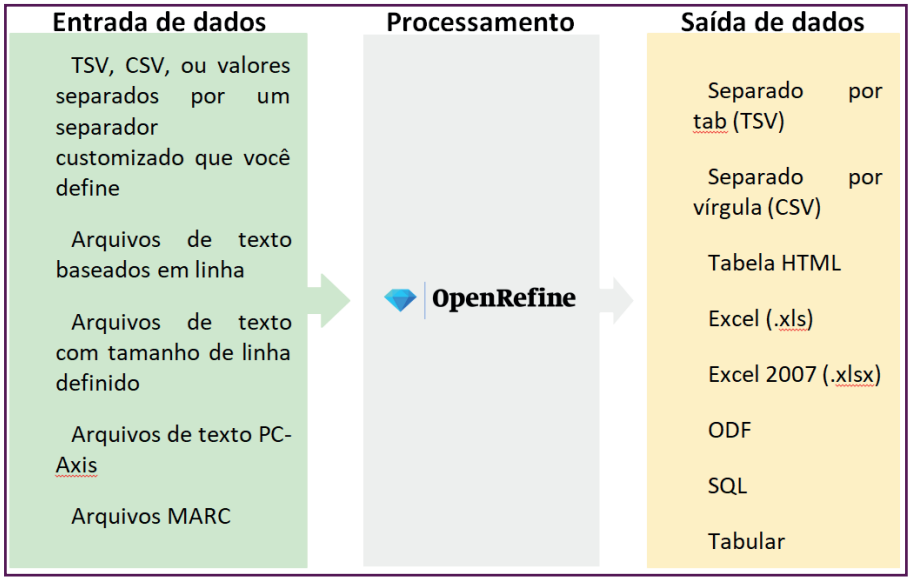
Verborgh e DeWilde (2013) relatam que pela quantidade de dados existentes, parte sem muita organização, o *OpenRefine* oferta uma grande gama de funcionalidades voltadas para o tratamento desses dados, como limpeza para corrigir problemas, remover falhas e corrigir despadronização. Com isso, pode ser utilizado para preparar dados para análises automatizadas ou não, principalmente quando se tem uma grande quantidade, como no caso dos registros bibliográficos de uma ou mais bibliotecas.

O *OpenRefine* é um *software livre* e pode ser baixado gratuitamente no *site oficial*[94] para ser instalado nos sistemas operacionais mais comuns, como o *Windows*, *Mac* e *Linux*. Na sua instalação, o *OpenRefine* possui dependências do *Java Runtime Environment* (JRE), um ambiente para execução de aplicativos escritos em *Java*. A sua instalação é simples, baseada em pacotes a serem baixados, diferentes para cada tipo de sistema operacional, descompactados e instalados por comando.

Seu funcionamento é semelhante a *páginas web*, com funcionamento simples, em um fluxo de entrada de dados, processamento e saída de dados, como apresentado na Figura 1. Para tanto, pode receber dados em uma grande variedade de formatos, incluindo conexões com gerenciados de banco de dados. Da mesma forma, pode ter os dados extraídos em vários formatos, facilitando o uso posterior em análise ou importação em outras ferramentas. O processamento requer a aplicação de fórmulas que processam os dados.

94 Disponível em: https://openrefine.org/. Acesso em: 27 set. 2023.

Figura 1 - Fluxo de entrada e saída de dados no *OpenRefine*.



Fonte: Elaboração dos autores (2023).

A versatilidade do *OpenRefine* repousa nas grandes possibilidades de entrada e saída de dados e, principalmente, nas possibilidade de aplicação das fórmulas para o tratamento deles. Sendo assim, as fórmulas conseguem processar em todos os dados da linha, escritas por meio de linguagem *General Refine Expression Language* (GREL), em *Jython* (implementação da linguagem *Python* para *Java*) e em *Clojure*.

8.4 EXEMPLO DO USO DO OPENREFINE PARA MANIPULAÇÃO DE REGISTROS BIBLIOGRÁFICOS

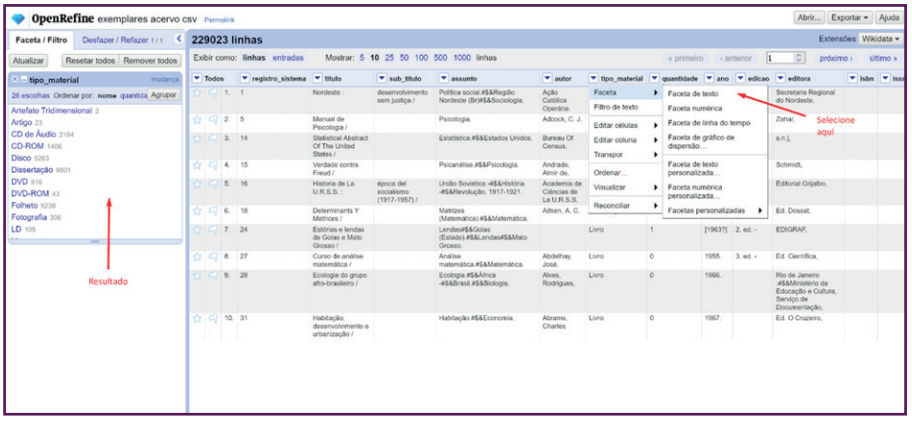
A utilização do *OpenRefine* no contexto de processamento de registros bibliográficos pode ser dividida em duas grandes etapas:

1. **Limpeza de Dados**: Muitas vezes, os registros bibliográficos existentes podem estar sujeitos a inconsistências, erros de digitação e formatos variados. O *OpenRefine* permite a detecção e correção eficaz desses problemas, garantindo consistência para os registros bibliográficos.

2. **Transformação de Dados**: Alguns *softwares* para análise de registros bibliográficos podem os requerer em um determinado padrão. O *OpenRefine* permite que os bibliotecários transformem facilmente os dados para atender a esses requisitos, seja na formatação de datas, padronização de campos ou qualquer outra necessidade específica.

A seguir, apresenta-se algumas das funcionalidades da ferramenta. A Figura 2 apresenta a funcionalidade de geração automática de facetas, tal recurso envolve a criação de categorias ou grupos com base nos valores de uma coluna de dados específica, facilitando o tratamento dos dados.

Figura 2 - Geração de facetas.



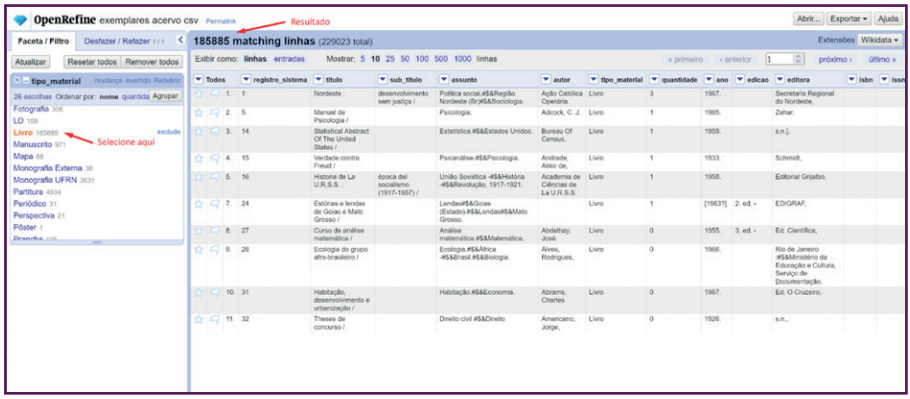
Fonte: Captura de tela (2023).

As facetas permitem organizar e analisar dados de maneira mais eficaz, especialmente quando você tem uma grande quantidade de informações não estruturadas. Por exemplo, imagine que você tem uma coluna com nomes de autores de livros em um conjunto de dados. Você pode criar facetas a partir dessa coluna para agrupar os autores por sua primeira letra inicial, facilitando a navegação e a pesquisa. Isso pode ser útil para bibliotecários e pesquisadores, pois ajuda na organização e recuperação de informações.

No *OpenRefine*, você pode criar facetas usando as operações de transformação, como dividir valores em partes ou aplicar expressões regulares para extrair informações específicas de uma coluna. Isso permite que você crie uma estrutura mais organizada para seus dados, o que pode ser valioso na área de Ciência da Informação e em tarefas de tratamento de informações.

Outra funcionalidade é a criação de filtros, conforme apresentado na Figura 3. Os filtros são ferramentas que permitem selecionar e exibir um subconjunto específico de dados com base em critérios definidos. Eles são úteis para analisar, limpar e transformar dados de maneira mais eficiente, permitindo que você foque apenas nos dados relevantes para suas tarefas.

Figura 3 - Geração de filtros.



Fonte: Captura de tela (2023).

Existem diferentes tipos de filtros disponíveis no *OpenRefine*, incluindo:

- **Filtro de Texto**: Esse tipo de filtro permite que você pesquise e filtre dados com base em valores de texto em uma coluna. Você pode procurar por palavras-chave ou usar expressões regulares para refinar os resultados.

- **Filtro Numérico**: Usado para filtrar dados em colunas numéricas com base em valores específicos, como intervalos numéricos, valores maiores ou menores que um número dado etc.

- **Filtro de Data**: Esse tipo de filtro é útil quando você lida com datas. Permite que você filtre dados com base em datas específicas, intervalos de datas e muito mais.

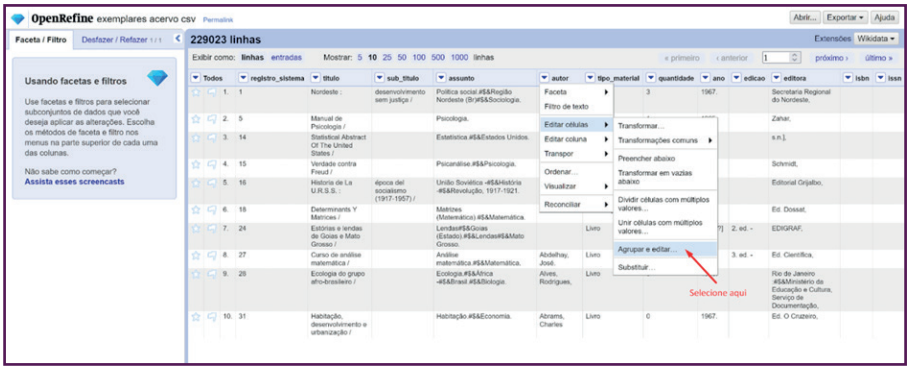
- **Filtro de Faceta de Texto**: Como mencionado anteriormente, as facetas são categorias criadas a partir dos valores de uma coluna. Você pode usar filtros de facetas de texto para selecionar rapidamente grupos de dados relacionados a essas categorias.

- **Filtro Personalizado**: Além dos filtros padrão mencionados acima, você também pode criar filtros personalizados usando expressões *GREL* (*General Refine Expression Language*) para aplicar condições específicas aos seus dados.

Os filtros são uma parte essencial do processo de limpeza e transformação de dados no *OpenRefine*. Eles ajudam a focar nos dados relevantes, permitindo que você realize operações específicas em subconjuntos de dados, incluindo edições em lote, o que é útil para bibliotecários e pesquisadores que trabalham com tratamento de informações em Ciência da Informação.

A terceira funcionalidade é chamada de *Clusterização*. Tal termo não é uma palavra da língua portuguesa. Ela é derivada do termo em inglês "*clustering*", que se refere ao processo de agrupar ou criar *clusters* em análises de dados. No *OpenRefine*, essa funcionalidade permite identificar e agrupar automaticamente valores semelhantes em uma coluna de dados, conforme Figura 4.

Figura 4 - Funcionalidade de *Clusterização*.



Fonte: Captura de tela (2023).

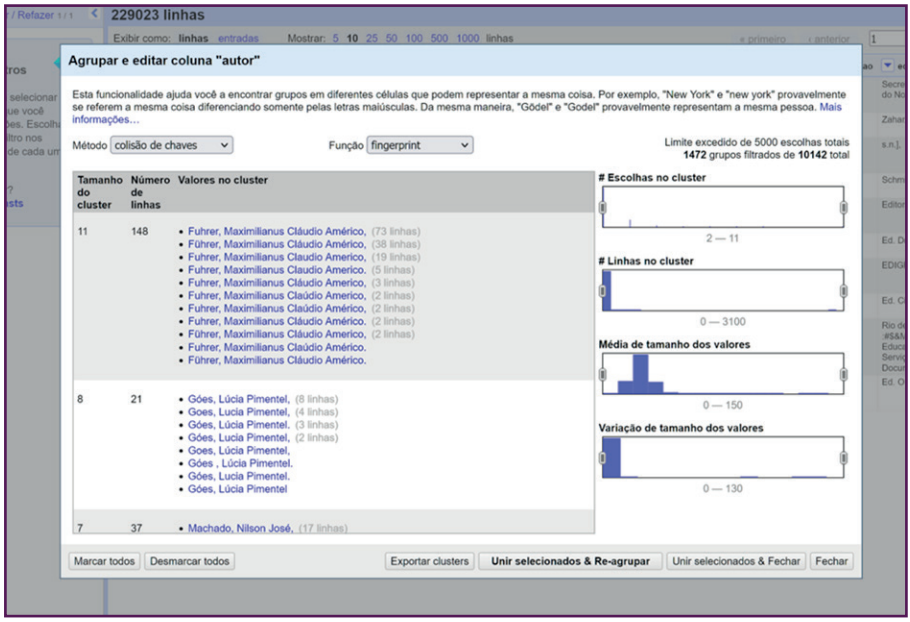
Essa funcionalidade funciona da seguinte forma:

1. **Identificação de valores semelhantes**: O *OpenRefine* analisa os valores únicos em uma coluna e identifica aqueles que são semelhantes com base em critérios como ortografia, distância de edição e outros métodos de comparação.

2. **Agrupamento**: Os valores semelhantes são agrupados automaticamente em *clusters*. Cada *cluster* representa um grupo de valores que são considerados equivalentes com base nos critérios de semelhança definidos.

3. **Revisão e Fusão**: Após a identificação dos *clusters*, você tem a oportunidade de revisar os agrupamentos propostos e, se necessário, fundir valores de diferentes *clusters* em um único valor corrigido.

Figura 5 - Funcionalidade de Clusterização aplicada à coluna "autor".

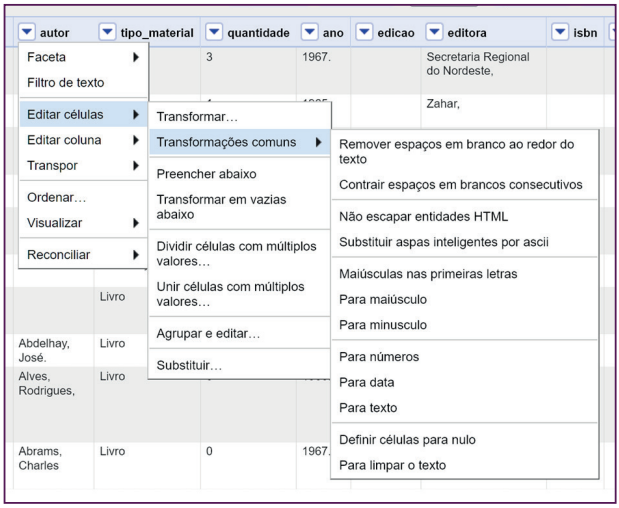


Fonte: Captura de tela (2023).

Essa funcionalidade é especialmente útil para tarefas de limpeza de dados, na qual você pode ter variações nos nomes de autores, por exemplo, devido a erros de digitação ou abreviações. O *clustering* ajuda a consolidar esses valores, tornando o processo de limpeza mais eficiente.

Existem outras funcionalidades que ajudam a edição das células, como Transformações comuns em células, conforme a Figura 6.

Figura 6 - Função de transformação comum de células.



Fonte: Captura de tela (2023).

Essa funcionalidade permite aplicar transformações simples e comuns aos dados de uma célula. Tais transformações são frequentemente usadas para limpar, padronizar ou formatar os dados de maneira consistente, por exemplo, as entradas de autoria ou assuntos. Alguns exemplos de transformações comuns que podem ser aplicadas usando essa função incluem:

- Para maiúsculo: Converte todo o texto em maiúsculas.
- Para minúsculo: Converte todo o texto em minúsculas.
- Maiúsculas nas primeiras letras: Torna a primeira letra de cada palavra maiúscula e as demais em minúsculas.
- Remover Espaços em Branco: Remove espaços em branco extras no início ou no final de um valor.
- Substituir Texto: Substitui um valor específico por outro em toda a coluna.
- Editar Célula: Permite editar o conteúdo de uma célula individualmente.

- Dividir células com múltiplos valores: Divide uma coluna em várias colunas com base em um separador, como vírgulas.
- Unir células com múltiplos valores: Mescla os valores de várias colunas em uma única coluna.
- Para data: Formata datas de acordo com um padrão específico.
- Para números: Converte valores de texto em números.
- Para células em branco: Preenche toda a seleção para valores em branco.

A função "Transformações comuns" é uma ferramenta poderosa para a limpeza e a preparação de dados no *OpenRefine*, tornando mais fácil garantir que os dados estejam em um formato adequado para análises subsequentes. Ela oferece uma série de opções para realizar essas transformações de maneira eficiente e consistente.

Além da função mencionada anteriormente, também há outras edições possíveis nas células, como apresentado na Figura 7.

Figura 7 - Função de edição das células.

Fonte: Captura de tela (2023).

Essas funcionalidade são:

- Preencher abaixo: Essa funcionalidade permite preencher valores em branco em uma coluna com base no valor da célula acima. Isso é útil para preencher lacunas em uma coluna com valores repetidos ou sequenciais.

- Transformar em vazias abaixo: Com essa função, você pode limpar o conteúdo das células abaixo de uma célula selecionada. Normalmente utilizado para excluir dados repetidos em células, a fim de facilitar o tratamento em outras colunas.

- Dividir células com múltiplos valores: Essa funcionalidade é útil quando você tem células que contêm múltiplos valores separados por um delimitador, como vírgulas. Ela permite dividir esses valores em células individuais ou em várias colunas diferentes.

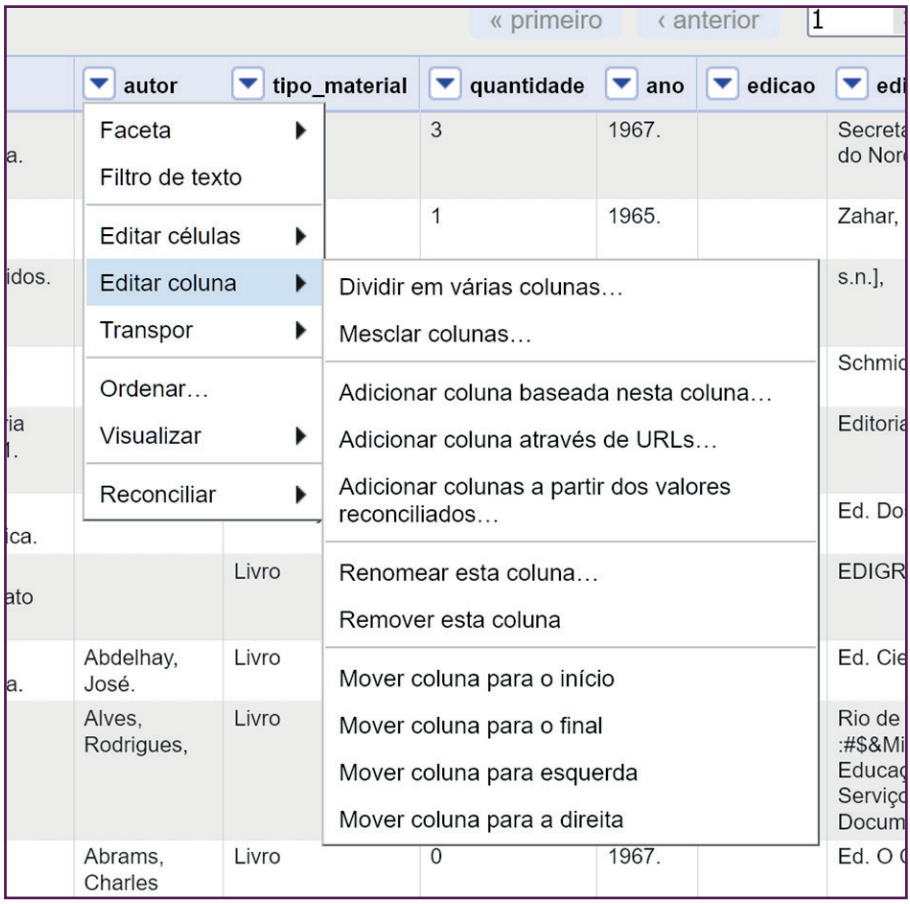
- Unir células com múltiplos valores: Ao contrário da divisão, essa funcionalidade permite unir células com múltiplos valores em uma única célula, separando os valores por um delimitador específico. Isso é útil quando você deseja consolidar informações em uma única célula.

- Substituir: Com a função "Substituir", você pode encontrar e substituir valores específicos em uma coluna. Isso é útil para correções em massa ou para padronizar valores em todo o conjunto de dados.

Essas funcionalidades são valiosas para a manipulação e preparação de dados no *OpenRefine*, ajudando a garantir que seus dados estejam em um formato adequado para análises posteriores ou para atender a requisitos específicos.

Além das células, é possível realizar edição nas colunas, conforme Figura 8. É uma funcionalidade que permite criar edições personalizadas nos valores de uma coluna de dados. Ela é útil quando é preciso realizar ajustes específicos nos valores da coluna que não podem ser realizados com as Transformações Comuns.

Figura 8 - Função de edição de colunas.



Fonte: Captura de tela (2023).

Essa funcionalidades são as seguintes:

- Dividir em diversas colunas: Essa funcionalidade permite dividir os valores de uma coluna em várias colunas com base em um separador específico. É útil quando você tem valores que estão concatenados e deseja organizá-los em colunas separadas.

- Mesclar colunas: Com essa funcionalidade, você pode mesclar os valores de duas ou mais colunas em uma única coluna. É útil para combinar informações de várias fontes em uma única coluna.

- Adicionar coluna baseada nesta coluna: Essa funcionalidade permite criar uma nova coluna com base nos valores de uma coluna existente. Você pode aplicar expressões *GREL* para calcular ou transformar os valores da nova coluna com base nos valores da coluna original.

- Adicionar coluna através de *URLs*: Com essa funcionalidade, você pode criar uma nova coluna recuperando dados de uma *URL* externa. Isso é útil quando você deseja enriquecer seus dados com informações obtidas da *web*, como informações de geolocalização, cotações de moeda etc.

- Renomear esta coluna: Essa funcionalidade permite renomear uma coluna, dando-lhe um novo nome que seja mais descritivo ou adequado ao seu projeto.

- Remover esta coluna: Com essa funcionalidade, você pode remover uma coluna inteira do conjunto de dados. Isso é útil quando você tem colunas desnecessárias ou duplicadas.

- Mover: A funcionalidade "Mover" permite reorganizar a ordem das colunas no conjunto de dados. Você pode arrastar e soltar as colunas para posicioná-las onde desejar, facilitando a visualização e análise dos dados.

Essas funcionalidades são utilizadas para a limpeza, transformação e preparação de dados. Permitindo a edição dos valores de forma eficiente de acordo com as necessidades.

Outra ferramenta que auxilia na organização dos dados é a "Transpor", que permite reorganizar os dados de uma tabela trocando as linhas por colunas e vice-versa. Isso é especialmente útil quando você deseja *pivotar* ou girar a estrutura dos seus dados para uma melhor visualização ou análise. São elas:

- Transpor células de linhas para colunas: A funcionalidade permite trocar as linhas da tabela pelas colunas e vice-versa. Isso é útil quando seus

dados estão dispostos horizontalmente, mas você precisa deles na vertical ou vice-versa.

- Transpor células de colunas para linhas: Você pode escolher quais colunas deseja transpor. Isso é importante quando você deseja manter algumas colunas inalteradas e transpor apenas um subconjunto delas.

- Você pode especificar o nome da nova coluna que será criada para armazenar os valores transpostos. Isso permite dar um nome descritivo e significativo à nova estrutura de dados.

As funcionalidades mencionadas anteriormente são aplicadas de forma combinada. Durante o processo de migração, essa ferramenta é utilizada quando os dados de origem são transformados em formato tabular. Essa tabela é então formatada de acordo com as normas *MARC* e exportada em um formato tabular para a posterior migração para o formato *MARC*, utilizando uma ferramenta externa, como o *Librecat*/*Catmandu*[95] ou o *MARCEdit*[96].

Exemplos de operações realizadas com o *OpenRefine* durante o processo de migração incluem:

Conversão de codificação: O formato *MARC* requer que os dados sejam codificados em campos específicos e siga uma codificação única. Por exemplo, o país de publicação deve ser codificado como "bl" para Brasil no *MARC*. No entanto, em sistemas legados, os dados podem variar, incluindo "Brasil", "Brazil" ou até mesmo conter erros de digitação. O *OpenRefine* oferece uma solução para corrigir esses dados, permitindo a separação dos dados em uma única coluna, o uso de facetas para criar conjuntos de dados e, em seguida, a alteração dos dados para os novos códigos. Essa funcionalidade permite a correção de todas as ocorrências na coluna de forma eficiente, algo que não seria facilmente executado em outras ferramentas.

Separação de campos: Em sistemas anteriores, informações como título e subtítulo podem estar contidas no mesmo campo, separadas por um

95 Disponível em: https://librecat.org/. Acesso em: 27 set. 2023.

96 Disponível em: https://marcedit.reeset.net/. Acesso em: 27 set. 2023.

caractere como ":" (dois pontos). No entanto, o formato *MARC* requer campos específicos para o título ($245a) e subtítulo ($245b). O *OpenRefine* oferece a funcionalidade de dividir colunas por separadores, o que possibilita a criação de duas colunas distintas para o título e o subtítulo, seguindo as exigências do formato *MARC*. Isso facilita a organização e a migração dos dados de maneira precisa.

8.5 CONSIDERAÇÕES FINAIS

O *OpenRefine* é uma ferramenta valiosa para o tratamento de dados bibliográficos, principalmente, em situações de migração de sistemas de informação. Sua capacidade de limpar, transformar e reconciliar dados de forma eficiente ajuda as bibliotecas a garantirem a qualidade e consistência de seus registros. Além disso, a flexibilidade do *OpenRefine* permite que as instituições personalizem o processo de migração de acordo com suas necessidades específicas, tornando-o uma escolha eficaz para bibliotecas de diversos tamanhos e contextos.

A integração bem-sucedida do *OpenRefine* no fluxo de trabalho para tratamento de registros bibliográficos pode resultar em economia de tempo e recursos, uma vez que automatiza tarefas repetitivas e facilita a correção de erros nos registros. No entanto, é importante que as equipes envolvidas recebam o treinamento adequado para aproveitar ao máximo essa ferramenta. Em resumo, o *OpenRefine* desempenha um papel significativo na simplificação do processo de migração de registros bibliográficos para sistemas de informação, contribuindo para uma gestão mais eficiente da informação nas bibliotecas.

Por fim, o *OpenRefine* se apresenta como mais uma ferramenta que pode ser útil na etapa de tratamento de dados em pesquisas científicas que tenham como dados os registros bibliográficos ou similares. Com isso é possível juntar dados provenientes de várias fontes, assim como transformar para padrões desejados para facilitar as análises. Em alguns casos de pesquisas científicas, nem toda a coleta de dados obtém os registros no formato desejado, requerendo processamento para passar para a próxima etapa, e o *OpenRefine* pode atender a essa demanda.

REFERÊNCIA

CUNHA, Murilo Bastos da; CAVALCANTI, Cordélia Robalinho de Oliveira. **Dicionário de Biblioteconomia e Arquivologia**. Brasília, DF: Briquet de Lemos, 2008. 451 p. Disponível em: https://repositorio.unb.br/handle/10482/34113. Acesso em: 21 set. 2023.

FERRAZ, Iraneuda Maria Cardinalli. Uso do catálogo de biblioteca: uma abordagem histórica. **TransInformação**, Campinas, v. 3, n. 1/3, p. 90-114, jan./dez. 1991. Disponível em: https://brapci.inf.br/index.php/res/download/55433. Acesso em: 1 mar. 2023.

FIUZA, Marysia Malheiros. A catalogação bibliográfica até o advento das novas tecnologias. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 16, n. 1, p. 45-53, mar. 1987.

HANCOCK-BEAULIEU, Micheline. Evaluating the impact of an online library catalogue on subject searching behavior at the catalog and the shelves. **Journal of Documentation**, Bingley, v. 46, n. 4, p. 318-338, Apr. 1990. DOI 10.1108/eb026863. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/eb026863/full/html. Acesso em: 21 set. 2023.

MACHADO, Raildo; ZAFALON, Zaira. **Catalogação:** dos princípios e teorias ao RDA e IFLA LRM. João Pessoa, PB: Editora UFPB, 2020. 128 p. Disponível em: http://eprints.rclis.org/43200/. Acesso em: 21 set. 2023.

MEY, Eliane Serrão Alves. **Não brigue com a catalogação**. Brasília, DF: Briquet de Lemos, 2003. 186 p.

OPENREFINE. [2023, *online*]. Disponível em: https://openrefine.org/. Acesso em: 1 mar. 2023.

RANGANATHAN, Shiyali Ramamrita. **As Cinco Leis da Biblioteconomia**. Brasília, DF: Briquet de Lemos, 2009. 336 p.

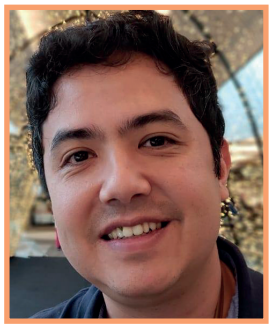
SELBACH, Clarissa Jesinska; FERREIRA, Anamaria; KERN, Lucas Martins; NOVAK, Loiva Duarte. Catalogação com Resource Description and Access (RDA): relato de experiência na Biblioteca Central Irmão José

Otão (PUCRS). **Revista ACB**: Biblioteconomia em Santa Catarina, Florianópolis, v. 25, n. 3, p. 729-733, 2020. Disponível em: https://brapci.inf.br/index.php/res/v/151853. Acesso em: 1 mar. 2023.

VERBORGH, Ruben; DE WILDE, Max. **Using OpenRefine**. Birmingham, UK: Packt Publishing, 2013.

DADOS DOS AUTORES:

Tiago Rodrigo Marçal Murakami



Tiago Rodrigo Marçal Murakami é Bibliotecário na Escola de Comunicações e Artes da Universidade de São Paulo. Graduado em Biblioteconomia pela Universidade de São Paulo.

https://orcid.org/0000-0003-1942-6434
trmurakami@usp.br

Ingrid Torres Schiessl

Ingrid Torres Schiessl é Mestre em Ciência da Informação e bacharela em Biblioteconomia pela Universidade de Brasília (UnB). Bibliotecária e pesquisadora no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

https://orcid.org/0000-0001-5815-2574
ingridschiessl@ibict.br

Diego José Macêdo

Diego José Macêdo é Mestre em Ciência da Informação pela Universidade de Brasília. Bacharel em Sistema de Informação pela Universidade Católica de Brasília. Atualmente é tecnologista do Instituto Brasileiro de Informações em Ciência e Tecnologia - Ibict.

diegomacedo@ibict.br
https://orcid.org/0000-0002-5696-0639

Milton Shintaku

Milton Shintaku é Doutor em Ciência da Informação pela Universidade de Brasília. Coordenador de Tecnologia para Informação (Cotec) do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

shintaku@ibict.br
https://orcid.org/0000-0002-6476-4953

Como referenciar o capítulo 8:

MURAKAMI, Tiago Rodrigo Marçal; SCHIESSL, Ingrid Torres; MACÊDO, Diego José; SHINTAKU, Milton. OpenRefine como ferramenta para tratamento de registros bibliográficos. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 8. p. 225-244. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap8.

9. ORANGE DATA MINING: UMA FERRAMENTA PARA INSERÇÃO DE INTELIGÊNCIA ARTIFICIAL NA PESQUISA CIENTÍFICA

Caio Saraiva Coneglian
Henrique Leal Tavares
Diego José Macedo
Milton Shintaku

9.1 INTRODUÇÃO

A presença da tecnologia nas atividades humanas remonta à aurora da espécie, com a criação de ferramentas que apoiam atividades diárias. Com a especialização das atividades, ferramentas foram criadas exclusivamente para algumas profissões, mas que muitas vezes são adaptadas para outras. Em outros casos, algumas ferramentas já nascem fadadas a serem generalistas, podendo ser utilizadas em uma grande gama de atividades e profissões. Esse último tipo é o caso dos computadores, que por serem programáveis, são flexíveis para serem utilizados em quase todas as atividades humanas.

Dentre as atividades que podem fazer uso de computadores está a pesquisa científica, principalmente as que requerem o processamento automatizado de dados e informações. Tanto que, desde os primórdios da computação, as universidades e institutos de pesquisa possuíam computadores para serem utilizados pelos seus pesquisadores. Em alguns casos, universidades construíram os seus próprios computadores, em projetos de pesquisa ousados e inovadores como no caso do conhecido "Patinho Feio", primeiro computador brasileiro desenvolvido 100% pelo Laboratório de Sistemas Digitais (LSD) do Departamento de Engenharia Politécnica da Universidade de São Paulo (USP), lançado em julho de 1972 (Cardi; Barreto, 2012).

Inicialmente os computadores, como são conhecidos por grande parte da população atual, nasceram para processar dados numéricos e estruturados,

quase como uma calculadora programável. Tanto que, muitas vezes os computadores nas pesquisas eram utilizados para realizar cálculos repetitivos com grande quantidade de números, nem sempre complexos pelas limitações tecnológicas da época. Por isso, grande parte dos estudos que utilizam computadores nos primórdios da computação era voltada para as ciências rígidas, engenharia e estatística.

Com a evolução tecnológica, os computadores adotaram novas formas de atuação, com processamento de textos, imagens, áudio e vídeo. A criação dos chamados gerenciadores de banco de dados e linguagens de programação modernas e flexíveis contribuiu para o que é denominado de dados pudessem transcender tipos e formatos. Nesse caminho, o processamento de dados pode atender a todo o tipo de objeto digital, ou seja, tudo que pode ser codificado em formato digital.

Com a mudança de século e a popularização da *internet* e *web* novas possibilidades de pesquisa científica atendem a todas as áreas de conhecimento, em praticamente todas as etapas dos estudos. Pode-se afirmar que os computadores e seu ambiente digital estão presentes desde a criação de propostas de estudos, até as publicações e uso dos resultados. Com isso, pesquisas tendem a ser efetuadas de forma mais rápida e eficaz, assim como a disseminação dos seus resultados.

O uso da computação nas ciências se tornou tão comum, que já há consenso sobre a chamada ciência virtual em contraposição às ciências naturais. Realidade virtual, projeções e simulações são comuns nas pesquisas científicas, facilitando resolver problemas. Tanto que, por meio de pesquisas é possível indicar tendências que as organizações e instituições vão seguir na computação, incluindo as ciências para um futuro próximo, *A Gartner Group*, empresa de consultoria especializada em computação, prevê que a partir de 2023, o uso da inteligência artificial será cada vez mais constante, principalmente com as chamadas aplicações em inteligência artificial adaptáveis.

9.2 INTELIGÊNCIA ARTIFICIAL NO PROCESSO DA PESQUISA CIENTÍFICA

A pesquisa científica, ao longo de décadas, tem sido conduzida com base em métodos tradicionais que envolvem coleta manual de dados, revisões literárias extensas e análises estatísticas complexas. Em especial, no âmbito das Ciências Sociais Aplicadas, há uma série de métodos de pesquisa que podem ser utilizados para o desenvolvimento de pesquisas aplicadas.

Com a evolução da tecnologia, tendo como destaque a Inteligência Artificial, foram desenvolvidas novas técnicas que podem apoiar o desenvolvimento de pesquisas, em especial utilizando dados. Técnicas como *Machine Learning*, *Text Learning* e *Data Mining* têm desempenhado um papel fundamental na transformação da pesquisa científica, proporcionando ferramentas poderosas para extrair conhecimento a partir de grandes conjuntos de dados.

O campo da Inteligência Artificial que mais tem impactado a sociedade é *Machine Learning*. Esse campo tem a capacidade de revolucionar a pesquisa científica ao permitir que computadores aprendam com dados e façam previsões mais precisas. Ademais, tal capacidade de identificar padrões complexos e gerar insights tem aplicações em diversas áreas, desde a medicina, onde auxilia na identificação de diagnósticos precisos, até a física, onde ajuda a entender fenômenos complexos.

Uma definição de *Machine Learning* é dada por Jordan e Mitchell (2015, p. 255, tradução dos autores)[97]:

> O aprendizado de máquina é uma disciplina focada em duas questões inter-relacionadas: Como construir sistemas de computador que melhoram automaticamente com a experiência? e Quais são as leis fundamentais da teoria estatística da informação computacional que governam todos

97 **Trecho original:** *Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations? The study of machine learning is important both for addressing these fundamental scientific and engineering questions and for the highly practical computer software it has produced and fielded across many applications.*

> os sistemas de aprendizagem, incluindo computadores, seres humanos e organizações? O estudo do aprendizado de máquina é importante tanto para abordar essas questões científicas e de engenharia fundamentais, quanto para o *software* de computador altamente prático que ele produziu e utilizou em vários aplicativos.

Outra vertente, que está vinculada ao *Machine Learning*, mas com foco em tratamento de texto é o *Text Learning*. Esse campo concentra-se na análise de texto escrito, tornando possível a extração de informações valiosas a partir de documentos científicos extensos. Essa abordagem é especialmente relevante em um mundo inundado de informações, onde cientistas precisam navegar por vastos repositórios de literatura científica para manter-se atualizados e encontrar pistas para suas pesquisas.

Tommasel e Godoy (2019, p. 1, tradução dos autores)[98] apontam que *Text mining*:

> [...] refere-se a um processo de descoberta de conhecimento que visa a extração de padrões interessantes e não triviais da linguagem natural. Este processo inclui múltiplas áreas, como análise de texto, processamento de linguagem natural e recuperação de informação, entre outras.

Para o autor, o *Text Mining* é um guarda chuva que se relaciona a outras técnicas voltadas a processamento de texto para extração de informação.

Por fim, *Data Mining* é uma técnica que busca padrões ocultos em grandes conjuntos de dados, oferecendo uma abordagem que apoia na identificação de *insights* e tendências nas mais diversas áreas. Garcia, Luengo e Herrera (2015, p. 1, tradução nossa)[99] apontam que

> *DM* [*Data Mining*] trata, de modo geral, de resolver problemas por meio da análise de dados presentes em bancos de dados reais. Hoje em dia,

98 Trecho original: [...] *refers to a knowledge discovery process aiming at the extraction of interesting and non-trivial patterns from natural language. This process includes multiple fields, such as text analysis, natural language processing and information retrieval, amongst others.*

99 Trecho original: *DM is, roughly speaking, about solving problems by analyzing data present in real databases. Nowadays, it is qualified as science and technology for exploring data to discover already present unknown patterns.*

> qualifica-se como ciência e tecnologia para explorar dados para descobrir padrões desconhecidos já presentes.

Partindo desses três campos, *Machine Learning*, *Text Learning* e *Data Mining*, identifica-se que a pesquisa científica nas ciências sociais aplicadas pode usufruir de novas técnicas, em especial no âmbito de pesquisas aplicadas. Assim, entra-se em detalhes sobre como essas técnicas nas próximas subseções.

9.2.1 MACHINE LEARNING NA PESQUISA CIENTÍFICA

Machine Learning é uma subárea da inteligência artificial que se concentra no desenvolvimento de *algoritmos* e modelos que permitem que os sistemas computacionais aprendam e melhorem com a experiência. No contexto das Ciências Sociais Aplicadas, *Machine Learning* oferece uma abordagem capaz de analisar dados e compreender o comportamento humano em uma variedade de contextos.

Em especial, *Machine Learning* envolve a capacidade de computadores aprenderem com dados históricos, identificando padrões e fazendo previsões ou tomando decisões com base nesses padrões. Isso é feito por meio do treinamento de *algoritmos* em conjuntos de dados que contêm exemplos passados e resultados conhecidos. À medida que o *algoritmo* é exposto a mais dados, este ajusta seus parâmetros internos para melhorar seu desempenho, tornando-se mais preciso e eficiente em tarefas específicas.

As técnicas de *Machine Learning* podem ser usadas nos seguintes processos de Pesquisa nas Ciências Sociais Aplicadas:

- **Análise de Dados Complexos:** Em pesquisas que envolvem grandes volumes de dados, como pesquisas de opinião pública, o *Machine Learning* pode ser usado para identificar tendências, padrões de comportamento e *insights* ocultos que seriam difíceis de extrair por meio de métodos tradicionais. (Alinejad-Rokny; Sadroddiny; Scaria, 2018).

- **Previsão de Tendências:** Os *algoritmos* de *Machine Learning* podem ser aplicados para prever tendências sociais, econômicas e políticas.

Isso é particularmente necessário em previsões de eleições, análise de mercado de trabalho e estimativas de demanda por produtos e serviços (Lim; Zohren, 2021).

- **Segmentação de Público-alvo:** Em *marketing* e pesquisa de mercado, o *Machine Learning* é usado para segmentar o público-alvo com base em características demográficas, comportamentais e de consumo. Isso ajuda a direcionar campanhas publicitárias de maneira mais eficaz (Ernawati; Baharin; Kasmin, 2021).

- **Análise de Sentimento e Opinião:** Em análises de mídia social e pesquisas de opinião, o *Machine Learning* é aplicado para analisar o sentimento do público em relação a produtos, serviços ou questões políticas. Isso fornece uma análise das opiniões públicas e das tendências de opinião (Wankhade; Rao; Kulkarni, 2022).

- **Detecção de Fraudes e Anomalias:** Em finanças e segurança, o *Machine Learning* é usado para detectar fraudes e comportamentos anômalos em transações financeiras e atividades online. Tal aspecto ajuda sistemas e recursos contra atividades fraudulentas (Pourhabibi *et al.*, 2020).

Destaca-se que o *Machine Learning* oferece uma nova abordagem para análise de dados e pesquisa nas Ciências Sociais Aplicadas. Essa tecnologia permite que os pesquisadores extraiam insights mais profundos, façam previsões mais precisas e compreendam melhor o comportamento humano em uma ampla variedade de contextos, enriquecendo a pesquisa nessa área de estudo.

9.2.2 TEXT LEARNING: DA LITERATURA CIENTÍFICA AO CONHECIMENTO AVANÇADO

Text Learning, ou aprendizado de texto, é um dos campos vinculados à *Machine Learning* que se concentra na análise e na interpretação de texto escrito. Nas Ciências Sociais Aplicadas, essa abordagem tem se tornado mais relevante, proporcionando uma nova dimensão na pesquisa e compreensão de uma variedade de tópicos. Explora-se a seguir, como o *Text Learning* pode ser aplicado no processo de pesquisa.

Text Learning envolve a utilização de *algoritmos* e técnicas de Processamento de Linguagem Natural (PLN) para extrair informações, padrões e *insights* de documentos de texto. A capacidade do *Text Learning* de compreender a linguagem humana se torna uma ferramenta importante na análise de vastos conjuntos de dados textuais encontrados em documentos, redes sociais, entrevistas, discursos e muito mais.

Nesse sentido, o *Text Learning* pode ser utilizado para:

- **Análise de Opiniões e Sentimentos:** Em pesquisas de mercado e estudos de opinião pública, o *Text Learning* é usado para analisar opiniões e sentimentos expressos em redes sociais, comentários de clientes e pesquisas online. A partir desta análise, é possível ter uma compreensão mais profunda da percepção pública sobre produtos, serviços ou questões sociais (Wankhade; Rao; Kulkarni, 2022).

- **Revisão de Literatura Automatizada:** Na revisão de literatura em Ciências Sociais, o *Text Learning* pode ser aplicado para identificar e resumir automaticamente artigos científicos relevantes em uma área específica. Por meio desta revisão, é possível economizar tempo e ajudar os pesquisadores a manter-se atualizados com os avanços em seu campo (Portenoy; West, 2020).

- **Extração de Conceitos e Relações:** Em estudos das Ciências Sociais Aplicadas, o *Text Learning* pode ser usado para identificar conceitos-chave e relações entre entidades em documentos políticos, discursos políticos ou registros de reuniões, contribuindo para uma análise mais profunda de eventos políticos e sociais.

- **Análise de Texto Qualitativo:** Em pesquisas qualitativas, o *Text Learning* pode auxiliar na categorização e organização de dados de entrevistas, permitindo que os pesquisadores identifiquem tendências e padrões em narrativas humanas. (Rutkowski, 2022).

- **Detecção de Discurso de Ódio e Preconceito:** Em estudos sobre discriminação e preconceito, o *Text Learning* é utilizado para identificar e classificar discurso de ódio em textos *online*, contribuindo para uma análise mais aprofundada das dinâmicas sociais (Poletto *et al.*, 2021).

Destaca-se que o *Text Learning* é uma técnica que pode ser utilizada em uma série de ferramentas nas Ciências Sociais Aplicadas, permitindo que os pesquisadores analisem e compreendam melhor os conjuntos de dados textuais que permeiam as áreas de estudo. Essa abordagem está transformando a maneira como a pesquisa é conduzida, oferecendo novas maneiras de extrair *insights* significativos de documentos escritos e enriquecendo o conhecimento nas Ciências Sociais.

9.2.3 DATA MINING: DESCOBRINDO CONHECIMENTO EM DADOS MASSIVOS

Data Mining, ou mineração de dados, é uma técnica poderosa que desvenda padrões, relações e informações ocultas em grandes conjuntos de dados. Nas Ciências Sociais Aplicadas, essa abordagem se tornou uma ferramenta valiosa para compreender fenômenos sociais complexos e informar a pesquisa. A seguir, explora-se o *Data Mining* e como esta técnica pode ser aplicada no processo de pesquisa nas Ciências Sociais Aplicadas.

Data Mining envolve a análise sistemática de dados para identificar tendências, padrões, correlações e informações relevantes que não seriam facilmente percebidas com métodos convencionais.

Nesse contexto, pode-se fazer uso do *Data Mining* para:

- **Análise de Redes Sociais:** Em sociologia e estudos sociais, o *Data Mining* é usado para analisar redes sociais e interações humanas. Ele pode identificar influenciadores, comunidades e dinâmicas de grupo, fornecendo uma compreensão mais profunda das relações sociais e suas implicações (Serrat, 2017).

- **Previsão de Tendências Sociais:** O *Data Mining* é aplicado para prever tendências sociais, econômicas e políticas. Isso é útil em estudos de políticas públicas, onde os resultados podem informar decisões governamentais e alocar recursos de forma mais eficaz (Serrat, 2017).

- **Identificação de Fatores de Risco:** Em pesquisas de saúde pública e epidemiologia, o *Data Mining* ajuda a identificar fatores de risco em grandes conjuntos de dados de saúde, contribuindo para a prevenção e o controle de doenças.

- **Detecção de Anomalias:** Em segurança cibernética e na detecção de fraudes, o *Data Mining* é utilizado para identificar comportamentos anômalos em transações financeiras e atividades *online*, protegendo sistemas contra ameaças (Pang *et al.*, 2021).

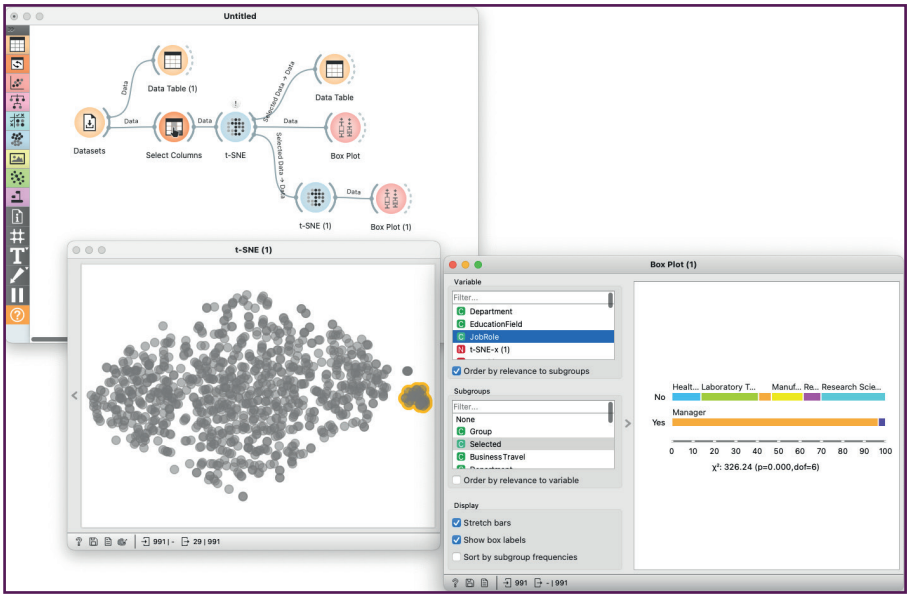
- **Modelagem de Comportamento:** Em psicologia e estudos comportamentais, o *Data Mining* pode ser aplicado para modelar o comportamento humano em diferentes contextos, ajudando a entender e prever respostas a estímulos específicos.

O *Data Mining* oferece uma nova perspectiva na pesquisa nas Ciências Sociais Aplicadas, permitindo que os pesquisadores descubram conhecimento valioso e padrões ocultos em grandes conjuntos de dados. Essa abordagem está transformando a maneira como a pesquisa é conduzida nessas áreas, fornecendo uma base sólida para a tomada de decisões informadas e o avanço do conhecimento nas Ciências Sociais.

9.3 ORANGE DATA MINING

O *Orange Data Mining*, também conhecido como *Orange*, é uma ferramenta visual projetada para ajudar os profissionais de ciência de dados e pesquisadores a explorarem e analisarem os dados de forma eficiente. Tal ferramenta é de código aberto e oferece uma ampla gama de recursos para tarefas de mineração de dados, análise exploratória, modelagem de aprendizado de máquina e visualização de resultados, tornando-a uma escolha popular na comunidade de análise de dados. A Figura 1 apresenta algumas telas da ferramenta.

Figura 1 - Telas da ferramenta *Orange Data Mining*



Fonte: Orange Data Mining (2023)[100]

O *Orange* é conhecido por sua interface intuitiva de arrastar e soltar, que permite aos usuários criar fluxos de trabalho de análise de dados sem a necessidade de codificação extensiva. A ferramenta combina a facilidade de uso com a flexibilidade de personalização, tornando-o adequado tanto para iniciantes quanto para usuários avançados. Além disso, suporta a linguagem de programação *Python*, o que significa que você pode integrar facilmente *scripts Python* personalizados em seus projetos *Orange*.

9.3.1 FUNCIONAMENTO DA FERRAMENTA

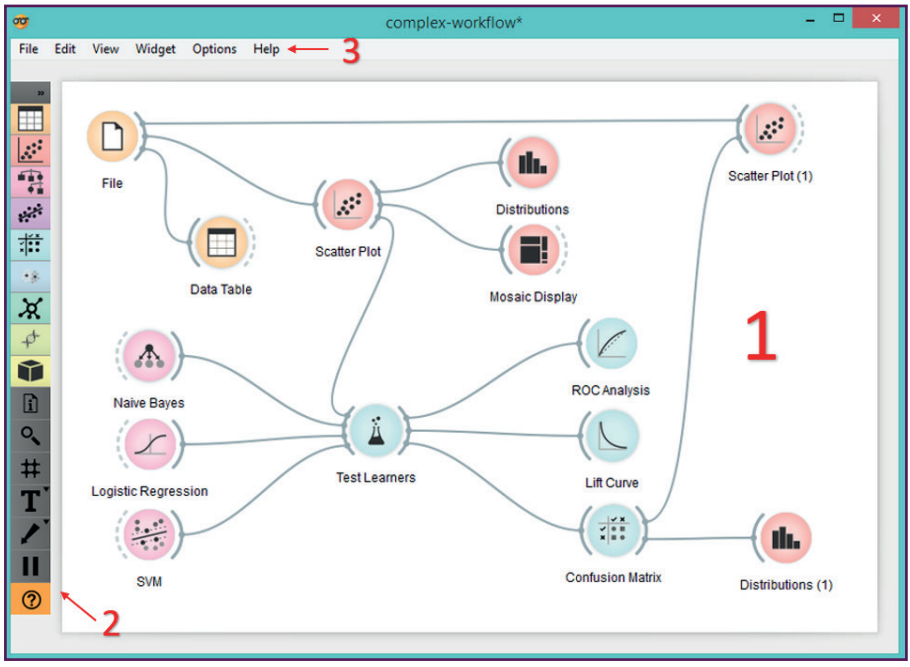
Antes de apresentar o potencial da ferramenta, apresenta-se os principais componentes da interface do *Orange*, buscando demonstrar a navegação e a organização da ferramenta.

100 Disponível em: https://orangedatamining.com/. Acesso em: 28 set. 2023.

9.3.1.1 A INTERFACE ORANGE

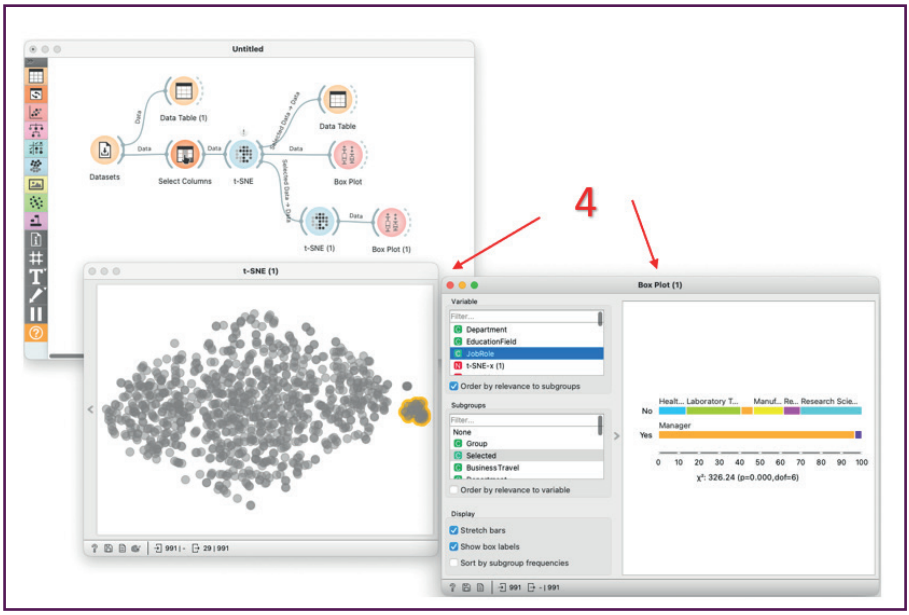
Ao iniciar o *Orange*, é possível ter controle de uma interface amigável composta por várias janelas e painéis. Neste texto, serão identificados numericamente os elementos principais, como ilustrado nas Figuras 2 e 3.

Figura 2 - Tela principal do *Orange Data Mining*



Fonte: Criado pelos autores (2023).

Figura 3 - Continuação da apresentação da tela principal do *Orange Data Mining*



Fonte: Criado pelos autores (2023).

- **Área de Canvas Principal:**
    - Esta é a zona central na qual são construídos e visualizados os fluxos de trabalho. É neste espaço que o usuário arrasta e solta os componentes para criar as suas análises.

- **Painel de Componentes:**
    - À esquerda da interface, o usuário encontra uma variedade de componentes que podem ser usados em seus fluxos de trabalho, como fontes de dados, *algoritmos* de aprendizado de máquina, visualizações e muito mais. Para utilizar tais componentes, basta arrastá-lo para a área de canvas para começar a construir seu fluxo de trabalho.

- **Barra de Ferramentas:**
    - A parte superior da interface contém uma barra de ferramentas com comandos como abrir, salvar e executar fluxos de trabalho, bem como outras funcionalidades essenciais.
- **Janelas de Visualização:**
    - O *Orange* oferece várias janelas de visualização para inspecionar seus dados e resultados, como gráficos, tabelas e painéis de pré-visualização de dados. Essas janelas são abertas automaticamente quando você executa componentes relevantes.

O *Orange Data Mining*, ainda, permite a personalização da interface de acordo com as preferências do usuário. Ademais, é possível ajustar a disposição das janelas, escolher um esquema de cores e até mesmo criar atalhos para as tarefas frequentes. Essa flexibilidade torna a experiência de uso do *Orange* altamente adaptável às suas necessidades específicas.

9.3.2 MANIPULAÇÃO DE DADOS

A manipulação de dados é uma etapa fundamental em qualquer projeto de análise de dados. O *Orange Data Mining* oferece uma série de recursos para importar, preparar e transformar dados, permitindo trabalhar de forma eficiente com conjuntos de dados de diferentes origens e formatos. Neste subcapítulo, explora-se em detalhes como o pesquisador pode realizar a manipulação de dados no *Orange*.

9.3.2.1 IMPORTAÇÃO DE CONJUNTO DE DADOS

A primeira etapa para qualquer análise de dados, é a importação dos conjuntos de dados para o *Orange*. A ferramenta suporta uma ampla variedade de formatos de arquivo, incluindo *CSV* (Comma-separated values), *Excel*, *SQL* (Structured Query Language), e até mesmo a conexão direta com fontes de dados *online*.

Carregar um conjunto de dados no *Orange* é uma tarefa simples, tendo que clicar no *widget 'File'* e selecionar o conjunto de dados na pasta em que ele está armazenado, tornando o processo semelhante ao abrir um arquivo no *Excel*. Aqui estão os passos básicos para importar um conjunto de dados:

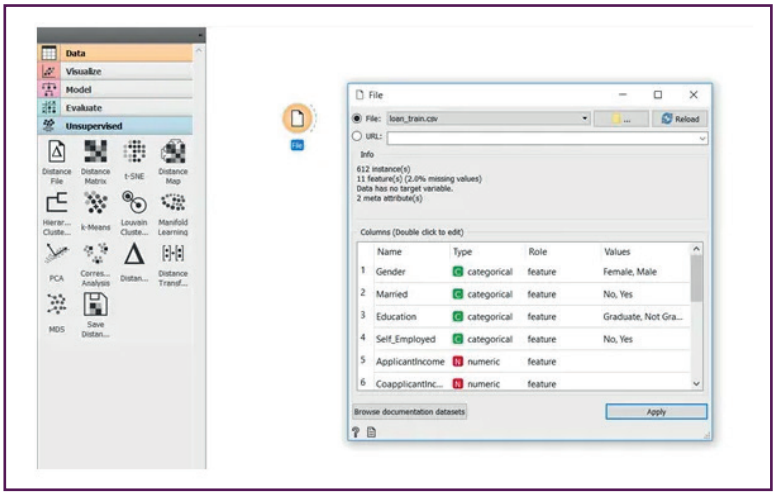
1. **Abrindo um Conjunto de Dados**: No painel de Componentes, tem um componente chamado "Arquivo" (File), sendo necessário arrastá-lo para a área de *Canvas*.

2. **Configurando o Componente de Arquivo**: Após selecionar componente de arquivo, é necessário escolher o arquivo desejado e definir as configurações de importação, como o tipo de delimitador (para *CSV*) ou as credenciais de conexão (para fontes de dados *online*).

3. **Conectando Componentes**: Para continuar a análise, há a necessidade de conectar o componente de arquivo a outros componentes, como gráficos ou *algoritmos* de aprendizado de máquina.

A Figura 4 exemplifica a tela de carregar informações dentro do elemento *File*.

Figura 4 - Tela de carregamento de dados



Fonte: Batista (2019)[101]

101 Disponível em: https://acesse.dev/1kWv4. Acesso em: 28 set. 2023.

9.3.2.2 LIMPEZA E TRANSFORMAÇÃO DE DADOS

Após o processo de importação de dados, a próxima etapa é limpar e transformá-los, preparando-os para análises mais avançadas. O *Orange* oferece uma variedade de componentes para ajudar nessa tarefa:

- **Filtros:** Utilização de filtros para remover dados irrelevantes, duplicados ou outliers do seu conjunto de dados.
- **Normalização e Padronização:** Normalizar ou padronizar atributos numéricos é essencial para muitos *algoritmos* de aprendizado de máquina. O *Orange* fornece componentes para executar essas operações.
- **Transformação de Atributos:** É possível criar novos atributos ou transformar atributos existentes usando funções matemáticas, de texto ou lógicas. Tal transformação ajuda na criação de características mais significativas para sua análise.
- **Amostragem:** Para grandes conjuntos de dados, pode ser útil realizar amostragens aleatórias ou estratificadas para tornar a análise mais ágil e economizar recursos computacionais.

9.3.2.3 SELEÇÃO DE ATRIBUTOS

Destaca-se que nem todos os atributos dos dados são igualmente informativos para as análises. Desta forma, algumas vezes, é necessário selecionar um subconjunto relevante de atributos para melhorar a eficiência do seu modelo. O *Orange* oferece maneiras de fazer isso:

- **Seleção Manual:** É possível selecionar manualmente os atributos que o usuário deseja manter ou remover da análise.
- **Seleção Automática:** O *Orange* também oferece métodos de seleção automática de atributos que identificam os atributos mais importantes com base em critérios estatísticos.

- **Redução de Dimensionalidade:** Em casos de conjuntos de dados de alta dimensionalidade, a redução de dimensionalidade pode ser aplicada para preservar informações essenciais enquanto reduz a complexidade.

A manipulação de dados é uma fase crítica em qualquer projeto de análise de dados, e o *Orange Data Mining* facilita essas tarefas. Compreender como importar, limpar, transformar e selecionar dados é fundamental para preparar seus dados para análises mais avançadas, como modelagem de aprendizado de máquina.

9.3.3 ANÁLISE EXPLORATÓRIA DE DADOS

Na sequência, tem-se uma etapa crítica em qualquer projeto de análise de dados, que é a análise exploratória de dados (AED). Esta etapa envolve a exploração e compreensão inicial dos dados antes de aplicar *algoritmos* de aprendizado de máquina ou realizar análises estatísticas mais avançadas. O *Orange Data Mining* oferece um conjunto robusto de ferramentas e recursos para ajudar na AED, permitindo que os usuários investiguem os dados, identifiquem padrões, avaliem a qualidade e ganhem *insights* valiosos. Neste capítulo, explora-se as diversas maneiras de realizar uma AED eficaz com o *Orange*.

9.3.3.1 TIPOS DE EXPLORAÇÃO VISUAL DOS DADOS

Uma das características distintivas do *Orange* é sua capacidade de permitir que os usuários explorem visualmente os dados de forma mais simples e intuitiva. A seguir, apresenta-se algumas destas opções:

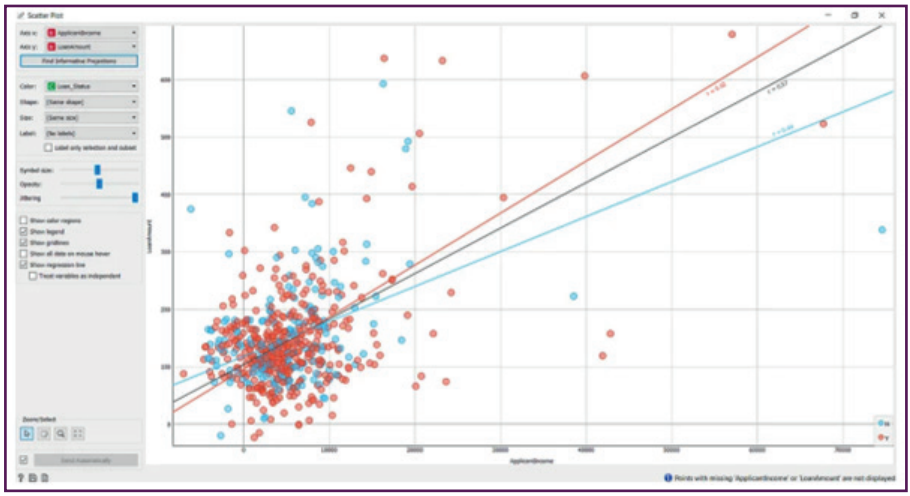
- **Gráficos de Dispersão (Scatter Plots)**: Os gráficos de dispersão são úteis para visualizar relações entre pares de atributos. No *Orange*, é possível criar gráficos de dispersão por meio dos atributos no *widget* de gráfico de dispersão.

- **Histogramas**: Os histogramas ajudam a entender a distribuição dos valores em um atributo. Eles podem ser criados com o *widget* de histograma e são uma ferramenta essencial para identificar a forma das distribuições de dados.

- **Visualizações de Dados Multidimensionais**: Com o *widget* "Projection", é possível criar visualizações de dados multidimensionais, como *PCA* (Análise de Componentes Principais) ou *t-SNE* (t-Distributed Stochastic Neighbor Embedding), para reduzir a dimensionalidade e visualizar agrupamentos e padrões em dados de alta dimensão.

A Figura 5 apresenta a tela de *scatter plot:*

Figura 5 - Tela de *scatter plot*



Fonte: Batista (2019)[102]

9.3.3.2 ESTATÍSTICAS DESCRITIVAS

Além da visualização, o *Orange* fornece informações estatísticas detalhadas sobre seus dados. Isso inclui:

- **Estatísticas Básicas**: É possível obter estatísticas descritivas básicas, como média, mediana, desvio padrão, mínimo e máximo para atributos numéricos.

102 Disponível em: https://encr.pw/1kWv4. Acesso: em 28 set. 2023.

- **Distribuição de Classe**: Em conjuntos de dados de classificação, é possível visualizar a distribuição de classes para entender o desequilíbrio de classes.

- **Correlação**: O *Orange* permite calcular a matriz de correlação para avaliar as relações entre os atributos numéricos.

A detecção de *outliers* e anomalias é fundamental na AED. O *Orange* oferece métodos e ferramentas para identificar pontos de dados que podem ser considerados valores atípicos:

- **Box Plots**: O *widget* "Box Plot" permite criar gráficos de caixa para identificar *outliers* em atributos numéricos.

- **Scatter Plots de Distância Mahalanobis**: Esses gráficos de dispersão ajudam a identificar pontos de dados que estão longe da média multivariada.

Vale destacar que tais análises não são válidas apenas para para dados numéricos, o *Orange* também fornece recursos para explorar dados categóricos:

- **Histogramas Categóricos**: Você pode criar histogramas para variáveis categóricas para entender a distribuição de categorias.

- **Tabelas de Contingência**: As tabelas de contingência ajudam a analisar as relações entre variáveis categóricas, destacando associações e dependências.

O *Orange Data Mining* oferece, ainda, outros recursos para análises geoespaciais, permitindo a obtenção de *insights* a partir de dados de localização. Além desta, os mapas de calor são eficazes para visualizar densidades geográficas e tendências em dados de localização. Com eles, é possível identificar áreas de alta atividade, concentração ou distribuição de eventos geográficos. O *Orange*, ainda, facilita a criação de mapas de calor interativos que destacam as áreas com maior densidade de ocorrências, tornando mais simples a identificação de padrões geoespaciais. Além disso, o *Orange* permite o mapeamento de pontos de dados diretamente em mapas interativos. Essa funcionalidade é especialmente útil para entender a distribuição espacial de eventos ou informações geográficas. Ao visualizar pontos de dados em um mapa, é possível explorar relações entre localizações e identificar *clusters* ou áreas de interesse. Essa compreensão espacial

mais profunda pode ser valiosa em uma variedade de aplicações, como análises de negócios, geografia de saúde pública ou pesquisa ambiental. Com a combinação dessas ferramentas, é possível ter uma visão para análise de dados geoespaciais de maneira eficaz, revelando informações importantes sobre padrões e tendências em informações de localização.

A análise exploratória de dados é uma etapa crucial para compreender a natureza dos seus dados e identificar padrões ou anomalias que podem orientar o restante do seu projeto de análise de dados. Com as ferramentas e recursos do *Orange*, é possível realizar uma AED eficaz de forma intuitiva e informada, estabelecendo uma base sólida para análises mais avançadas e modelagem de aprendizado de máquina.

9.3.4 MODELAGEM DE APRENDIZADO DE MÁQUINA

A modelagem de aprendizado de máquina é uma das etapas com mais potencial no contexto da análise de dados. Tal etapa permite a construção de modelos preditivos a partir de dados, o que pode ser aplicado a uma ampla variedade de cenários, desde classificação de *e-mails* como *spam* ou *não spam* até previsões de vendas de produtos. O *Orange Data Mining* oferece uma ampla gama de *algoritmos* de aprendizado de máquina e ferramentas para criar, avaliar e ajustar esses modelos. Assim, a seguir explora-se em detalhes como realizar modelagem de aprendizado de máquina no *Orange*.

9.3.4.1 ESCOLHA DE ALGORITMO

Na etapa de modelagem de aprendizado de máquina, a escolha do *algoritmo* apropriado é fundamental para o sucesso do projeto. O *Orange* oferece uma variedade de *algoritmos* de classificação, regressão, *clustering* e associação. Alguns dos *algoritmos* mais utilizados e destacados incluem:

- **Regressão Linear:** este *algoritmo* é utilizado para modelar relações lineares entre variáveis dependentes e independentes, sendo especialmente útil quando se deseja entender como uma variável afeta outra de maneira linear.

- **Árvores de Decisão**: árvores de decisão são excelentes escolhas para problemas de classificação, além de oferecerem uma interpretabilidade natural do modelo, permitindo compreender como as decisões são tomadas.

- **Random Forests**: esta é uma abordagem de *ensemble* que combina várias árvores de decisão para melhorar o desempenho do modelo. É particularmente eficaz para reduzir o *overfitting* e aumentar a precisão.

- **K-Means Clustering:** um *algoritmo* amplamente utilizado para tarefas de *clustering*, o *K-Means* segmenta os dados em grupos com base em suas similaridades. É valioso para identificar padrões e estruturas em dados não rotulados.

- **Regras de Associação:** essas regras são úteis para descobrir padrões de associação em conjuntos de dados, como análises de cestas de compras, onde se deseja entender quais itens tendem a ser comprados juntos.

A escolha do *algoritmo* adequado dependerá do tipo de problema enfrentado e dos objetivos da análise. Experimentar diferentes *algoritmos* e avaliar seu desempenho com métricas apropriadas, como mencionadas anteriormente, é uma prática comum para determinar qual *algoritmo* se adapta melhor aos seus dados e metas.

A avaliação de modelo desempenha um papel crucial na construção de sistemas de aprendizado de máquina robustos e eficazes. Esta permite medir o desempenho de um modelo e identificar áreas para melhorias. No *Orange Data Mining*, é possível encontrar um conjunto de ferramentas abrangentes para avaliar modelos de classificação, regressão, *clustering* e associação. Apresenta-se na sequência como realizar uma avaliação completa de modelo.

9.3.4.2 MÉTRICAS DE AVALIAÇÃO DE CLASSIFICAÇÃO

Para modelos de classificação, a escolha das métricas apropriadas é essencial para entender o quão bem o modelo está funcionando. Algumas das métricas de avaliação de classificação mais comuns incluem:

- **Precisão (Accuracy)**: Mede a proporção de instâncias classificadas corretamente em relação ao total de instâncias.

- **Recall (Sensibilidade)**: Calcula a proporção de instâncias positivas corretamente classificadas em relação ao total de instâncias positivas.

- **F1-Score**: Uma métrica que combina precisão e *recall*, sendo útil quando há desequilíbrio entre classes.

- **Matriz de Confusão**: Uma tabela que mostra as classificações corretas e incorretas feitas pelo modelo, fornecendo uma visão detalhada do desempenho em diferentes categorias.

- **Curvas ROC e AUC**: Permitem avaliar o desempenho do modelo em diferentes limiares de classificação e medir a capacidade de discriminação do modelo. A curva *ROC* (Receiver Operating Characteristic Curve) representa a taxa de verdadeiros positivos em relação à taxa de falsos positivos, enquanto a *AUC* (Area under the ROC Curve) quantifica a qualidade geral do modelo, sendo um valor entre 0 e 1, onde valores mais próximos de 1 indicam um modelo melhor.

Para modelos de regressão, as métricas de avaliação focam na precisão das previsões. Alguns exemplos de métricas de avaliação de regressão incluem:

- **Erro Quadrático Médio (RMSE)**: Mede a média dos erros quadrados das previsões em relação aos valores reais, fornecendo uma medida de quão bem as previsões se ajustam aos dados. Quanto menor o *RMSE* (Root Mean Squared Error), mais precisas são as previsões.

- **Erro Absoluto Médio (MAE)**: Calcula a média dos valores absolutos das diferenças entre as previsões e os valores reais, oferecendo uma medida direta da magnitude média dos erros de previsão. O *MAE* (Mean Absolute Error) é útil para entender o tamanho médio dos erros.

- **Coeficiente de Determinação (R^2)**: Avalia a proporção da variabilidade nos dados explicada pelo modelo. O R^2 varia de 0 a 1, onde 1 indica que o modelo explica toda a variabilidade e 0 indica que o modelo não explica nenhuma. É uma métrica importante para determinar o ajuste global do modelo aos dados.

Além destas, apresenta-se com mais detalhes, algumas técnicas que podem ser utilizadas para validação:

- **Validação Cruzada**

  - A validação cruzada é uma técnica essencial para avaliar o desempenho do modelo em dados não vistos. O *Orange* suporta diferentes estratégias de validação cruzada, como validação cruzada *k-fold* e validação cruzada estratificada. Essas abordagens auxiliam na prevenção do superajuste (overfitting) e permitem obter uma estimativa mais confiável do desempenho do modelo em dados futuros. Ao dividir o conjunto de dados em partes menores e testar o modelo em diferentes combinações de treinamento e teste, a validação cruzada fornece uma avaliação mais sólida da capacidade do modelo de generalizar para novos dados, aumentando a confiabilidade de suas conclusões e previsões.

- **Ajuste de Hiperparâmetros**

  - Os hiperparâmetros de um modelo são configurações que podem afetar significativamente seu desempenho. O *Orange* oferece ferramentas para ajustar automaticamente esses hiperparâmetros, como a busca em grade (grid search) e a otimização *bayesiana*. Isso permite encontrar a combinação ideal de configurações para o modelo, maximizando sua capacidade de generalização e precisão. O ajuste de hiperparâmetros é uma etapa crucial no desenvolvimento de modelos de aprendizado de máquina robustos e eficazes, e o *Orange* simplifica esse processo, economizando tempo e esforço dos cientistas de dados e pesquisadores.

- **Visualização de Resultados**

  - Além das métricas numéricas, o *Orange* fornece visualizações interativas para auxiliar na análise dos resultados do modelo. Por exemplo, é possível visualizar as curvas *ROC*, matrizes de confusão e gráficos de dispersão de resíduos para obter uma compreensão mais profunda do desempenho do modelo. Essas visualizações não apenas tornam os resultados mais acessíveis, mas também permitem identificar tendências,

anomalias ou áreas que podem exigir ajustes no modelo, contribuindo assim para aprimorar sua eficácia.

- **Exportação e Implantação**

    - Uma vez que se tenha construído e avaliado o modelo com sucesso no *Orange*, a ferramenta permite exportá-lo para ser utilizado em outros contextos. É possível exportar os modelos treinados em *Python* e integrá-los em aplicações ou fluxos de trabalho de produção.

Destaca-se que com as ferramentas e métricas disponíveis no *Orange Data Mining*, você pode medir, aperfeiçoar e compreender o desempenho dos seus modelos em detalhes. Esse processo é fundamental para assegurar que suas soluções de aprendizado de máquina atendam aos requisitos de qualidade e confiabilidade, garantindo que eles sejam eficazes e precisos em suas aplicações práticas.

9.4 CONSIDERAÇÕES

O capítulo ressalta a influência positiva da Inteligência Artificial (IA) nas Ciências Sociais Aplicadas, especificamente por meio das técnicas de *Machine Learning*, *Text Learning* e *Data Mining*. Essas tecnologias têm o potencial de revolucionar a pesquisa nesse campo, permitindo uma compreensão mais profunda e precisa de fenômenos sociais e comportamento humano.

Machine Learning oferece a capacidade de lidar com grandes volumes de dados, identificando padrões e fazendo previsões precisas, com aplicações em diversas áreas, como mercado de trabalho e análise de opiniões públicas. O *Text Learning* automatiza a revisão de literatura, economizando tempo na identificação de artigos relevantes, além de analisar sentimentos e opiniões expressos em textos. O *Data Mining* revela padrões em grandes conjuntos de dados, permitindo uma compreensão mais profunda das interações sociais e a previsão de tendências.

Importante destacar que essas técnicas não substituem os métodos tradicionais de pesquisa, mas complementam, sendo a experiência humana e

a ética essenciais na pesquisa científica. Questões éticas e de privacidade relacionadas ao uso de dados também devem ser consideradas.

O *Orange* é apresentado como uma ferramenta valiosa não apenas para profissionais e pesquisadores em ciência de dados, mas também para outros profissionais e pesquisadores das ciências sociais aplicadas. Ele possui uma interface intuitiva, suporte à linguagem *Python* e flexibilidade na personalização. A ferramenta abrange todas as fases do processo de análise de dados, desde a importação até a modelagem de aprendizado de máquina, com ênfase na avaliação de modelos.

Em suma, a IA, por meio das técnicas mencionadas, está enriquecendo a pesquisa nas Ciências Sociais Aplicadas, oferecendo novas abordagens para análise de dados. O *Orange Data Mining*, com sua usabilidade, flexibilidade e recursos, é uma ferramenta importante para profissionais e pesquisadores envolvidos em análise de dados, contribuindo para avanços nesse campo.

REFERÊNCIAS

ALINEJAD-ROKNY, Hamid; SADRODDINY, Esmaeil; SCARIA, Vinod. Machine learning and data mining techniques for medical complex data analysis. **Neurocomputing**, [*s. l.*], v. 276, n. 1, p. 1-2, 2018. DOI https://doi.org/10.1016/j.neucom.2017.09.027. Disponível em: https://www.sciencedirect.com/science/article/pii/S0925231217315400. Acesso em: 28 set. 2023.

CARDI, Marilza de Lourdes; BARRETO, Jorge Muniz. Primórdios da computação no Brasil. *In*: SIMPÓSIO DE HISTÓRIA DA INFORMÁTICA NA AMÉRICA LATINA E CARIBE (SHIALC), 2.; CLEI, 38., Medellin, 2012. **Anais** [...]. [*S. l.: s. n.*], 2012. Disponível em: https://www.cos.ufrj.br/shialc/2012/content/docs/shialc_2/clei2012_submission_126.pdf. Acesso em: 28 set. 2023.

ERNAWATI, E.; BAHARIN, S. S. K.; KASMIN, F. A review of data mining methods in RFM-based customer segmentation. **Journal of Physics**: Conference Series, [*s. l.*], v. 1869. p. 012085, 2021. Disponível em: https://iopscience.iop.org/article/10.1088/1742-6596/1869/1/012085/meta. Acesso em: 28 set. 2023.

GARCÍA, Salvador; LUENGO, Julián; HERRERA, Francisco. **Data preprocessing in data mining**. Cham: Springer, 2015. (Intelligent Systems Reference Library, v. 72). Disponível em: https://content.e-bookshelf.de/media/reading/L-3926777-b03bc1919c.pdf. Acesso em: 28 set. 2023.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, Washington, v. 349, n. 6245, p. 255-260, July 2015. Disponível em: https://www.cs.cmu.edu/~tom/pubs/Science-ML-2015.pdf. Acesso em: 30 ago. 2023.

LIM, Bryan; ZOHREN, Stefan. Time-series forecasting with deep learning: a survey. **Philosophical Transactions of the Royal Society A**, Londres, v. 379, n. 2194, p. 20200209, 2021. DOI https://doi.org/10.1098/rsta.2020.0209. Disponível em: https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0209. Acesso em: 30 ago. 2023.

PANG, Guansong; SHEN, Chunhua; CAO, Longbing; VAN DEN HENGEL, Anton. Deep learning for anomaly detection: A review. **ACM Computing Surveys**, [*s. l.*], v. 54, n. 2, p. 1-38, 2021. Disponível em: https://dl.acm.org/doi/10.1145/3439950. Acesso em: 30 ago. 2023.

POLETTO, Fabio; BASILE, Valerio; SANGUINETTI, Manuela; BOSCO, Cristina; PATTI, Viviana. Resources and benchmark corpora for hate speech detection: a systematic review. **Language Resources and Evaluation**, London, v. 55, p. 477-523, 2021. DOI https://doi.org/10.1007/s10579-020-09502-8 . Disponível em: https://link.springer.com/article/10.1007/s10579-020-09502-8. Acesso em: 28 set. 2023.

PORTENOY, Jason; WEST, Jevin D. Constructing and evaluating automated literature review systems. **Scientometrics**, Dordrecht, v. 125, n. 3, p. 3233-3251, 2020. DOI https://doi.org/10.1007/s11192-020-03490-w . Disponível em: https://link.springer.com/article/10.1007/s11192-020-03490-w . Acesso em: 25 set. 2023.

POURHABIBI, Tahereh; ONG, Kok-Leong; KAM, Booi H; BOO, Yee Ling. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. **Decision Support Systems**, [*s. l.*], v. 133, p. 113303, 2020. DOI https://doi.org/10.1016/j.dss.2020.113303.

Disponível em: https://www.sciencedirect.com/science/article/pii/S0167923620300580. Acesso em: 24 set. 2023.

RUTKOWSKI, Rachel A.; LEE, John D.; COLLER, Ryan J.; WERNER, Nicole E. How can text mining support qualitative data analysis?. *In*: HUMAN FACTORS AND ERGONOMICS SOCIETY INTERNATIONAL ANNUAL MEETING, 66th, Atlanta, 2022. **Proceedings** [...]. Los Angeles, CA: SAGE Publications, 2022. p. 2319-2323. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/1071181322661535. Acesso em: 1 out. 2023.

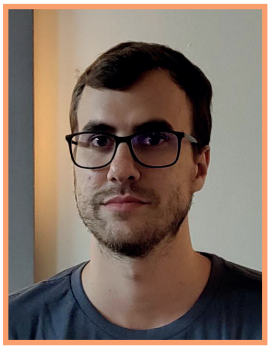
SERRAT, Olivier. Social Network Analysis. *In*: SERRAT, O. (ed.). **Knowledge Solutions**: Tools, Methods, and Approaches to Drive Organizational Performance. Singapore: Springer, 2017. p. 39-43.

TOMMASEL, Antonela; GODOY, Daniela. Short-text learning in social media: a review. **The Knowledge Engineering Review**, Cambridge, v. 34, p. e7, 2019. DOI https://doi.org/10.1017/S0269888919000018. Disponível em: https://www.cambridge.org/core/journals/knowledge-engineering-review/article/shorttext-learning-in-social-media-a-review/F2A5A1F47D512C94BD265A2D63AFF593. Acesso em: 1 out. 2023.

WANKHADE, Mayur; RAO, Annavarapu Chandra Sekhara; KULKARNI, Chaitanya. A survey on sentiment analysis methods, applications, and challenges. **Artificial Intelligence Review**, Oxford, v. 55, n. 7, p. 5731-5780, 2022. DOI https://doi.org/10.1007/s10462-022-10144-1. Disponível em: https://link.springer.com/article/10.1007/s10462-022-10144-1. Acesso em: 27 set. 2023.

DADOS DOS AUTORES:

Caio Saraiva Coneglian



Caio Saraiva Coneglian é Doutor e mestre em Ciência da Informação. Bacharel em Ciência da Computação. Docente da Universidade de Marília - UNIMAR. Pesquisador do Instituto Brasileiro de Informação em Ciência e Tecnologia- IBICT. Docente colaborador do PPGCI - UNESP.

https://orcid.org/0000-0002-6126-9113
caio.coneglian@gmail.com

Henrique Leal Tavares

Henrique Leal Tavares é Mestre em Ciência da Computação. Tecnólogo em Análise e Desenvolvimento de Sistemas. Docente da Universidade de Marília - UNIMAR.

https://orcid.org/0009-0006-8960-3386
henriquetavares@unimar.br

Diego José Macêdo

Diego José Macêdo é Mestre em Ciência da Informação pela Universidade de Brasília. Bacharel em Sistema de Informação pela Universidade Católica de Brasília. Atualmente é tecnologista do Instituto Brasileiro de Informações em Ciência e Tecnologia - Ibict.

diegomacedo@ibict.br
https://orcid.org/0000-0002-5696-0639

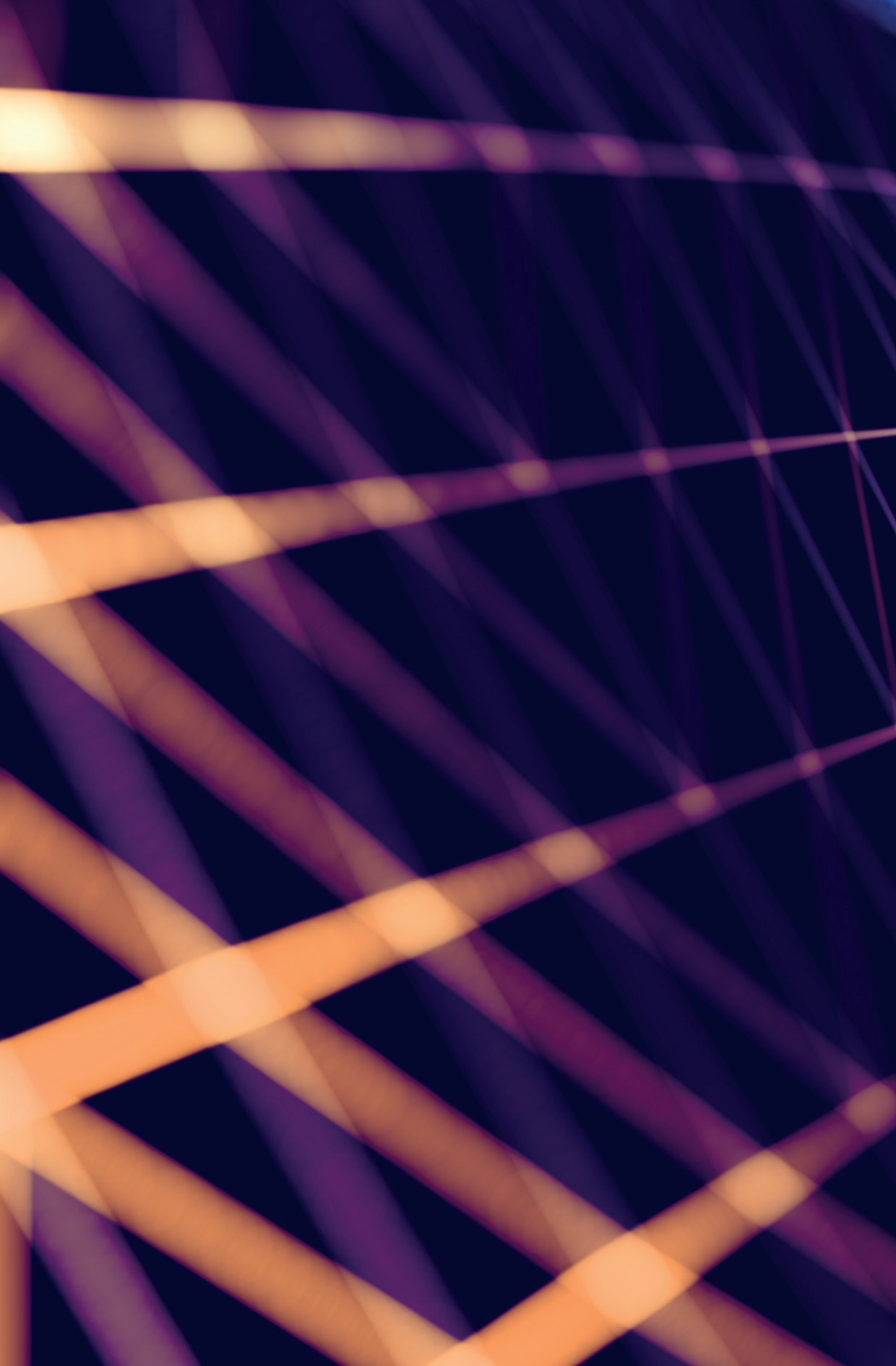
Milton Shintaku

Milton Shintaku é Doutor em Ciência da Informação pela Universidade de Brasília. Coordenador de Tecnologia para Informação (Cotec) do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict).

shintaku@ibict.br
https://orcid.org/0000-0002-6476-4953

Como referenciar o capítulo 9:

CONEGLIAN, Caio Saraiva; TAVARES, Henrique Leal; MACÊDO, Diego José; SHINTAKU, Milton. Orange Data Mining: Uma ferramenta para inserção de inteligência artificial na pesquisa científica. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 9. p. 245-273. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap9.



10. REVOLUCIONANDO A PESQUISA CIENTÍFICA COM A PLATAFORMA KNIME ANALYTICS

Fernanda Farinelli

10.1 INTRODUÇÃO

A pesquisa em Ciências Sociais Aplicadas (CSA) se concentra em compreender e analisar os fenômenos sociais, culturais e econômicos que afetam a sociedade. Essa área abrange um vasto campo de estudo, englobando disciplinas como Ciência da Informação, Economia, Sociologia, Psicologia, Administração, entre outras. Seu principal foco reside na compreensão das complexas dinâmicas sociais, econômicas e culturais que moldam nosso mundo. Ela desempenha um papel fundamental na busca por soluções para os desafios contemporâneos que enfrentamos, como a desigualdade social, a sustentabilidade ambiental, o desenvolvimento econômico e a melhoria da qualidade de vida (Gil, 2008; Minayo, 2001).

Nos últimos anos, a pesquisa no campo das CSA tem passado por uma transformação significativa com o uso crescente da Tecnologia da Informação (TI). A TI, incluindo a *internet*, dispositivos móveis e inteligência artificial trouxeram novas perspectivas e ferramentas para os pesquisadores dessa área. O casamento entre tal campo e a TI tem se revelado uma aliança poderosa, transformando a maneira como conduzimos pesquisas e compreendemos a sociedade.

O processo de pesquisa envolve diversas etapas essenciais para a produção de conhecimento. Inclui, dentre outras etapas, a coleta de dados, que pode ser conduzida por meio de métodos quantitativos ou qualitativos, seguida pelo tratamento ou elaboração dos dados para, então, passar pela etapa de análise dos dados com o intuito de identificar padrões e responder às questões de pesquisa. Posteriormente, os resultados são interpretados e as conclusões são comunicadas por meio de relatórios ou publicações acadêmicas, desempenhando um papel fundamental no

avanço do conhecimento em uma determinada área (Gil, 2002, 2008; Lakatos; Marconi, 2010; Marconi; Lakatos, 2002).

As tecnologias digitais têm se tornado recursos essenciais para os pesquisadores nos diversos campos de pesquisa. Observa-se que elas têm impactado a pesquisa científica ao trazer alternativas para a coleta e análise de dados em grande escala, assim como oferecer melhores possibilidades de tratamento, organização e visualização dos resultados (Swiech; Francisco; Lima, 2016).

As novas possibilidades da TI também revolucionaram a forma como os cientistas se comunicam, colaboram e compartilham suas pesquisas (Mueller, 1994). A TI também tem desempenhado um papel fundamental na disseminação de pesquisas. As publicações acadêmicas, agora, são facilmente encontradas *on-line*, tornando o conhecimento mais acessível a uma audiência global. Ademais, as mídias sociais e as plataformas de compartilhamento de informações permitem que os pesquisadores alcancem um público mais amplo e envolvam a sociedade em debates relevantes.

Além disso, a análise de *big data* permite aos pesquisadores examinarem grandes conjuntos de informações de forma mais eficiente e detalhada, abrindo novas possibilidades de pesquisa. As vastas quantidades de dados disponíveis *on-line* abriram um mundo de novas possibilidades para a pesquisa. Pesquisadores de diversas disciplinas podem explorar uma variedade de fontes de dados, desde os demográficos e econômicos até informações sobre comportamento humano e opiniões públicas (Leonelli, 2022). As redes sociais, por exemplo, fornecem uma rica fonte de dados sobre comportamentos sociais e interações humanas, que podem ser analisados para obter *insights* sobre tendências culturais e comportamentais.

Essa disponibilidade de dados, muitas vezes em tempo real, permite que os pesquisadores realizem estudos longitudinais abrangentes, analisem tendências globais e até mesmo investiguem fenômenos em escala macro, proporcionando uma compreensão mais profunda e holística dos desafios e oportunidades em seus campos de estudo (Leonelli, 2022; Moura; Amorim, 2014). Além disso, a capacidade de acessar dados de fontes diversas e geograficamente dispersas promove a colaboração interdisciplinar e internacional, enriquecendo ainda mais a pesquisa e impulsionando a inovação científica (Garcia; Vieira; Vivacqua; França; Dias, 2020).

Essa transformação não é apenas uma resposta às demandas da era digital, mas também uma oportunidade para impulsionar o rigor, a eficiência e a amplitude da pesquisa. Nesse contexto, este capítulo tem como objetivo explorar a influência e o impacto da TI na pesquisa e metodologia científica, destacando a *plataforma KNIME* (Konstanz Information Miner), *Knime Analytics Platform*, como uma ferramenta multifacetada e versátil que atende a uma variedade de etapas e necessidades na pesquisa científica sobretudo nas CSA.

À medida que se avança nessa jornada pela interseção entre a pesquisa em CSA e TI, são exploradas as diversas formas pelas quais a ferramenta *KNIME* pode enriquecer os estudos, identificando as etapas metodológicas em que sua aplicação é pertinente. A expectativa é que este capítulo possa capacitar os pesquisadores a adotarem esse tipo de tecnologia para o aprimoramento de suas pesquisas.

10.2 PLATAFORMA KNIME ANALYTICS

A análise de dados evoluiu de um campo especializado para uma ferramenta fundamental em quase todos os setores da sociedade. À medida que a era da informação avança, a quantidade de dados disponíveis se torna cada vez mais impressionante. Com isso, surge a necessidade de ferramentas que possam extrair significado e *insights* valiosos desse oceano de informações.

É nesse contexto que o *KNIME* se destaca como uma poderosa ferramenta de código aberto para análise de dados, processamento de informações e criação de fluxos de trabalho analíticos. A ferramenta *KNIME* é muito mais do que uma simples ferramenta, é uma plataforma completa e acessível que emerge como uma solução poderosa e versátil para lidar com os desafios da coleta, preparação, análise e visualização de dados. Atualmente, a ferramenta está na versão 5.1 com uma interface gráfica moderna e aprimorada, contribuindo para redução na curva de aprendizado para novos usuários (KNIME, 2022a).

10.2.1 DOWNLOAD E INSTALAÇÃO

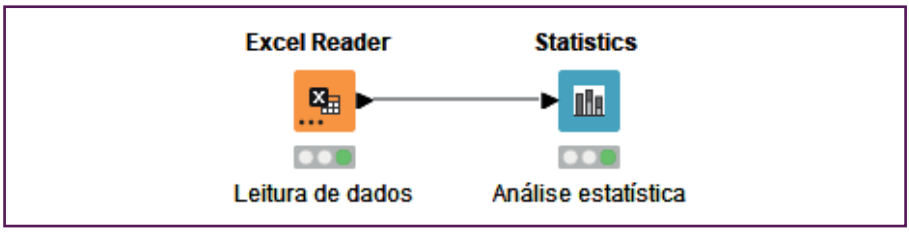
A versão mais recente da *Plataforma de Análise KNIME*, a versão 5.1.1, está disponível para *download* no *site* oficial[103], e é compatível com os sistemas operacionais para *Windows*, *Linux* e *macOS*. Para detalhes sobre a instalação consulte o guia oficial (KNIME, 2023b) fornecido pelo fabricante.

10.2.2 CONCEITOS BÁSICO DA FERRAMENTA

A *plataforma KNIME* se fundamenta na ideia do paradigma da programação visual na qual implica que um programa seja criado usando um grafo como um fluxo de dados (Khodnenko; Ivanov; PROKOFIEV; Lantseva, 2020; KNIME, 2023c).

Em seu ambiente de programação visual, adota-se o conceito de fluxos de dados, também conhecidos como *pipelines*. Isso permite que os usuários construam fluxos de trabalho (Figura 1) de forma intuitiva e eficiente, combinando diversos componentes e etapas para processar dados. Um fluxo de trabalho (workflow) é uma representação gráfica de uma sequência de etapas ou operações que descrevem todo o processo de análise ou processamento de dados (Berthold *et al.*, 2006; Berthold *et al.*, 2009; Hayasaka; Silipo, 2023; KNIME, 2022a; 2023c). É uma maneira visual de organizar e executar tarefas de análise de dados em um ambiente interativo.

igura 1 - Exemplo de fluxo de trabalho visual do *KNIME*



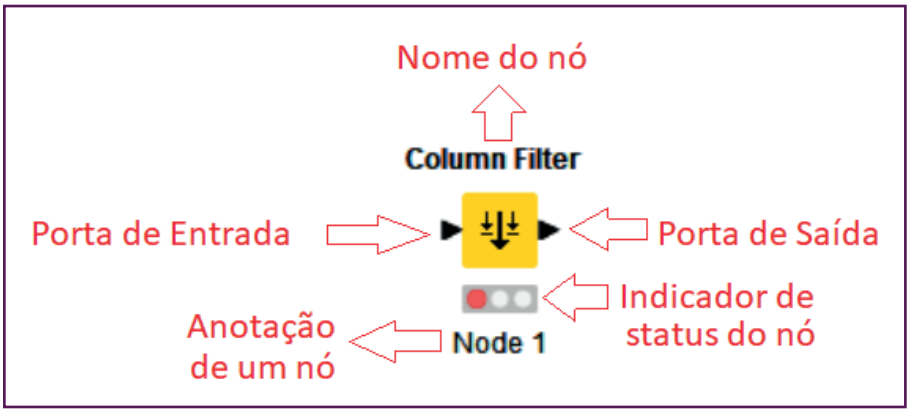
Fonte: Elaborado pela autora (2023).

103 Disponível em: https://www.knime.com/downloads. Acesso em: 17 set. 2023.

Nesse contexto, os usuários podem criar fluxos de trabalho compostos por nós que manipulam os dados, e esses dados são transportados entre os nós por meio de conexões. Um "nó" refere-se a uma unidade de processamento individual. Cada nó desempenha uma função específica em um fluxo de trabalho, e pode realizar ações como leitura de dados, pré-processamento, modelagem estatística, visualização, entre outros.

Na Figura 2, que se segue, apresenta-se uma minuciosa análise da anatomia de um nó, destacando seus componentes e características essenciais. Cada nó geralmente possui portas que servem como entrada e saída de dados respectivamente à esquerda e à direita do nó. Cada nó do repositório de nós ter um nome único que o identifica no contexto geral do *KNIME*. O "nome do nó" refere-se simplesmente à designação ou rótulo que é atribuído a um nó específico. Esses nomes são frequentemente escolhidos de forma a descrever sucintamente a função ou o propósito do nó, o que facilita a compreensão do usuário da ferramenta (KNIME, 2023c).

Figura 2 - Anatomia de um nó

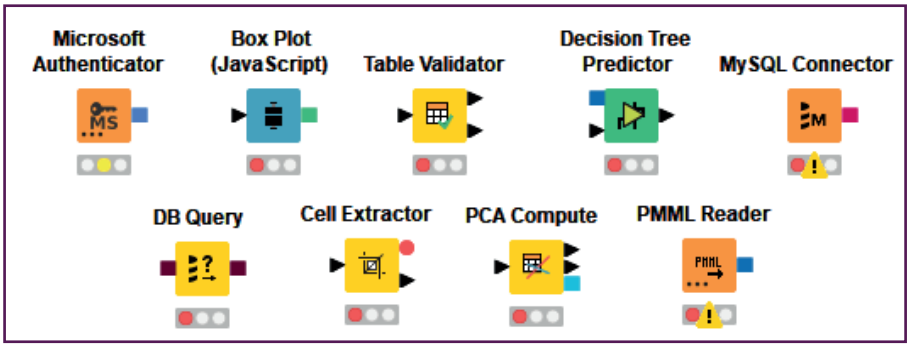


Fonte: Elaborado pela autora (2023).

Já a "anotação de um nó" é uma funcionalidade que permite aos usuários adicionarem informações textuais ou notas personalizadas a um nó específico em um fluxo de trabalho. Essas anotações são úteis para documentar e explicar o propósito ou a função de um nó, fornecendo detalhes adicionais que auxiliem na compreensão do fluxo de trabalho (KNIME, 2023c).

Existem diferentes tipos de portas de entrada e saída, normalmente, são representadas por diferentes símbolos (Figura 3). Os nós são diferenciados por cores, as quais refletem suas categorias. Como exemplo, todos os nós na tonalidade amarela destinam-se à manipulação de dados.

Figura 3 - Exemplo das diferentes portas de entrada e saída existentes



Fonte: Elaborado pela autora (2023).

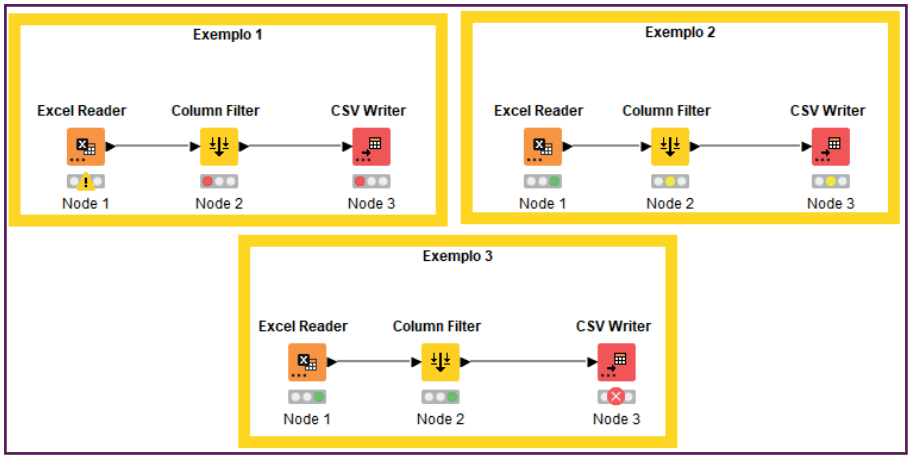
É importante ter em mente a compatibilidade entre essas portas ou, mais precisamente, entre os tipos de dados que elas manipulam. A regra fundamental é que uma porta de saída específica deve, obrigatoriamente, compartilhar o mesmo símbolo e cor com a porta de entrada correspondente para que sejam consideradas compatíveis.

A Figura 4 apresenta exemplos de três diferentes nós disponíveis no *KNIME*. Os nós "*Excel Reader*" e "*Column Filter*" possuem um triângulo preto à sua direita que representa as portas de saída. O nó "*Column Filter*" e o nó "*CSV Writer*" possuem um triângulo preto à sua esquerda que representa as portas de entrada de dados.

Um nó recebe um conjunto de dados como entrada, o processa e o torna disponível em sua porta de saída (Berthold *et al.*, 2009; Hayasaka; Silipo, 2023; KNIME, 2023c). Cada nó pode assumir quatro estados distintos, que seguem um esquema de cores semelhante a um semáforo de trânsito. Esses estados são descritos a seguir e ilustrados na Figura 4 (KNIME, 2023c):

- Inativo e não configurado -> Luz vermelha;
- Configurado, mas não executado -> Luz amarela;
- Executado com sucesso -> Luz verde;
- Executado com erros -> Luz vermelha com um "X".

Figura 4 - Exemplo dos diferentes estados dos nós



Fonte: Elaborado pela autora (2023).

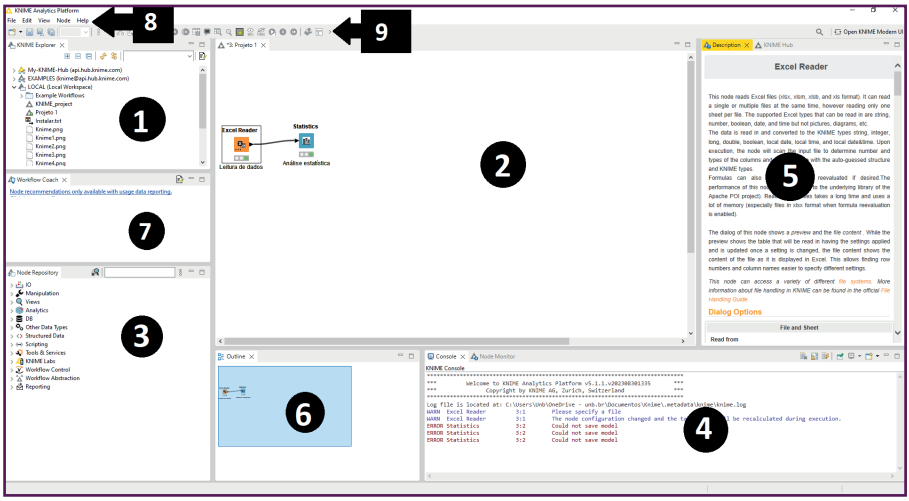
Na Figura 4, no exemplo 1, o primeiro nó "*Excel Reader*" apresenta um símbolo amarelo com um sinal de exclamação, que indica um alerta que o nó deve ser configurado, esse sinal sobrepõe o sinalizador do meio que está em amarelo. No Exemplo 1, ainda pode-se notar que os nós "*Column Filter*" e "*CSV Writer*" apresentam o *status* vermelho, indicando que estão inativos e não configurados. A configuração do primeiro nó nesse fluxo e sua respectiva execução alteram o *status* desse nó para executado com sucesso, conforme ilustrado no Exemplo 2. Automaticamente, pode-se ver que o segundo nó passa para o estado "configurado, mas não executado", pois esse nó não demanda nenhuma configuração especial. Já o terceiro nó deve ser configurado para que fique com a luz amarela. Por fim, no Exemplo 3, nota-se que os dois primeiros nós foram executados com sucesso e o terceiro nó foi executado com erro.

Na subseção seguinte, explora-se em detalhes a interface de trabalho do *KNIME*, examinando as características gerais da interface a fim de se obter uma compreensão completa da ferramenta.

10.2.3 INTERFACE DE TRABALHO

A ferramenta *KNIME* disponibiliza uma interface gráfica intuitiva (Figura 5) que confere à ferramenta acessibilidade, permitindo sua utilização por profissionais de diversas disciplinas, abrangendo não apenas programadores, mas também indivíduos de diferentes especializações (KNIME, 2023c).

Figura 5 -Interface gráfica do *KNIME* até a versão 4.7



Fonte: Elaborado pela autora (2023).

A interface gráfica ou área de trabalho do *KNIME* é composta pelos seguintes componentes, numerados de 1 a 9 na Figura 6 acima e explicados a seguir no Quadro 1.

Quadro 1 - Detalhes da interface gráfica da plataforma *KNIME*

Nº	Nome	Descrição
1	Explorador (KNIME Explorer)	Uma visão geral dos fluxos de trabalho disponíveis e grupos de fluxos de trabalho existentes nas áreas de trabalho que você tem ativo, ou seja, no seu espaço de trabalho local e nos servidores KNIME. O espaço de trabalho é onde os fluxos de trabalho são criados e organizados.
2	Editor de Fluxo de Trabalho (Workflow Editor)	Uma área de edição visual para o fluxo de trabalho ou workflow atualmente ativo. É aqui que você cria, edita e configura os fluxos de trabalho. Você pode ter múltiplas abas de edição abertas, uma para cada workflow que está trabalhando.
3	Repositório de Nós (Node Repository)	Todos os nós disponíveis na Plataforma KNIME que são instalados junto com a instalação padrão e os nós das extensões que você instalou estão listados aqui. Os nós são organizados por categorias, mas você também pode usar a caixa de pesquisa na parte superior do repositório de nós para encontrar nós. Os usuários podem pesquisar e arrastar nós do navegador para a área de trabalho para adicioná-los ao fluxo de trabalho.
4	Console	É a área onde são exibidas as mensagens de execução que indicam o que está acontecendo nos bastidores durante a execução de um fluxo de trabalho. Isso pode incluir informações sobre o progresso da execução, erros encontrados, resultados produzidos e outras mensagens relevantes que auxiliam os usuários no monitoramento e na depuração de seus fluxos de trabalho.

Nº	Nome	Descrição
5	Descrição do Nó (Node Description)	Área onde é apresentada a descrição de um nó selecionado (no Editor de Fluxo de Trabalho ou no Repositório de Nós) ou do fluxo de trabalho em si (caso nenhum nó do fluxo seja selecionado).
6	Visão Geral (Outline)	Essa área apresenta uma visão geral do fluxo de trabalho atualmente ativo. Seu uso é interessante quando um workflow fica grande, com diversos nós, e não é facilmente visível na área de edição.
7	Assistente de Fluxo de Trabalho (Workflow Coach)	Essa área é destinada à lista recomendações de nós com base nos fluxos de trabalho criados pela comunidade de usuários. Ela fica inativa se você não permitir que o KNIME colete estatísticas de uso. Essa janela pode ser fechada e, quando desejado, ser reaberta.
8	Barra de Menu	Contém uma série de menus suspensos que fornecem acesso a uma ampla gama de comandos e funcionalidades.
9	Barra de Ferramentas	Contém botões para ações comuns, como salvar, abrir, executar e interromper fluxos de trabalho.

Fonte: Elaborado pela autora a partir de (KNIME, 2023c).

Com um entendimento sólido da interface, pode-se explorar como essa ferramenta permite construir fluxos de trabalho personalizados e executar análises de dados eficazes.

10.2.4 CONSTRUÇÃO DE UM FLUXO DE TRABALHO

O processo de construção de um fluxo de trabalho normalmente é iniciado por um nó de origem, que, em geral, tem a finalidade de coletar ou extrair dados provenientes de uma fonte de dados específica, por exemplo, um

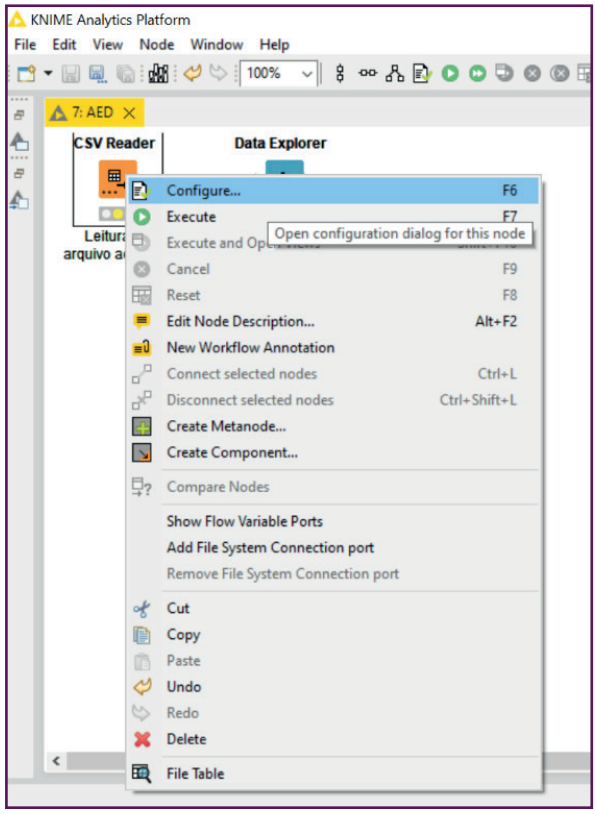
arquivo, uma *página web*, um banco de dados, dentre outras fontes. Os dados coletados são armazenados em um formato interno baseado em tabelas, no qual cada coluna possui um tipo de dado específico. Na sequência, essas tabelas de dados são encaminhadas a outros elementos por meio de conexões, possibilitando que esses elementos realizem uma série de operações, tais como modificações, transformações, modelagem ou geração de representações visuais fundamentadas nos dados previamente processados.

Um fluxo de trabalho da *plataforma KNIME* é criado simplesmente arrastando e soltando os nós existentes no repositório de nós para o editor de fluxo de trabalho ou clicando duas vezes sobre os nós. Dessa forma, à medida que os nós são adicionados sequencialmente, o *pipeline* é rapidamente construído. Cada nó deve ser configurado conforme necessário e, em seguida, o fluxo de trabalho pode ser executado para processar os dados.

10.2.4.1 CONFIGURAÇÃO DE UM NÓ

Até a versão 5.1 do *KNIME*, para configurar um nó basta clicar duas vezes sobre ele ou pressionar a tecla F6 com o nó selecionado (Figura 6).

Figura 6 - Acionando a configuração de um nó

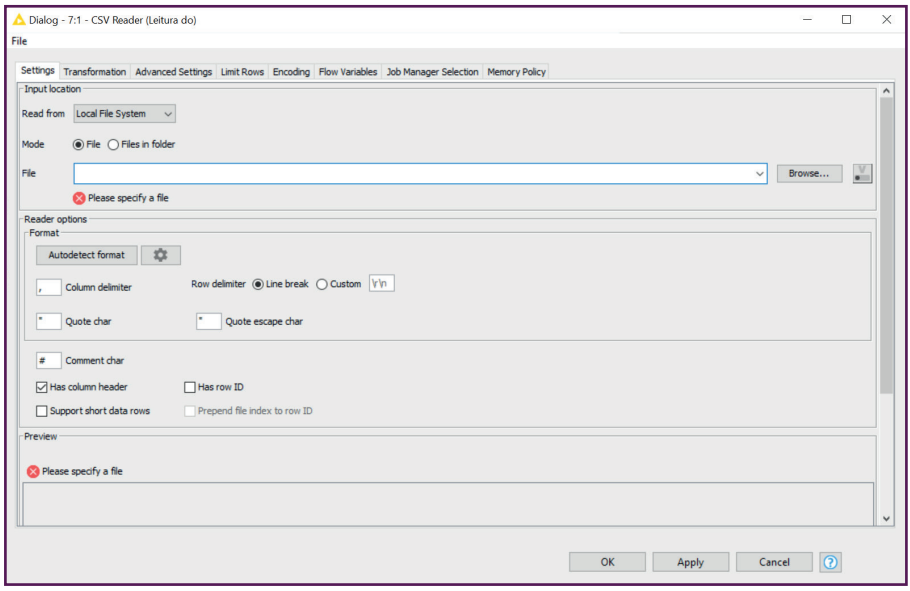


Fonte: Elaborado pela autora (2023).

Dentro da janela de configurações, tem-se acesso a um conjunto de opções e parâmetros que podem ser adaptados de acordo com as exigências do projeto. A diversidade dessas opções está vinculada ao tipo de nó em questão e às funcionalidades específicas que ele proporciona. Por exemplo, é possível definir as entradas de dados, configurar os parâmetros de processamento, especificar as ações a serem executadas e outras configurações pertinentes ao funcionamento do nó.

Para ilustrar a configuração de um nó, a Figura 7, a seguir, apresenta a janela de configurações do nó "CSV Reader". Esse nó é usado para ler dados de um arquivo *CSV* (Comma-Separated Values) (KNIME, 2022b).

Figura 7 - Janela de configuração do nó CSV *Reader*



Fonte: Elaborado pela autora (2023).

Existem várias opções de configuração para especificar como os dados do arquivo *CSV* devem ser lidos e interpretados. Conforme a Figura 7, alguns exemplos de configuração para esse nó são (KNIME, 2022b):

- No campo "*File*" informe o nome e caminho completo do arquivo a ser lido;
- Escolha o caractere ou sequência de caracteres que delimita os campos no arquivo *CSV*. O caractere padrão é a vírgula (,), mas pode ser configurado para outros caracteres, como ponto e vírgula (;) ou tabulação (\t);
- Para adivinhar automaticamente a estrutura do arquivo, clique no botão "*Autodetect format*";
- Se a primeira linha do arquivo *CSV* contiver nomes de coluna, você pode marcar a opção "*Has column header*".

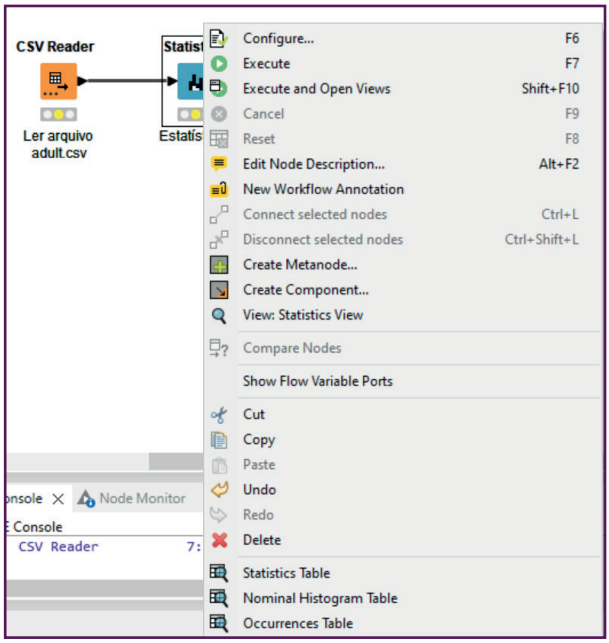
Informações detalhadas sobre cada nó podem ser encontradas na janela "*Description*". Ao consultar essa seção, os pesquisadores podem obter orientações detalhadas sobre como configurar cada nó. Dessa forma, para

cada nó, explore as opções de configuração disponíveis e faça os ajustes necessários para configurar o nó de acordo com os requisitos do seu fluxo de trabalho. Certifique-se de entender o impacto de cada configuração antes de aplicá-la.

10.2.4.2 EXECUÇÃO DO NÓ OU DO WORKFLOW

Após configurar os nós do fluxo de trabalho, o próximo passo essencial é executar o nó ou até mesmo o *workflow* completo. Para isso, basta clicar com o botão direito do mouse sobre o nó desejado e o mesmo menu exibido no ato da configuração será aberto. As opções "*Execute*" e "*Execute and Open Views*" estão ativas (Figura 8).

Figura 8 - Detalhe do menu para acesso a configuração e execução do nó



Fonte: Elaborado pela autora (2023).

Ao selecionar "*Execute*", o nó será processado de acordo com as configurações previamente definidas. Durante sua execução será possível observar o progresso da execução. É importante observar qualquer mensagem de erro

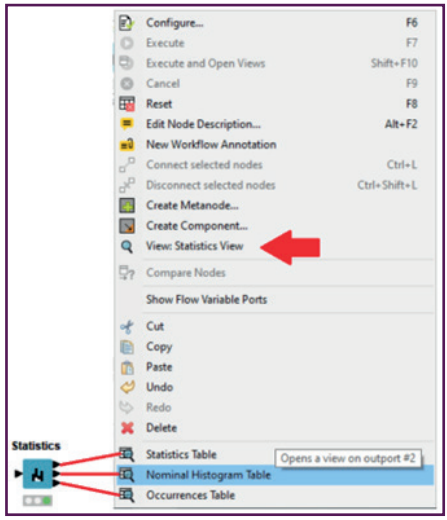
ou alerta, pois isso pode indicar problemas nos dados ou na configuração do nó. Ao utilizar a opção "*Execute and Open Views*", tão logo encerrada a execução, se os dados foram processados corretamente, será exibida uma janela com os resultados imediatamente após o processamento dos dados.

10.2.4.3 VERIFICAÇÃO DE RESULTADOS

Durante a construção e execução do *workflow* é importante proceder a uma inspeção dos resultados intermediários após a execução de cada nó. Essa prática ajuda a entender a tarefa encapsulada em cada nó e ajuda a assegurar a precisão e validade das operações realizadas. Essa verificação é realizada por meio da análise das portas de saída associadas a cada nó no fluxo de trabalho. Cada nó apresenta portas de saída que encapsulam os dados processados por ele.

Para visualizar os resultados, clicar sobre o nó com o botão direito do mouse, no menu que é exibido, assim é possível acessar o resultado de cada porta ao selecionar a respectiva opção localizada no inferior desse menu (Figura 9). Quando um nó possui múltiplas portas, a lista de resultados segue a mesma ordem que as portas, basta selecionar a porta de saída relevante e examinar os dados apresentados. O nó "*Statistics*" possui três portas de saída conforme indicado na Figura 9, e as três opções inferiores do menu equivalem aos resultados da execução desse nó, cada tabela é a saída de uma das portas. Adicionalmente, conforme destacado pela seta vermelha, em geral, nós que têm a opção de "*Execute and Open Views*" possuem a opção de visão interativa dos resultados, geralmente identificado pelo ícone de lupa.

Figura 9 - Exemplo da relação de postas e os respectivos resultados



Fonte: Elaborado pela autora (2023).

Dessa forma, é possível realizar uma validação dos resultados em relação às expectativas previamente estabelecidas. Saber interpretar e validar esses resultados é fundamental para garantir que as análises estejam corretas e confiáveis.

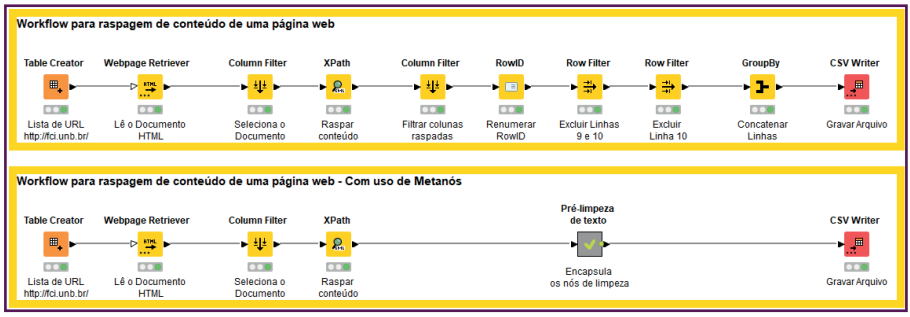
10.2.5 COMPONENTES E METANÓS

No *KNIME*, Componentes (em inglês *Components*) e Metanós (em inglês *Metanodes*) são recursos que permitem a organização, reutilização e modularização de partes de seus *workflows*, tornando o processo de construção e gerenciamento mais eficiente e organizado. Eles são particularmente úteis quando se deseja criar módulos de análise personalizados que podem ser aplicados em várias partes do seu processo de análise (KNIME, 2023f).

O exemplo da Figura 10 apresenta dois *workflows* que fazem a coleta de dados disponíveis em uma *página web*. No primeiro, existem cinco nós que são responsáveis por realizar uma pré-limpeza do conteúdo raspado, são os nós de "*Column Filter*" até o "*GroupBy*". Já no segundo *workflow*,

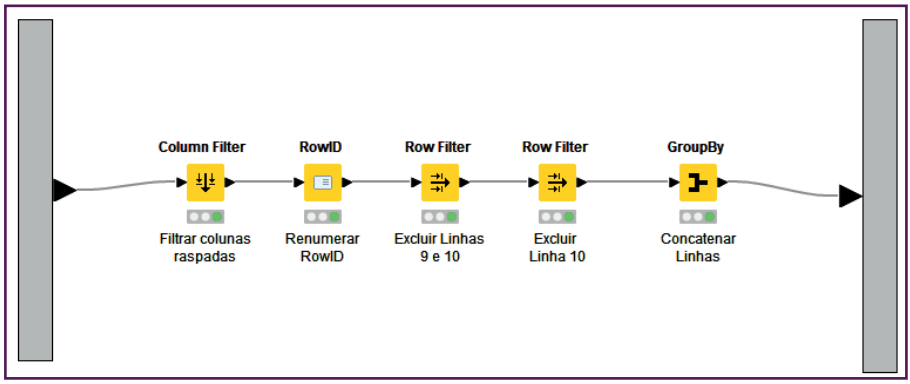
tem-se apenas o metanó identificado por "*Pré-limpeza de texto*", encapsulando esses nós conforme pode ser visualizado na Figura 11.

Figura 10 - Exemplo de *workflow* com e sem metanó



Fonte: Elaborado pela autora (2023).

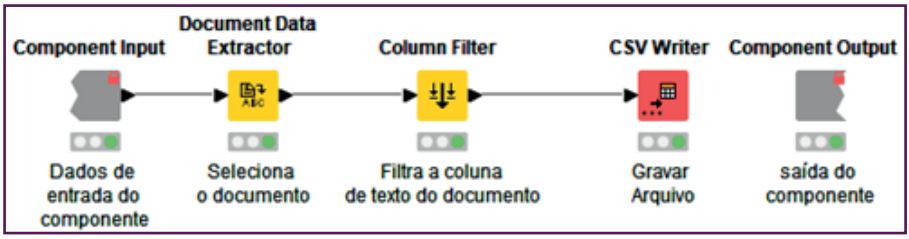
Figura 11 - *Workflow* encapsulado no metanó



Fonte: Elaborado pela autora (2023).

Na Figura 12 é apresentado um exemplo de *workflow* encapsulado em um componente. Nota-se a distinção em relação a Figura 11 pela existência de dois nós específicos: o "*Component Input*", que representa a entrada de dados no componente, e o "*Component Output*", que representa a saída dos dados processados pelo componente. Esses pontos de entrada e saída funcionam como pontos de integração entre o componente e os *workflows* que o utilizam, facilitando a comunicação eficaz de dados entre diferentes partes da solução.

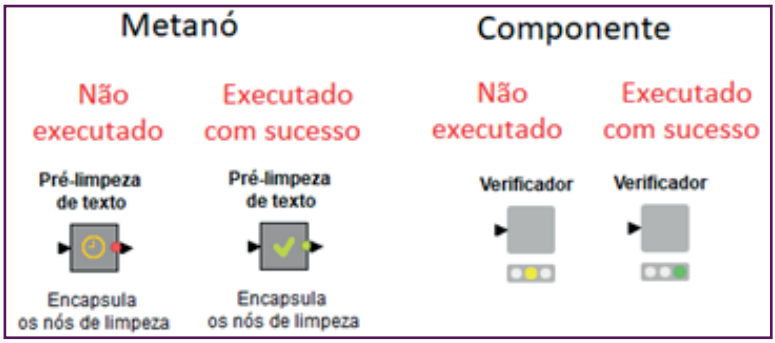
Figura 12 - Exemplo de um *workflow* encapsulado em um componente



Fonte: Elaborado pela autora (2023).

Os componentes e metanós encapsulam *workflows* ou partes específicas, com o objetivo de agrupar um conjunto de nós que, quando combinados, representam uma operação lógica específica. No entanto, existem diferenças significativas entre eles. Por exemplo, diferente dos metanós, os componentes têm a capacidade de possuir uma janela de configuração personalizada, com uma interface de diálogo dedicada e podem ainda apresentar saídas de visualizações interativas e complexas. Além disso, os componentes podem ser reutilizáveis no mesmo fluxo ou até mesmo em fluxos de trabalho separados. Eles podem ser compartilhados com outros usuários por meio do *KNIME Community Hub* (KNIME, 2022a, 2023e, 2023f). A Figura 13 ilustra a distinção visual entre metanós e componentes, ambos em sua representação gráfica antes da execução e após uma execução bem-sucedida.

Figura 13 - Exemplo de metanó e componente



Fonte: Elaborado pela autora (2023).

10.2.6 EXTENSÕES E INTEGRAÇÕES

As extensões e integrações são recursos adicionais que podem ser incorporados ao ambiente *KNIME* para estender suas funcionalidades e integrá-lo com outras tecnologias (KNIME, 2023a, 2023c, 2023e).

As extensões são pacotes de *plugins*[104] que adicionam funcionalidades específicas à plataforma. Esses *plugins* podem ser desenvolvidos pela comunidade ou pela equipe própria *KNIME AG*. Podem incluir nós (nodes) adicionais para processamento de dados, *algoritmos* de aprendizado de máquina, conectores para diferentes fontes de dados, visualizações personalizadas e outras funcionalidades (KNIME, 2023a). As extensões desenvolvidas e mantidas pela *KNIME AG* são de código aberto e disponíveis na página de extensões[105].

Já as integrações referem-se à capacidade da plataforma se integrar com outras ferramentas, tecnologias, linguagens de programação ou serviços. Por exemplo, o *KNIME* pode ser usado na preparação e processamento dados e, em seguida, enviar dados a uma ferramenta de visualização como o *Power BI*[106] para criar gráficos interativos e painéis informativos. As integrações também são desenvolvidas e mantidas pela *KNIME AG*, ou pelos fornecedores das ferramentas que elas integram. As integrações podem ser visualizadas e consultadas na página de integrações[107].

104 *Plugins* são componentes de software modulares e extensíveis que adicionam funcionalidades específicas a uma aplicação ou sistema maior. Eles são desenvolvidos separadamente e podem ser integrados facilmente, permitindo a extensão ou personalização da funcionalidade principal de um *software* sem a necessidade de alterar seu código-fonte fundamental. Em essência, *plugins* são peças de software que expandem as capacidades de um programa principal.

105 Disponível em: https://www.knime.com/knime-extensions. Acesso em: 20 ago. 2023.

106 Para mais detalhes consulte: https://powerbi.microsoft.com/pt-br/. Acesso em: 20 ago. 2023.

107 Disponível em: https://www.knime.com/knime-integrations. Acesso em: 20 ago. 2023.

10.3 INTEGRAÇÃO DO KNIME NA METODOLOGIA DE PESQUISA

A necessidade da adoção de TI para apoiar as várias etapas metodológicas do processo de pesquisa tem se tornado cada vez mais evidente. Na era atual da pesquisa científica, na qual grandes volumes de dados são gerados e necessitam ser analisados, ferramentas avançadas são essenciais para extrair respostas significativas e impulsionar ainda mais o desenvolvimento da ciência. Na primeira parte deste capítulo, foi apresentada a *Plataforma KNIME Analytics*, discutindo suas principais capacidades e funcionalidades. Nesta seção, será explorada como a plataforma pode ser habilmente incorporada na pesquisa científica, destacando suas capacidades.

10.3.1 ETAPAS GERAL DA PESQUISA CIENTÍFICA

A condução de uma pesquisa científica envolve várias etapas cruciais para garantir que os dados coletados sejam confiáveis, relevantes e úteis para responder às perguntas de pesquisa. A seguir foram sintetizadas quatro etapas envolvidas no processo de pesquisa: aquisição ou coleta de dados, preparação ou tratamento de dados, análise de dados e visualização de dados.

10.3.1.1 AQUISIÇÃO/COLETA DE DADOS

A etapa inicial da pesquisa científica engloba a aquisição ou coleta de dados, sendo uma etapa fundamental para reunir informações que servirão como alicerces para todo o processo de pesquisa. Independentemente dos métodos ou técnicas empregados, toda pesquisa demanda a obtenção de dados de uma variedade de fontes. Nessa etapa, os pesquisadores obtêm as informações essenciais necessárias para responder às questões de pesquisa formuladas e para testar as hipóteses estabelecidas. É importante que a coleta de dados seja precisa e bem planejada, uma vez que essa precisão é vital para assegurar a validade e a confiabilidade dos resultados alcançados durante a pesquisa (Gil, 2002, 2008; Lakatos; Marconi, 2010; Marconi; Lakatos, 2002).

A coleta de dados em pesquisas pode variar consideravelmente de acordo com o tipo de pesquisa e as circunstâncias envolvidas. Dentre os procedimentos adotados para essa coleta, destacam-se:

- **Entrevistas:** Entrevistas estruturadas ou semiestruturadas podem ser conduzidas para obter informações de participantes ou especialistas no campo de estudo. Após a transcrição das respostas, os dados são organizados em planilhas, arquivos ou bancos de dados, podem ser codificados e categorizados para identificar padrões e temas;

- **Questionários:** O uso de questionários padronizados é comum para coletar dados de grandes grupos de pessoas de maneira consistente. Esses questionários incluem perguntas fechadas, nas quais os participantes escolhem entre opções predefinidas e perguntas abertas, que permitem respostas em suas próprias palavras. As respostas obtidas são frequentemente registradas em planilhas e bancos de dados para análise posterior;

- **Observação:** Os pesquisadores podem observar eventos, comportamentos ou fenômenos diretamente, registrando suas observações de forma sistemática. Após a coleta, os dados são organizados de forma sistemática descrevendo detalhes do contexto, comportamentos ou eventos relevantes. Os pesquisadores, muitas vezes, utilizam códigos ou símbolos para categorizar as observações. Esses dados podem ser registrados em formato textual, em bancos de dados e planilhas.

É importante ressaltar que as coletas de dados podem ser realizadas em diversas fontes de informação acadêmica, tais como artigos científicos, teses e dissertações. Essas fontes desempenham um papel crucial ao fornecer fundamentação teórica e prática para a pesquisa científica. Além disso, é comum conduzir pesquisas como estudos métricos da informação e revisões sistemáticas de literatura, visando essencialmente examinar a produção científica existente nessas fontes de informação. Tais fontes são categorizadas em três tipos distintos: primárias, secundárias e terciárias, cada uma com suas características específicas. A escolha pelo tipo de fonte é determinada pelos objetivos da pesquisa.

As fontes primárias são aquelas que apresentam informações originais e inéditas, como artigos científicos, teses e dissertações. As fontes secundárias são aquelas que compilam informações de diversas fontes primárias, como bases de dados, revisões sistemáticas e meta-análises. Já as fontes terciárias são aquelas que fornecem informações resumidas e

de fácil acesso, como *catálogos on-line*, manuais, guias, enciclopédias e dicionários (Souza *et al.*, 2022).

Em estudos métricos da informação e revisões sistemáticas de literatura, a coleta e aquisição de dados envolvem métodos específicos para analisar a produção científica existente em uma área de estudo.

10.3.1.2 PREPARAÇÃO E/OU TRATAMENTO DE DADOS

Após a coleta dos dados, os dados devem ser organizados e tratados para posterior análise e interpretação. A preparação ou elaboração dos dados devem seguir os seguintes passos: seleção, codificação, tabulação (Lakatos; Marconi, 2010; Marconi; Lakatos, 1990).

Durante a seleção, são escolhidos os dados pertinentes para a pesquisa, enquanto na codificação, eles são transformados em categorias ou códigos, facilitando sua interpretação quantitativa. Posteriormente, na tabulação, os dados são organizados de forma estruturada em tabelas, preparando-os para análises detalhadas e interpretações significativas. Além disso, a tabulação facilita a representação gráfica dos dados, possibilitando uma compreensão e interpretação mais rápidas.

Em pesquisas que envolvem estudos métricos da informação e revisões sistemáticas de literatura, a etapa de elaboração ou preparação dos dados é fundamental para garantir a confiabilidade e validade das análises. Nessa fase, os pesquisadores realizam uma seleção criteriosa de estudos relevantes e os organizam. Para essa organização, os pesquisadores precisam extrair sistematicamente as informações pertinentes de cada estudo. Isso pode incluir a criação de uma planilha ou banco de dados eletrônico contendo detalhes de cada estudo, como título, autores, instituições afiliadas, palavras-chave, local e ano de publicação, métodos, resultados e conclusões. Durante o processo de preparação dos dados, os pesquisadores também podem realizar análises da qualidade metodológica dos estudos incluídos, corrigindo dados conforme necessário, garantindo, assim, a integridade dos dados utilizados na pesquisa métrica ou revisão sistemática de literatura (Biolchini *et al.*, 2005; Galvão; Ricarte, 2019; Grácio *et al.*, 2020; Kitchenham, 2004).

De acordo com a natureza da pesquisa em questão e a tipologia dos dados coletados, a fase de preparação de dados pode implicar na execução de diversas atividades, tais como:

- **Limpeza de dados:** Identificar, corrigir e remover erros ou valores inconsistentes nos conjuntos de dados para garantir que sejam precisos, confiáveis e adequados para análise estatística. Por exemplo, podem ser tratados erros de erros de digitação e valores extremos (outliers);

- **Tratamento de Valores Ausentes:** Caso haja dados ausentes, os pesquisadores podem decidir preencher esses valores usando técnicas como a imputação de dados (substituindo os valores ausentes por estimativas baseadas em dados existentes) ou remover as entradas com dados ausentes, dependendo da natureza do estudo;

- **Transformação de dados:** Os dados podem ser reformatados, agregados ou transformados para torná-los adequados para análise. Por exemplo, em alguns casos, variáveis contínuas podem ser discretizadas, transformando-as em variáveis categóricas ou ordinais. Ou ainda a inclusão de novas variáveis derivadas dos dados originais para capturar características específicas ou padrões que não são diretamente observáveis nos dados brutos;

- **Padronização:** Garantir que os dados estejam em um formato uniforme e compatível para análise. Por exemplo, as datas podem ser formatadas de maneira consistente (dd/mm/aaaa ou mm/dd/aaaa) e as categorias podem ser padronizadas para evitar redundâncias ou sobreposições;

- **Codificação:** Atribuir códigos ou categorias a dados qualitativos para facilitar sua análise e interpretação. Essa tarefa é especialmente relevante quando se lida com respostas abertas de questionários, entrevistas ou outras formas de dados qualitativos. Por exemplo, a codificação intervalar é uma técnica de codificação de dados em pesquisas qualitativas que envolve atribuir códigos a segmentos específicos de dados que estão dentro de intervalos pré-definidos.

10.3.1.3 ANÁLISE DE DADOS

A etapa de análise de dados na pesquisa científica é um processo complexo que envolve a aplicação de métodos estatísticos, análises qualitativas e técnicas específicas para extrair *insights* significativos dos dados coletados. Na pesquisa científica, a análise e interpretação dos dados representam o núcleo central do processo. A importância dos dados reside não apenas em si mesmos, mas na capacidade de fornecer respostas às investigações propostas (Gil, 2002, 2008; Lakatos; Marconi, 2010; Marconi; Lakatos, 2002).

É crucial definir a técnica de análise apropriada para o desenvolvimento do seu trabalho de acordo com o tipo de pesquisa que está sendo conduzida, seja ela quantitativa ou qualitativa. A pesquisa quantitativa lida principalmente com dados numéricos, assim, as técnicas de análise quantitativa envolvem o uso de métodos estatísticos para analisar padrões e relações nos dados. Já a pesquisa qualitativa, em geral, trabalha com dados categóricos (não numéricos) e emprega técnicas de análise que envolvem interpretar significados, padrões e contextos dos dados (Prodanov; Freitas, 2013).

Nessa fase, os pesquisadores exploram uma diversidade de técnicas tanto qualitativas quanto quantitativas para extrair conhecimento substancial dos dados coletados. Lembre-se de que em alguns estudos, uma abordagem mista, combinando elementos quantitativos e qualitativos, pode ser adotada para obter uma compreensão mais abrangente do fenômeno em estudo.

Em pesquisa de caráter quantitativo, os dados são tratados numericamente para discernir padrões e relações significativas. Nesse sentido é comum a adoção de técnicas estatísticas como: i) análise descritiva ou Análise Exploratória de Dados (AED) para resumir características fundamentais dos dados, como medidas de posição e medidas de dispersão; ii) análise de regressão para examinar relações entre variáveis; iii) e análise de séries temporais para identificar tendências ao longo do tempo.

Já em pesquisas de caráter qualitativo o foco está na compreensão aprofundada e na interpretação dos dados. Algumas das técnicas empregadas são: i) análise de conteúdo que busca significados subjacentes em dados textuais; ii) a modelagem de tópicos que revela temas latentes em grandes conjuntos de documentos ou textos em geral; iii) a análise de rede social

que e examina a estrutura das relações sociais entre indivíduos, grupos ou organizações; iv) análise de sentimentos que é usada para determinar a tonalidade emocional do texto, seja positiva, negativa ou neutra e; v) a análise de discurso que examina o uso da linguagem para entender como as pessoas constroem significado e representam o mundo ao seu redor.

Estudos métricos da informação e revisões sistemáticas de literatura são tipos de pesquisas especializadas que demandam uma análise aprofundada da produção científica em uma área ou assunto de estudo.

A análise de dados em estudos métricos da informação lida com indicadores de produção, de ligação e de citação como variáveis quantitativas e/ou qualitativas. Esses indicadores contribuem para entender não apenas a produção científica, mas também as relações entre pesquisadores, instituições, países e áreas de conhecimento (Prado; Castanha, 2020). Os indicadores são analisados por meio de várias técnicas e métodos estatísticos.

Em pesquisas dedicadas à revisão sistemática de literatura, os pesquisadores têm como objetivo gerar novos conhecimentos a partir dos estudos selecionados para revisão. Esse processo implica em uma leitura sintópica, na qual relações entre os textos são estabelecidas e os dados são transformados em resumos significativos, buscando responder à questão que motivou a revisão (Brizola; Fantin, 2016). Algumas técnicas de análise qualitativa como a modelagem de tópicos e a análise de conteúdo podem ser empregadas para analisar esses estudos. A mineração de texto[108] também pode ser uma técnica automatizada para contribuir na geração dos resultados esperados.

10.3.1.4 VISUALIZAÇÃO DE DADOS

A etapa de visualização de dados em uma pesquisa científica refere-se ao processo de representar informações complexas e volumosas por meio

108 Mineração de textos é uma área de processo de descoberta de conhecimento que fornece técnicas efetivas de descoberta de conhecimento em bases de dados não estruturados, como textos. A mineração de textos é responsável por identificar informações úteis e implícitas nos textos, permitindo a extração de conhecimento e *insights* valiosos (Morais; Ambrósio, 2007).

de gráficos, tabelas, mapas ou outras formas visuais. É uma etapa que interliga intimamente com a análise de dados e o processo de apresentação no relatório de pesquisa.

Uma representação ideal dos resultados da pesquisa envolve a descrição dos dados, geralmente realizada por artefatos gráficos como tabelas, quadros e gráficos, acompanhados por textos explicativos. A adoção de técnicas de visuais de apresentação de resultados proporciona ao leitor uma compreensão e interpretação rápidas da grande quantidade de informações (Gil, 2008; Lakatos; Marconi, 2010).

Gráficos, quadros, tabelas, mapas conceituais, mapas de calor, *Bag of Words* (BoW), diagramas e nuvens de palavras são algumas das técnicas frequentemente usadas para transformar dados brutos em informações significativas e facilmente interpretáveis. Essas técnicas não apenas facilitam a compreensão dos resultados, mas também fornecem evidências visuais que sustentam as conclusões e os argumentos apresentados no relatório.

10.3.1.5 EXECUÇÃO DA PESQUISA CIENTÍFICA EM SÍNTESE

A execução da pesquisa científica é um processo complexo que segue várias etapas essenciais para assegurar a confiabilidade e relevância dos resultados da pesquisa. A pesquisa é executada por meio de quatro etapas cruciais no processo de pesquisa: aquisição de dados, preparação ou tratamento de dados, análise de dados e visualização de dados.

Começando com a coleta precisa de informações de fontes variadas, seguindo para a preparação e tratamento dos dados, envolvendo desde a limpeza até a padronização, e sua subsequente análise quantitativa ou qualitativa. Finalmente, a visualização dos dados, por meio de gráficos, tabelas e outras formas visuais, desempenha um papel crucial na apresentação clara e envolvente dos resultados. Essas etapas, combinadas, transformam dados brutos em conhecimento sólido, contribuindo significativamente para o avanço do conhecimento científico em diversas áreas de estudo.

Todas essas etapas do processo de pesquisa científica podem ser aprimoradas e otimizadas por meio do uso de ferramentas de TI. Com o auxílio de *softwares* especializados, pesquisadores podem coletar dados de forma

mais eficiente, organizar e preparar informações de maneira automatizada, realizar análises estatísticas complexas com precisão e criar visualizações de dados interativas e informativas. Além disso, o uso de ferramentas de TI permite a gestão eficaz de grandes volumes de dados, facilitando a identificação de padrões e tendências significativas. Essas tecnologias desempenham um papel fundamental na aceleração do processo de pesquisa, proporcionando aos cientistas recursos valiosos para conduzir estudos mais detalhados e avançados em diversas áreas do conhecimento.

10.3.2 USO DO *KNIME* EM CADA ETAPAS DA METODOLOGIA DE PESQUISA

A integração de ferramentas de TI tornou-se indispensável para pesquisadores em diversas áreas, especialmente em CSA. O avanço tecnológico revolucionou a maneira como os dados são coletados, analisados e interpretados, proporcionando aos pesquisadores uma gama de recursos poderosos para explorar fenômenos complexos e extrair *insights* valiosos. Nas CSA, em que a compreensão profunda de comportamentos humanos, padrões sociais e dinâmicas organizacionais é fundamental, as ferramentas de TI oferecem uma vantagem significativa.

Nesse contexto, tanto a *plataforma KNIME* quanto ferramentas similares, como o *OpenRefine* e o *Orange,* destacam-se como soluções multifuncionais essenciais para pesquisadores. Nesta seção, explora-se como o *KNIME* se encaixa de forma precisa em cada etapa da metodologia de pesquisa científica. Para ilustrar sua aplicação prática e impacto no campo, apresenta-se exemplos de *workflows* que podem ser usados em casos reais de estudos em CSA.

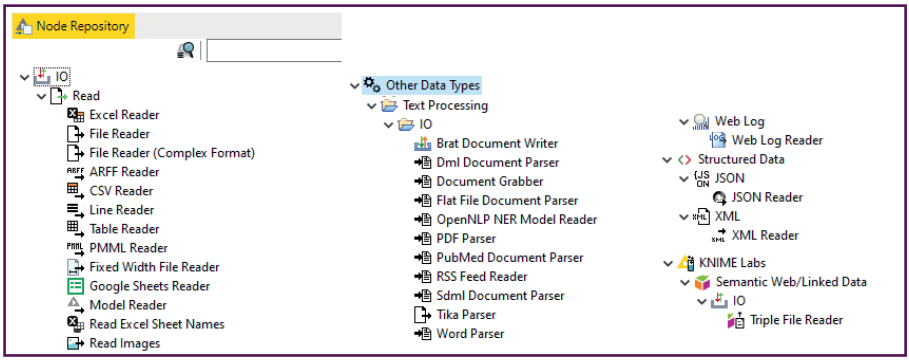
10.3.2.1 COLETANDO DADOS COM O KNIME

O *KNIME* se destaca por sua capacidade de coletar dados de diversas fontes, adaptando-se a diferentes formatos e origens. É possível coletar dados de arquivos localizados na mesma máquina em que a ferramenta está instalada, como arquivos disponíveis remotamente em servidores ou na *web*, por exemplo, arquivos disponíveis no Portal Brasileiro de Dados Abertos. Para arquivos, os pesquisadores podem utilizar nós específicos

para formatos como *CSV, Excel, XML, PDF* e *JSON*, garantindo a integração perfeita de dados de documentos variados.

A Figura 14 apresenta uma lista de nós que podem ser usados para realizar a coleta de dados em arquivos de diferentes formatos. A maior parte dos nós de entrada e saída (*IO* do inglês *In Out*) já são instalados em conjunto com a instalação básica. Entretanto, alguns nós só estarão disponíveis após a instalação de suas respectivas extensões ou integrações. Por exemplo, os nós "*PDF Parser*" e "*Tika Parser*" que fazem a leitura de arquivos em formato *PDF* necessitam da instalação da extensão "*KNIME Textprocessing*", o nó "*XML Reader*" para ler arquivos em formato *XML* e o nó "*JSON Reader*" só estarão disponíveis após a instalação das extensões "*KNIME XML-Processing*" e "*KNIME JSON-Processing*" respectivamente.

Figura 14 -Exemplos de nós para coleta de dados em arquivos



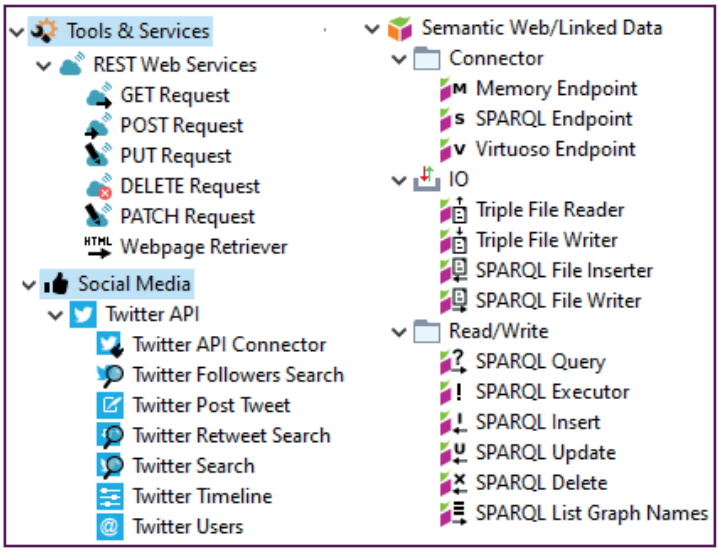
Fonte: Elaborado pela autora (2023).

Em relação a bancos de dados, a extensão "*KNIME Database*" proporciona uma gama de nós para bancos de dados relacionais, como *MySQL, PostgreSQL* e *SQLite*, permitindo a extração eficaz de dados estruturados. Essa extensão permite conectar-se a bancos de dados compatíveis com JDBC (Java Database Connectivity). Não é necessário instalar nenhuma extensão adicional, pois essa extensão já é instalada em conjunto com a instalação básica. Esses nós estão localizados na categoria "*DB*" no repositório de nós, na qual você pode encontrar vários nós para acesso, manipulação e escrita de banco de dados (KNIME, 2023g).

Para lidar com bancos de dados não estruturados, existem extensões e integrações específicas. Por exemplo, a extensão "*KNIME Big Data Extensions*" integra o *Apache Spark* e o ecossistema *Apache Hadoop*. Ela oferece aos usuários a capacidade de acessar uma variedade de bancos de dados, incluindo *Amazon Redshift, H2, Hive, Impala*, entre outros (KNIME, 2023d). Essa extensão pode ser instalada em sua totalidade ou apenas em partes conforme necessidade do pesquisador.

O *KNIME* permite ainda que os pesquisadores raspem dados de *páginas web* por meio de nós específicos para raspagem da *web* (web scraping) ou de redes sociais como *Twitter*. No geral, esses nós são fornecidos por extensões e integrações específicas que não fazem parte da instalação básica. A Figura 15 apresenta uma síntese de nós que podem ser usados para coletar dados disponíveis na *web*.

Figura 15 - Exemplos de nós para coleta de dados na *web* e redes sociais



Fonte: Elaborado pela autora (2023).

Na coluna da esquerda da Figura 15, são visualizados nós das extensões "*KNIME REST Client Extension*" e "*KNIME Twitter Connectors*". A primeira, oferece uma variedade de nós que possibilitam aos pesquisadores interagir de forma direta com serviços *web* que seguem o padrão *REST*. Esses nós

permitem realizar solicitações *HTTP*, enviar parâmetros, receber dados de resposta e manipular informações provenientes de *APIs RESTful*. Por meio do nó "*Webpage Retriever*" é possível recuperar *páginas da web*, emitindo solicitações *HTTP GET* e analisando as *páginas da web HTML* solicitadas. Por padrão, a tabela de saída conterá uma coluna com o *HTML* analisado convertido em *XHTML*. Ao conectar essa saída ao nó *Xpath* do *KNIME*, o pesquisador pode configurar os dados que deseja extrair do documento *XHTML*. Isso é útil quando você precisa acessar dados específicos dentro de documentos *XML* ou *XHTML*. Um exemplo de *workflows* com esses nós foi apresentado na Figura 10.

Já a extensão "*KNIME Twitter Connectors*" oferece aos pesquisadores uma maneira direta de interagir com a *API* (Application Programming Interface) do *Twitter* em seus projetos de pesquisa. Ao utilizar essa extensão, os pesquisadores podem extrair dados do *Twitter*, como *tweets*, informações de perfil de usuário, tendências e muito mais. Os nós disponíveis nessa extensão permitem pesquisas específicas e coleta de dados em tempo real. Nota-se que essa extensão está sujeita às limitações das políticas da plataforma *Twitter*, as quais devem ser observadas pelos usuários. Essas limitações podem incluir restrições no volume de dados que pode ser acessado, bem como nos tipos de dados que podem ser coletados, conforme as diretrizes e políticas de uso do *Twitter*[109]. Tal extensão é bem útil para pesquisas que focam em análise de sentimentos ou mineração da opinião na rede social *Twitter.*

Na coluna da direita da Figura 15, é listado um conjunto de nós específicos para trabalhar *web semântica* e dados interligados (linked data) da extensão "*KNIME Semantic Web*". Com ela, os pesquisadores terão a capacidade de acessar e manipular recursos da *web semântica*, incluindo *endpoints SPARQL*. Assim, os pesquisadores podem explorar dados e informações em ontologias, triplas *RDF*[110] (Resource Description Framework) e outros recursos semânticos na *web*. Isso amplia significativamente as

109 Detalhes no site oficial do *Twitter:* https://developer.twitter.com/en. Acesso em: 25 set. 2023.

110 Triplas RDF: são estruturas de dados fundamentais compostas por três partes: o sujeito (representando o recurso), o predicado (descrevendo a relação) e o objeto (indicando o valor ou recurso), usadas na *web semântica* para a integração e troca de informações semânticas.

possibilidades de pesquisa, permitindo uma análise mais profunda e detalhada de dados relacionados à *web semântica*.

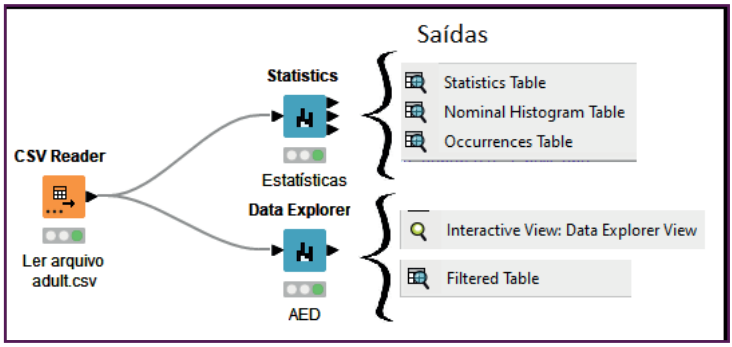
Existem outras extensões e integrações que possuem nós destinados a coletar dados em diferentes fontes de dados, veja a lista completa de extensões[111] disponíveis no *KNIME Community Hub*.

10.3.2.2 PREPARAÇÃO DE DADOS COM KNIME

A preparação de dados é uma etapa fundamental em qualquer projeto de pesquisa científica, pois dados bem-preparados são essenciais para análises precisas e resultados confiáveis. A *plataforma KNIME* oferece diversos nós e extensões que atendem as técnicas para limpar, transformar e organizar dados de acordo com as necessidades dos pesquisadores. Para cada tarefa de preparação pode-se visualizar diversas funcionalidades conforme apresentadas a seguir.

Na **identificação de erros**, é possível realizar uma análise descritiva ou exploratória dos dados por meio da utilização dos nós "*Statistics*" e "*Data Explorer*", conforme exemplo de *workflow* da Figura 16.

Figura 16 - Exemplo de *workflow* para identificação de erros e suas saídas



Fonte: Elaborado pela autora (2023).

111 Veja as extensões disponíveis no *Knime Community Hub* em: https://hub.knime.com/knime/extensions.

Esses nós permitem calcular estatísticas descritivas em colunas numéricas, incluindo valores como contagem, mínimo, máximo, média, mediana, desvio padrão e variância, e medidas de distribuição como assimetria (Skewness) e Curtose (Kurtosis). Ambos os nós ajudam a identificar valores ausentes em colunas nominais e numéricas, fornecendo informações sobre o número de valores ausentes (missing values) em cada variável e ainda para valores numéricos o número de ocorrências do valor 0 (zero). Esses nós ainda calculam o número de valores únicos e a contagem de cada valor nominal presente nas variáveis categóricas. Ele oferece uma visão quantitativa das variáveis nominais, indicando quantas vezes cada categoria aparece no conjunto de dados por meio de gráficos do tipo histograma. O nó "*Statistics*" ainda gera uma tabela de saída de ocorrências indicando para cada valor de cada variável quantas vezes ele ocorre, proporcionando uma compreensão abrangente da distribuição dos dados e ajudando na identificação de possíveis valores discrepantes ou *outliers*. O nó "*Data Explorer*" permite que os usuários interajam com os gráficos. Por exemplo, os usuários podem clicar em uma categoria no gráfico de barras para filtrar os dados e observar como outras variáveis numéricas se comportam em relação a essa categoria específica.

Na tarefa de **Correção de Erros**, diversas estratégias podem ser empregadas de acordo com as necessidades específicas do processo de preparação de dados. No repositório de nós da pasta "*Manipulation*", encontram-se diversas ferramentas designadas para a manipulação eficaz dos dados. Destacam-se alguns nós particularmente úteis nesse contexto, tais como "*Row Filter*", "*Missing Value*", "*Rule Engine*", "*String Replacer*", "*Math Formula*", "*String Manipulation*", "*Duplicate Row Filter*", entre outros que se encontram disponíveis.

Esses nós fornecem uma variedade de funcionalidades, permitindo desde a filtragem de linhas com base em critérios específicos ("Row Filter" e "Duplicate Row Filter"), a substituição de padrões de texto ("*String Replacer*" e "S*tring Manipulation*"). Por exemplo, o nó "*Row Filter*" pode ser empregado para excluir as linhas em que o valor de uma variável específica está ausente ou possui um valor específico.

O nó "*Missing Value*" é um nó mais avançado para lidar com valores ausentes encontrados nas células da tabela de entrada. Ele oferece opções de tratamento padrão para todas as colunas de um tipo específico ou opções

para tratamento individualizado de cada coluna. Por exemplo, imagine um conjunto de dados com uma coluna de idade e algumas entradas estão em branco. É possível preencher os valores ausentes com a média arredondada para valor inteiro das idades existentes.

O nó "*Math Formula*" avalia uma expressão matemática com base nos valores em uma linha. Os resultados computados podem ser adicionados como uma nova coluna ou utilizados para substituir uma coluna de entrada. Por exemplo, em um conjunto de dados com colunas "Preço" e "Quantidade", pode ser criada uma nova coluna chamada "Total" com a fórmula: Preço x Quantidade. Ou ainda, em um conjunto de dados que tenha uma coluna chamada "Data de Nascimento" e deseje calcular a idade das pessoas com base nessa informação.

O nó "*Rule Engine*" permite que o usuário forneça uma lista de regras personalizadas e as aplique a cada linha na tabela de entrada. Quando uma regra é correspondida, o valor do seu resultado é adicionado em uma nova coluna ou substitui o valor resultante na própria coluna da tabela. Por exemplo, em um conjunto de dados de alunos, pode-se criar uma regra para atribuir um status "Aprovado" aos alunos com notas superiores a 60 e um *status* "Reprovado" aos alunos com notas inferiores ou iguais a 60.

O *KNIME* também oferece uma variedade de nós especializados para facilitar a tarefa de **integração de dados**, por exemplo os nós "*Joiner*", "*Concatenate*", "*GroupBy*" e "*Column Combiner*". O nó "*Joiner*" é instrumental na combinação de dados de duas ou mais tabelas com base em chaves específicas. Os dados resultantes são agregados de acordo com as chaves de junção, proporcionando uma visão dos dados combinados. O nó "*Concatenate*" mescla dados verticalmente, empilhando linhas de várias tabelas ou conjuntos de dados. Isso é valioso quando é necessário combinar dados de fontes semelhantes, pois mantém as colunas correspondentes das tabelas de entrada, criando um conjunto de dados contínuo.

O nó "*Column Combiner*" é utilizado para combinar duas ou mais colunas em uma única coluna. Ele é útil quando há necessidade de criar uma nova variável que seja uma combinação ou concatenação de informações de diferentes colunas. Por exemplo, é possível combinar o nome e o sobrenome de uma pessoa em uma única coluna de nome completo. O nó "*GroupBy*" é fundamental para análises agregadas. Ele agrupa os dados

com base em uma ou mais colunas de referência e permite calcular estatísticas ou realizar operações em grupos específicos de dados. Por exemplo, pode-se agrupar dados por região geográfica e calcular a média de vendas em cada região.

Na tarefa de **padronização e normalização,** o *KNIME* oferece um conjunto de opções, além dos nós já apresentados para correção de erros que também podem desempenhar o papel de padronização de dados. Nós como o "*Number To String*" e o "*Category To Number*" realizam a transformação de dados, permitindo a conversão entre tipos de dados diferentes. O nó "*Normalizer*" pode ser utilizado para ajustar os dados a uma escala específica, como a *escala Z* ou a *escala mín-máx*, garantindo que diferentes variáveis estejam na mesma escala para análises justas e equitativas. Em contrapartida, o nó "*Denormalizer*" pode reverter as transformações aplicadas aos dados normalizados.

Esses são apenas alguns exemplos dos nós disponíveis que podem favorecer a pesquisa científica apoiando a tarefa de preparação e elaboração de dados. Seja para lidar com valores ausentes, aplicar transformações complexas, filtrar dados específicos ou combinar diferentes colunas, existe um nó específico para cada cenário. Para encontrar os nós que se adequem às características de suas pesquisas, sugere-se explorar o *KNIME Hub* e o repositório de nós. Dessa forma, pesquisadores podem escolher e utilizar os nós mais adequados para atender às suas necessidades específicas de preparação de dados antes de iniciar suas análises, garantindo, assim, a integridade e a precisão dos dados durante todo o processo de pesquisa.

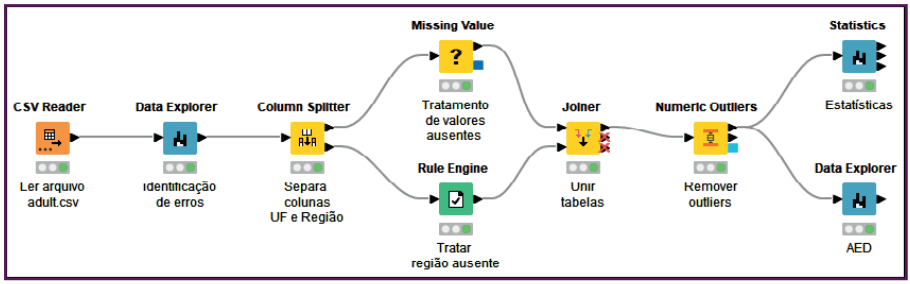
10.3.2.3 ANÁLISE DE DADOS COM *KNIME*

Na tarefa de análise de dados, o *KNIME* oferece funcionalidades que cobrem uma variedade de abordagens, desde métodos qualitativos que exploram nuances subjetivas até métodos quantitativos que se baseiam em dados numéricos e estatísticas precisas.

No contexto da AED, pode-se trabalhar com um *workflow* simples como já foi apresentado anteriormente e ilustrado na Figura 16. Os nós "*Statistics*" e "*Data Explorer*" não só podem ser usados para compreender os dados, mas também podem ser usados para tratar uma abordagem quantitativa

na pesquisa científica. A Figura 17 ilustra um exemplo no qual o nó "*Data Explorer*" primeiro foi usado para identificação de erros e necessidade de tratamento nos dados e, ao final, após a preparação de dados, ele é usado novamente e paralelo ao nós "*Statistics*" com o intuito de gerar uma análise estatística dos dados.

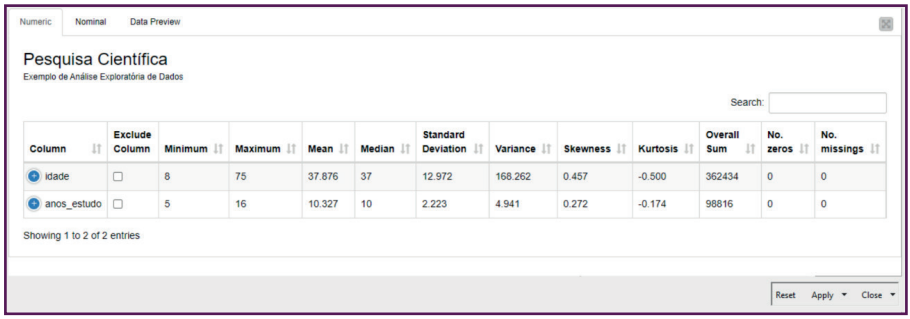
Figura 17 - Exemplo de *workflow* de AED



Fonte: Elaborado pela autora (2023).

O nó "*Data Explorer*" gera como saída um sumário com os resultados separados para as variáveis numéricas e as variáveis nominais ou categóricas. A saída numérica (Figura 18) mostra as principais propriedades estatísticas dos dados numéricos, como mínimo, máximo, mediana (opcional), desvio padrão, variância, assimetria, curtose, soma total, número de valores zero, número de valores ausentes.

Figura 18 - Exemplo de resultado do nó "Data Explorer" para variáveis numéricas


Numeric | Nominal | Data Preview

Pesquisa Científica

Exemplo de Análise Exploratória de Dados

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Median	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings
idade	☐	8	75	37.876	37	12.972	168.262	0.457	-0.500	362434	0	0
anos_estudo	☐	5	16	10.327	10	2.223	4.941	0.272	-0.174	98816	0	0

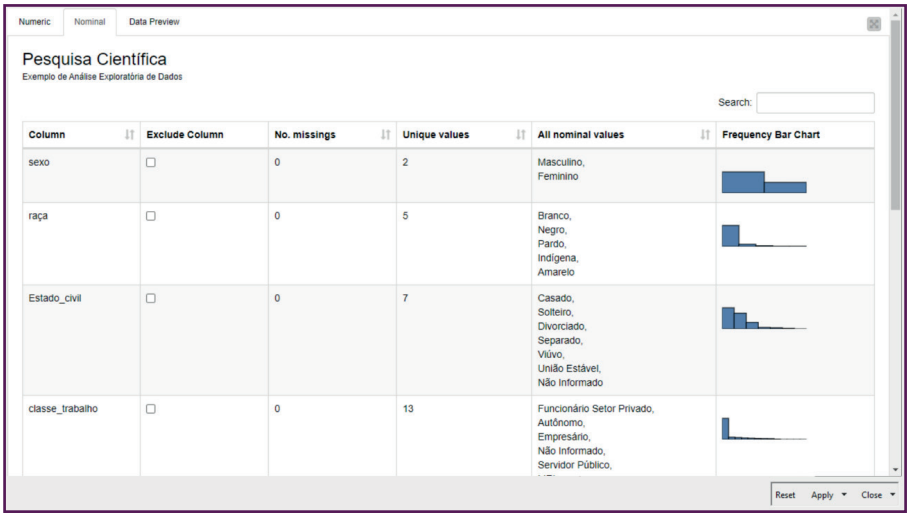
Showing 1 to 2 of 2 entries

Reset | Apply | Close

Fonte: Elaborado pela autora (2023).

A saída nominal (Figura 19) lista as principais propriedades dos valores nominais, como o número de valores ausentes, valores únicos e os *n valores* mais frequentes e menos frequentes. Para cada coluna, o nó também calcula um histograma mostrando a distribuição dos valores numéricos ou as frequências com que os valores nominais ocorrem.

Figura 19 - Exemplo de resultado do nó "*Data Explorer*" para variáveis nominais

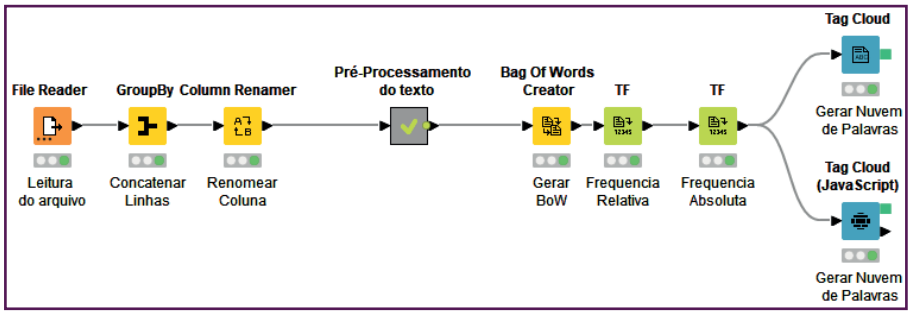


Fonte: Elaborado pela autora (2023).

Além de análises estatísticas, o *KNIME* oferece uma variedade de técnicas de análise de dados por meio de nós específicos. Desde análises descritivas até modelos estatísticos avançados, os pesquisadores podem explorar padrões e relacionamentos em seus dados. Alguns nós e extensões úteis para o pesquisador são apresentadas a seguir.

A extensão "*KNIME Textprocessing*" permite aos pesquisadores realizarem análises de texto dentro do ambiente do *KNIME*. Ela oferece um conjunto de nós para geração de *BoW*, cálculo de frequência de palavras, identificação de entidades nomeadas e etiquetas *Part-of-Speech tags* (POS) e geração de nuvem de palavras, entre outros, conforme exemplo apresentado na Figura 20.

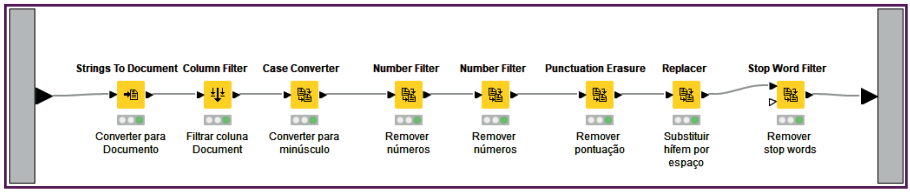
Figura 20 - Exemplo de *workflow* para análise de texto



Fonte: Elaborado pela autora (2023).

Essa extensão também apresenta um conjunto de nós para o pré-processamento do texto que permite realizar tarefas como como *tokenização* (dividir o texto em palavras ou frases), remoção de números, pontuação e *stop words*, *stemming* (reduzir palavras à sua forma raiz) e *lematização* (transformar palavras em suas formas base ou lemas). No exemplo da Figura 20, alguns desses nós foram incluídos no metanó "Pré-Processamento do texto" que apresenta sua visualização expandida na Figura 21.

Figura 21 - Exemplo de metanó "Pré-Processamento do texto" expandido



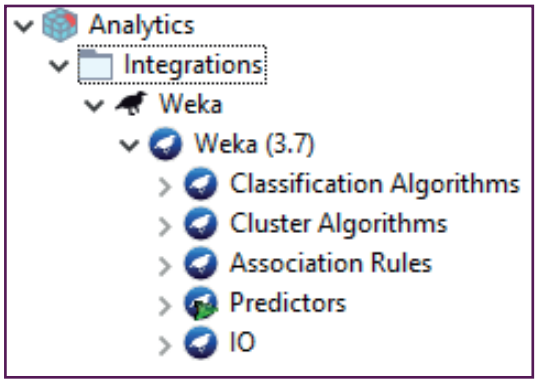
Fonte: Elaborado pela autora (2023).

Essa extensão é especialmente valiosa quando a pesquisa se trata de lidar com dados textuais em análises de texto, mineração de texto, Processamento de Linguagem Natural (PNL) e tarefas relacionadas.

Outra opção interessante é a integração "*KNIME Weka Data Mining Integration*" que incorpora as capacidades do *framework* de mineração de dados *Weka* adicionando ainda mais possibilidades de análise de dados. Essa integração permite aos pesquisadores explorarem uma gama ainda mais ampla de técnicas de mineração de dados e aprendizado de máquina

para descobrir *insights* valiosos para suas pesquisas. Após sua instalação, seus nós estarão disponíveis no repositório de nós na categoria "*Analytics --> Integrations*" conforme ilustrado na Figura 22.

Figura 22 -Exemplo de nós da integra com a ferramenta *Weka*

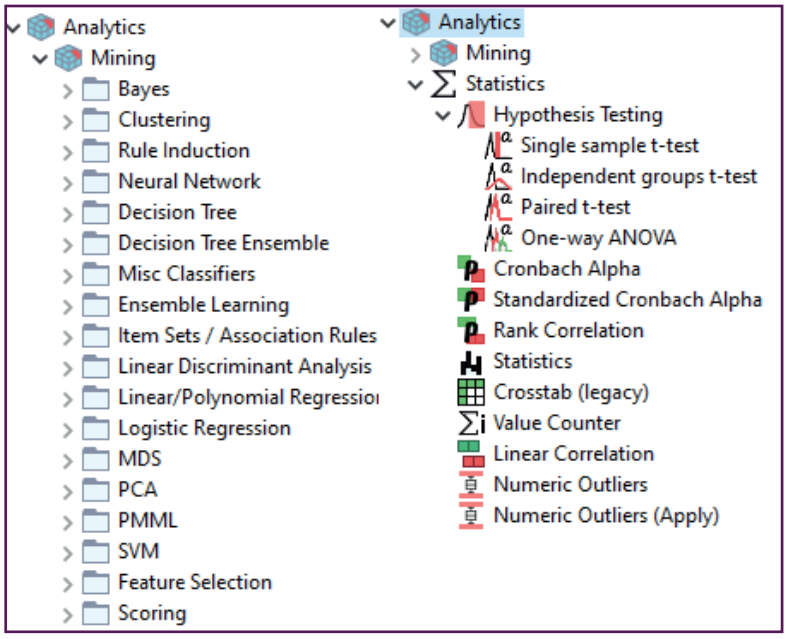


Fonte: Elaborado pela autora (2023).

No repositório de nós (Figura 23), na categoria "*Analytics --> Mining*", encontra-se diversas opções de nós para a análise de dados conforme a necessidade da pesquisa. Esses nós fornecem opções para explorar padrões, fazer previsões e outras possibilidades de análise nos dados. Para um entendimento completo e detalhado de cada nó é recomendável explorar a aba de descrição fornecida pela própria plataforma.

Na categoria "*Analytics --> Statistics*" do repositório de nós, encontra-se um conjunto de nós da extensão "*KNIME Statistics Nodes*" que oferece uma variedade de nós estatísticos para análise de dados (Figura 23). Esses nós possibilitam aos cientistas calcularem estatísticas descritivas precisas, realizar testes de hipóteses significativos e explorar correlações complexas entre variáveis.

Figura 23 - Exemplo de nós para análise de dados



Fonte: Elaborado pela autora (2023).

Além do nó "*Statistics*", observa-se o nó "*Linear Correlation*" para calcular a correlação linear entre duas variáveis contínuas, medindo a força e a direção da relação entre essas variáveis. O nó "*Rank Correlation*" calcula a correlação entre duas variáveis ordinais. O nó "C*ronbach Alpha*" calcula o coeficiente alfa de *Cronbach*, uma medida de consistência interna de um conjunto de itens ou perguntas em um questionário ou escala. Ele é amplamente utilizado na área de pesquisa para avaliar a confiabilidade de um teste ou instrumento de medição. A extensão engloba alguns nós para testes de hipóteses, como teste *t* e Análise de Variância (ANOVA) que ajudam a fazer inferências sobre populações com base em amostras de dados, e nós para análise de regressão linear, permitindo prever variáveis dependentes com base em variáveis independentes.

Para uma compreensão aprofundada de como utilizar o *KNIME* para análise de texto e aplicar essa ferramenta em pesquisas envolvendo análises métricas da informação e revisão sistemática de literatura, explore os *workflows* disponíveis no *KNIME Hub,* por exemplo, os *workflows* "*PubMed Network Analysis*", "*PubMed Literature Search*", "*Creating a Corpus of*

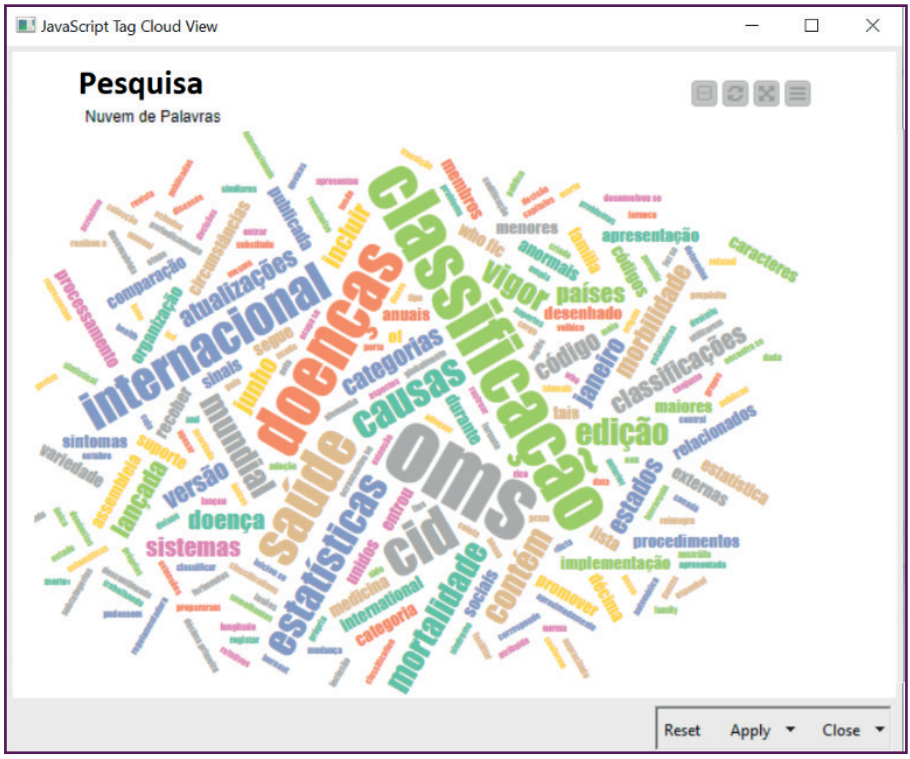
Documents" e "*Text mining techniques in life sciences literature*". Esses recursos são úteis para pesquisadores que necessitam analisar conjuntos de dados textuais, como artigos científicos e literatura acadêmica. Ao estudar esses *workflows*, os cientistas podem aprender técnicas avançadas de mineração de texto, análise de redes e criação de corpora de documentos, necessárias para realizar análises métricas detalhadas da informação e revisões sistemáticas de literatura de forma eficaz.

Pesquisadores que desejam realizar análises de modelagem de tópicos, por exemplo usando a técnica *Latent Dirichlet Allocation* (LDA), os *workflows* "*Topic Modeling on Biomedical Literature*", "*Topic Modeling with LDA*", "*Identify and visually represent a topic model*", "*Topic Detection LDA: Summarizing Romeo & Juliet or cataloging News*" e "*IS Literature Mining with Topic Detection*" são úteis. Para uma introdução prática e detalhada sobre como realizar análises de sentimentos e trabalhar com dados do *Twitter* e análise de sentimentos, existem diversos modelos de fluxos de trabalho disponíveis no *KNIME Hub*, por exemplo: "*Sentiment Analysis*", "*Twitter Data Collection*" e "*Twitter Data Analysis*".

10.3.2.4 VISUALIZAÇÃO DE DADOS COM KNIME

Na tarefa de visualização de dados, a *plataforma KNIME* oferece uma ampla gama de nós para visualização com gráficos, tabelas e outros. No *workflow* da Figura 20, observa-se o nó "*Tag Cloud* (Java Script)". Como visto na Figura 24, ele cria nuvens de palavras, nas quais o tamanho e a cor das palavras são determinados pela frequência dos termos no conjunto de dados.

Figura 24 - Exemplo de visualização de nuvem de palavras

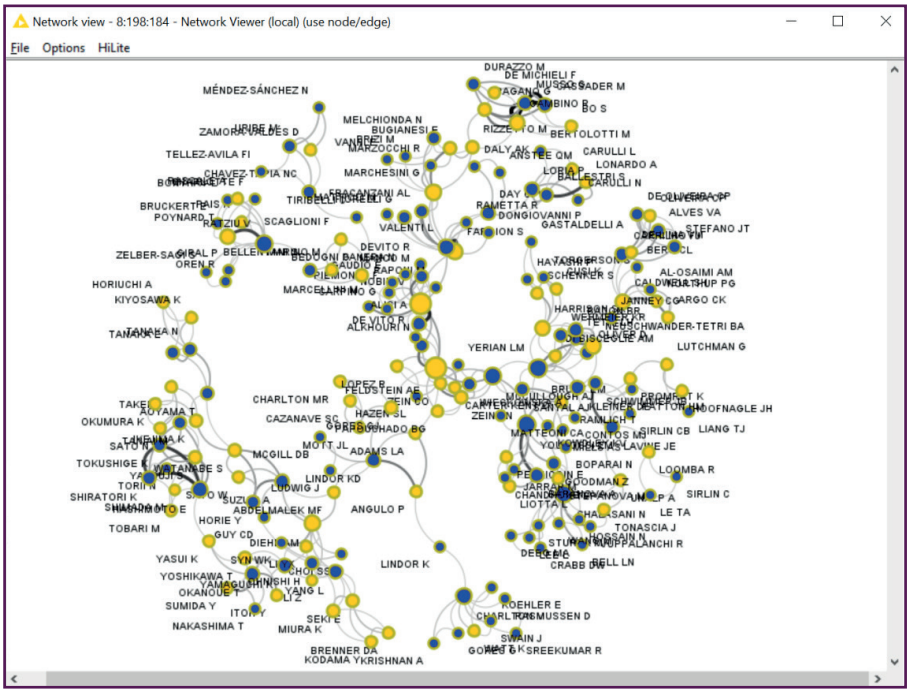


Fonte: Elaborado pela autora (2023).

Outra visualização muito usada em pesquisas que necessitam construir *gráficos de rede* para mostrar conexões entre diferentes entidades em um conjunto de dados é criada com o nó "*Network Viewer* (local)". Ele permite a criação desse tipo de visualização, conforme ilustrado na Figura 25, um exemplo de uma rede de colaboração de autores extraída do *workflow* "*PubMed Network Analysis*" [112].

112 Disponível em: https://hub.knime.com/knime/spaces/Examples/latest/08_Other_Analytics_Types/05_Network_Mining/07_Pubmed_Author_Network_Analysis~wWLJp6ofl-qo9VnUq. Acesso em: 20 set. 2023.

Figura 25 - Exemplo de um gráfico de rede de colaboração

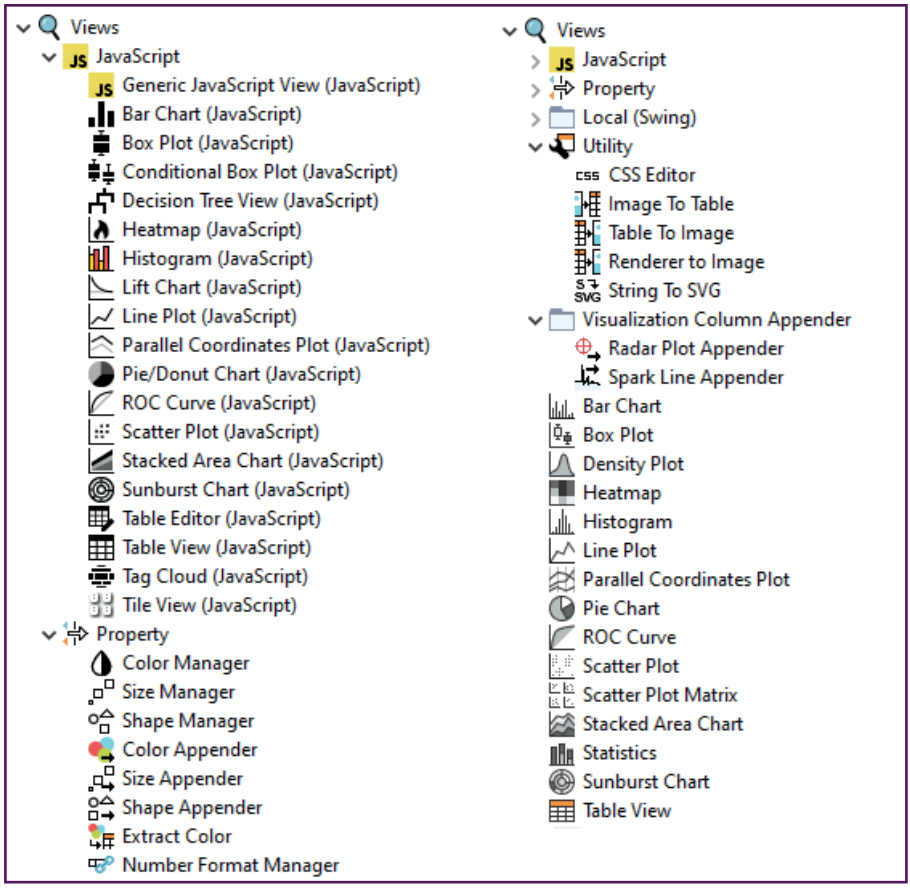


Fonte: Elaborado pela autora (2023).

Na categoria "*Views*" (Visualizações) do repositório de nós, existe uma variedade de nós dedicados à visualização de dados (Figura 26). Esses nós são projetados para criar diferentes tipos de gráficos, tabelas e outras representações visuais dos dados em seus fluxos de trabalho. Eles oferecem opções para personalizar a aparência e a interatividade das visualizações.

O nó "*Table View* (JavaScript)" permite exibir dados em formato de tabela. Ele é útil para uma inspeção detalhada dos dados tabulares, exibindo as informações em um formato claro e organizado, como as tabelas de saída do nós "*Data Explores*" da Figura 18 e Figura 19.

Figura 26 - Exemplo de nós para análise de dados



Fonte: Elaborado pela autora (2023).

Essa lista de nós permite criar diferentes tipos de gráficos, alguns deles que podem ser gerados no *KNIME* e os respectivos nós disponíveis incluem:

- Gráficos de Barras: Permitem comparar a frequência ou proporção de diferentes categorias de dados usando barras verticais ou horizontais. Podem ser usados os nós "*Bar Chart*";

- Gráficos de Linhas: São ideais para mostrar tendências ao longo do tempo ou em uma série de dados sequenciais. Podem ser criados pelos nós "*Line Plot*";

- Gráficos de Dispersão: Mostram a relação entre duas variáveis, sendo úteis para identificar padrões ou correlações nos dados e gerados pelos nós "*Scatter Plot*";

- Gráficos de Pizza (Pie) e de Rosca (Donut): *Gráficos de pizza* representam partes de um todo usando setores circulares, usados para mostrar proporções percentuais. Os *gráficos de Rosca* são semelhantes aos de pizza, mas com um buraco no centro, permitindo mostrar várias categorias e subcategorias. Podem ser usados os nós "*Pie Chart*" ou "*Pie/Donut Chart*" e, em sua janela de configuração, deve-se optar por uso o tipo pizza ou rosca;

- Histogramas: Exibem a distribuição de uma variável contínua por meio de barras verticais, mostrando a frequência de valores em intervalos específicos, criados pelos nós "*Histogram*";

- Gráficos de Caixa: Mostram a distribuição estatística dos dados, incluindo a mediana, quartis e possíveis *outliers*. Podem ser usados os nós "*Box Plot*";

- Mapas de Calor: Representam dados em uma matriz de cores, sendo úteis para visualizar padrões em grandes conjuntos de dados. Os nós "*Heatmaps*" são responsáveis por gerar essa visualização;

- Gráficos de Função de Densidade: Exibem a distribuição de dados contínuos, mostrando onde os dados são mais densos. Pode ser criado usando os nós "*Density Plot*";

- Gráfico de mosaico (treemap): Nesse gráfico, a hierarquia dos dados é representada por retângulos aninhados e cada retângulo equivale a uma categoria específica, sendo o tamanho do retângulo proporcional à quantidade ou ao valor numérico associado a essa categoria. O nó "*Tile View* (JavaScript)" pode gerar um gráfico desse tipo;

- Gráfico de Área Empilhada: Esse tipo de gráfico é útil para mostrar a composição e a tendência de múltiplas séries de dados ao longo do tempo ou de outras variáveis. Cada série é representada como uma área colorida no gráfico, empilhada sobre a série anterior. O nó "*Stacked Area Chart*" permite criar visualizações desse tipo.

Além dessas opções, algumas integrações como "*KNIME Power BI Integration*" e "*KNIME Tableau Integration*" são úteis para quem deseja usar o *KNIME* apenas para coleta, preparação e análise, e deixar as visualizações para outras ferramentas que permitem uma apresentação ainda mais rica e dinâmica.

A visualização de dados desempenha um papel importante na interpretação e comunicação de resultados de pesquisa em diversos contextos acadêmicos, desde artigos científicos a dissertações e teses. Antes de usar a ferramenta, é essencial que os pesquisadores compreendam como seus dados devem ser representados visualmente para transmitir seus resultados de pesquisa de forma clara. Após definir os requisitos visuais específicos para seu relatório de pesquisa, recomenda-se explorar o repositório *KNIME Hub* para encontrar recursos que atendam às suas necessidades específicas.

10.3.3 *KNIME COMMUNITY HUB*

O *KNIME Community Hub* ou apenas *KNIME Hub* é um repositório on-line integrado a *KNIME*, dedicado a reunir e compartilhar conhecimentos e recursos criados pela comunidade global de usuários da ferramenta. Esse repositório serve como um depósito centralizado, no qual pesquisadores e cientistas de dados podem acessar uma variedade de fluxos de trabalho, nós, componentes e extensões (KNIME, 2023e).

Para os pesquisadores, o *KNIME Community Hub*[113] oferece cerca de 18250 fluxos de trabalho, 4504 nós, 1503 componentes e 239 extensões. Eles podem encontrar fluxos de trabalho prontos para uso, desenvolvidos por especialistas em diferentes áreas de pesquisa. Além disso, há uma variedade de nós, componentes e extensões que podem ser incorporados diretamente nos fluxos de trabalho dos pesquisadores, permitindo-lhes criar soluções personalizadas para suas pesquisas científicas.

113 Disponível em: https://hub.knime.com/. Acesso em: 20 set. 2023.

10.4 CONSIDERAÇÕES FINAIS

O presente capítulo mergulhou nas possibilidades oferecidas pela *Plataforma KNIME Analytics*, destacando-a como uma ferramenta de valor para ser usada nas pesquisas científicas, sobretudo na área de CSA. Em um cenário no qual as tecnologias de informação e comunicação desempenham um papel central, esse estudo demonstrou de forma clara o papel dessa plataforma na coleta, preparação, análise e visualização de dados durante as etapas metodológicas de pesquisas acadêmicas.

Em resumo, o *KNIME* se sobressai por sua capacidade de integração, flexibilidade e variedade de extensões, tornando-o uma boa escolha para pesquisadores em diversos campos de pesquisa. Sua flexibilidade também é uma vantagem fundamental, permitindo que os pesquisadores adaptem suas metodologias de acordo com as necessidades específicas de seus projetos. Ao oferecer suporte em todas as fases da pesquisa científica, a plataforma possibilita uma abordagem mais eficiente e completa para a análise de dados, economizando tempo valioso que pode ser direcionado para a análise e interpretação dos resultados e para descobertas significativas.

Entretanto, enquanto celebramos as vantagens da plataforma, é imperativo reconhecer os desafios éticos e metodológicos associados ao uso da TI em pesquisas em CSA. A preservação da privacidade dos dados, a avaliação rigorosa da confiabilidade das fontes *on-line* e a interpretação cuidadosa dos resultados são questões que devem permanecer no centro das discussões dos pesquisadores. A tecnologia é uma aliada poderosa, mas, somente quando associada a uma abordagem ética e crítica, ela pode verdadeiramente impulsionar a pesquisa científica para novos horizontes.

REFERÊNCIAS

BERTHOLD, Michael R *et al*. KNIME: the konstanz information miner. In: **4th Annual Industrial Simulation Conference (ISC)**, Palermo, Itália: CentAUR, p. 58-61, 2006. Disponível em: https://centaur.reading.ac.uk/6139/1/2006_DiFatta06-MASS-ISC.pdf. Acesso em: 10 ago. 2023.

BERTHOLD, M. R. *et al*. KNIME - the Konstanz information miner: version 2.0 and beyond. **Acm Sigkdd Explorations Newsletter**, [*s. l.*], v. 11, n. 1, p. 26-31, 16 nov. 2009. Association for Computing Machinery (ACM). DOI: https://doi.org/10.1145/1656274.1656280. Disponível em: https://dl.acm.org/doi/10.1145/1656274.1656280. Acesso em: 10 out. 2023.

BIOLCHINI, J. *et al*. Systematic review in software engineering. **Technical report**, Rio de Janeiro, v. 679, n. 05, p. 45, May 2005. Disponível em: https://www.cos.ufrj.br/uploadfile/es67905.pdf. Acesso em: 10 out. 2023.

BRIZOLA, Jairo; FANTIN, Nádia. Revisão da literatura e revisão sistemática da literatura. **Revista de Educação do Vale do Arinos**, Juara, v. 3, n. 2, p. 23-39, jul./dez. 2016. Disponível em: https://periodicos.unemat.br/index.php/relva/article/view/1738/1630. Acesso em: 10 ago. 2023.

GALVÃO, Maria Cristiane Barbosa; RICARTE, Ivan Luiz Marques. Revisão sistemática da literatura: conceituação, produção e publicação. **Logeion**: Filosofia da Informação, [*s. l.*], v. 6, n. 1, p. 57-73, 15 set. 2019. Disponível em: https://revista.ibict.br/fiinf/article/view/4835/4187. Acesso em: 10 out. 2023.

GARCIA, Ana Cristina B. *et al*. Groupware 4.0: avanços e desafios da computação social. **Jornada de Atualização em Informática**. Porto Alegre: SBC, 2020, v. 39, p. 142-186. Disponível em: https://sol.sbc.org.br/livros/index.php/sbc/catalog/view/57/252/492-1. Acesso em: 11 out. 2023.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GRÁCIO, M. C. C. *et al*. **Tópicos da bibliometria para bibliotecas universitárias**. São Paulo: UNESP, 2020.

HAYASAKA, S.; SILIPO, R. **KNIME Analytics Platform for Beginners**. Zurich, Switzerland: KNIME Press, 2023.

KHODNENKO, Ivan *et al*. **A Lightweight Visual Programming tool for Machine Learning and Data Manipulation**. Las Vegas: IEEE, 2020, p. 981-985. Disponível em: https://ieeexplore.ieee.org/abstract/document/9458229. Acesso em: 10 out. 2023.

KITCHENHAM, Barbara. **Procedures for Performing Systematic Reviews**. Keele: Keele University, 2004, v. 33, 28 p. Disponível em: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf. Acesso em: 11 out. 2023.

KNIME. **KNIME Best Practices Guide**. Version 5.1. Zurich, Switzerland: KNIME AG, 2022a, p. 20. Disponível em: https://docs.knime.com/latest/analytics_platform_best_practices_guide/analytics_platform_best_practices_guide.pdf. Acesso em: 11 out. 2023.

KNIME. **KNIME File Handling Guide**. Version 5.1. Zurich, Switzerland: KNIME AG, 2022b, p. 41. Disponível em: https://docs.knime.com/latest/analytics_platform_file_handling_guide/analytics_platform_file_handling_guide.pdf. Acesso em: 11 out. 2023.

KNIME. **Extensions and Integrations Guide**. Version 5.1. Zurich, Switzerland: KNIME AG, 2023a, p. 4. Disponível em: https://docs.knime.com/latest/analytics_platform_extensions_and_integrations/analytics_platform_extensions_and_integrations.pdf. Acesso em: 11 out. 2023.

KNIME. **KNIME Analytics Platform Installation Guide**. Version 5.1. Zurich, Switzerland: KNIME AG, 2023b, p. 29. Disponível em: https://docs.knime.com/latest/analytics_platform_installation_guide/analytics_platform_installation_guide.pdf. Acesso em: 11 out. 2023.

KNIME. **KNIME Analytics Platform User Guide**. Version 5.1. Zurich, Switzerland: KNIME AG, 2023c, p. 50. Disponível em: https://docs.knime.com/latest/analytics_platform_user_guide/analytics_platform_user_guide.pdf. Acesso em: 11 out 2023.

KNIME. **KNIME Big Data Extensions User Guide**. Version 5.1. Zurich, Switzerland: KNIME AG, 2023d, p. 19. Disponível em: https://docs.

knime.com/latest/bigdata_extensions_user_guide/bigdata_extensions_user_guide.pdf. Acesso em: 11 out 2023.

KNIME. **KNIME Community Hub User Guide**. Version 1.6. Zurich, Switzerland: KNIME AG, 2023e, p. 35. Disponível em: https://docs.knime.com/2023-07/hub_user_guide/index.pdf. Acesso em: 11 out. 2023.

KNIME. **KNIME Components Guide**. Version 5.1. Zurich, Switzerland: KNIME AG, 2023f, p. 59. Disponível em: https://docs.knime.com/latest/analytics_platform_components_guide/analytics_platform_components_guide.pdf. Acesso em: 11 out.2023.

KNIME. **KNIME Database Extension Guide**. Version 5.1. Zurich, Switzerland: KNIME AG, 2023g, p. 89. Disponível em: https://docs.knime.com/latest/db_extension_guide/db_extension_guide.pdf. Acesso em: 11 out. 2023.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos da metodologia científica**. 7. ed. São Paulo: Atlas, 2010.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Metodologia do trabalho científico**: procedimentos básicos, pesquisa bibliográfica, projeto e relatório, publicações e trabalhos científicos. 3. ed. São Paulo: Atlas, 1990.

LEONELLI, S. **A pesquisa científica na era do Big Data**: cinco maneiras que mostram como o big data prejudica a ciência, e como podemos salvá-la. Rio de Janeiro: FIOCRUZ, 2022.

MARCONI, Marina de. Andrade; LAKATOS, Eva Maria. **Técnicas de pesquisa**. 5. ed. São Paulo: Atlas, 2002.

MINAYO, M. C. S. (org.). **Pesquisa Social**: teoria,método e criatividade. 21. ed. Petrópolis: Vozes, 2002.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007. Disponível em: https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf. Acesso em: 11 out. 2023.

MOURA, A. L. T.; AMORIM, D. G. Big Data: o impacto e sua funcionalidade na sociedade tecnológica. **Revista Opara**, Petrolina,, v. 4, n. 1, p. 53-64, 2014.

MUELLER, S. P. M. O impacto das tecnologias de informação na geração do artigo científico: tópicos para estudo. **Ciência da Informação**, Brasília, DF, v. 23, n. 3, p. 309-317, set./dez. 1994. DOI: https://doi.org/10.18225/ci.inf.v23i3.528. Disponível em: https://revista.ibict.br/ciinf/article/view/528. Acesso em: 11 out. 2023.

PRADO, M. A. R. do; CASTANHA, R. C. G. Indicadores: conceitos fundamentais e importância em CT&I. *In*: GRÁCIO, M. C. C. (org.). **Tópicos da bibliometria para bibliotecas universitárias**, São Paulo: Cultura Acadêmica, 2020, p. 50-71.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013.

SOUZA, A. D. *et al.* A tipologia das fontes de informação em saúde: suporte à tomada de decisão. **Asklepion**: Informação em Saúde, Rio de Janeiro, v. 2, n. 1, p. 51-74, 28 jun. 2022. Disponível em: https://brapci.inf.br/index.php/res/v/198107. Acesso em: 11 out. 2023.

SWIECH, C.; FRANCISCO, A. C. A.; LIMA, S. A. A tecnologia da informação e comunicação transformando e inovando a prática da pesquisa científica. **Revista ESPACIOS**, [*s. l.*], v. 37, n. 11, p. 14, 30. jan. 2016. Disponível em: https://www.revistaespacios.com/a16v37n11/16371115.html. Acesso em: 11 out. 2023.

DADOS DA AUTORA:

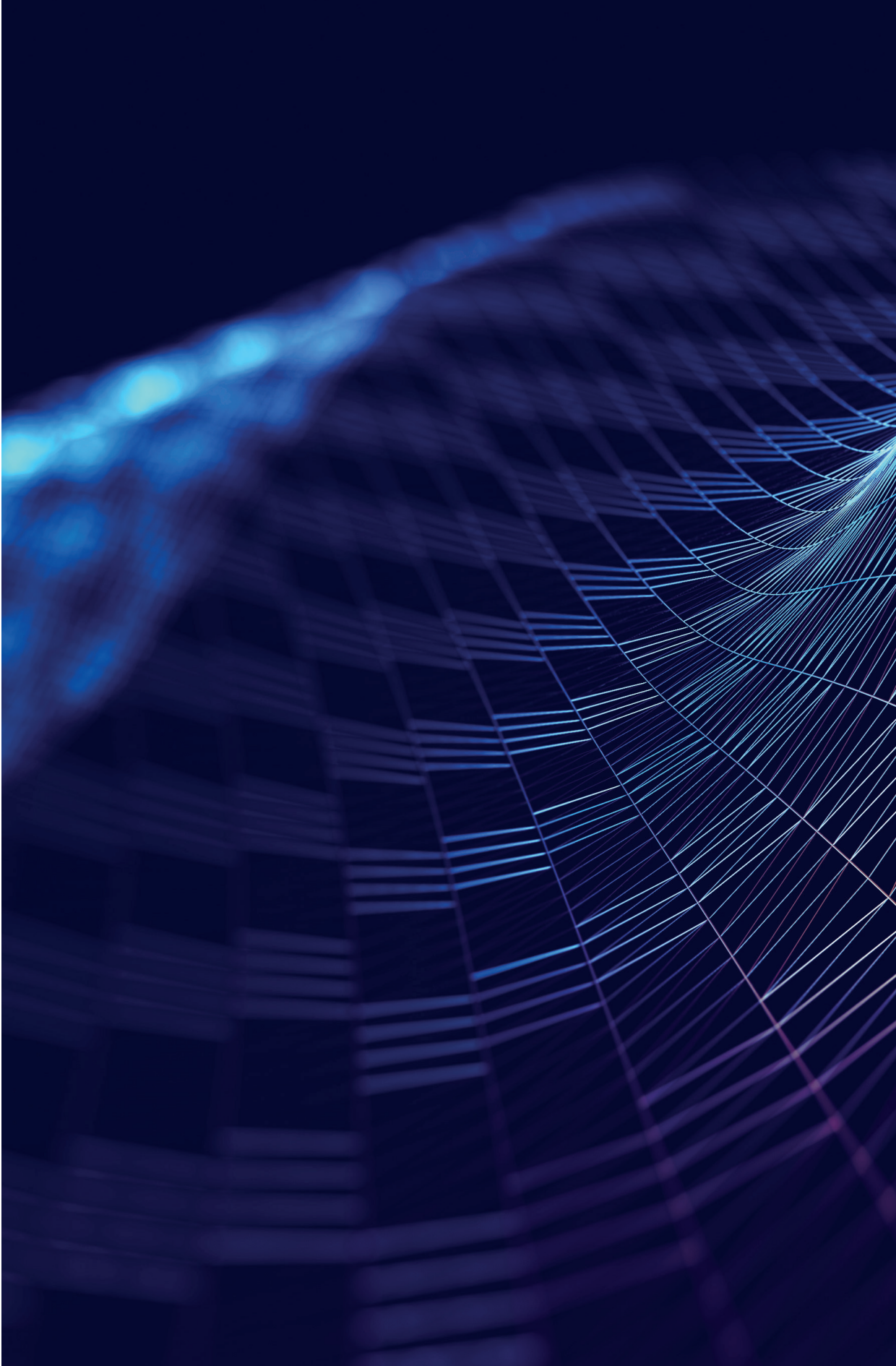
Fernanda Farinelli

Fernanda Farinelli é Professora Adjunta na Faculdade de Ciência da Informação da UnB. Doutora em Gestão e Organização do Conhecimento pela Escola de Ciência da Informação da UFMG pesquisando o tema ontologias formais realistas como solução de integração semântica de dados. Ontologista responsável pelo projeto da OntONeo (Ontologia do domínio obstétrico e neonatal). Pesquisadora visitante no Departamento de Filosofia e no Departamento de Informática Biomédica da Universidade Estadual de Nova York em Buffalo entre 05/2015 e 04/2017. Mestre em Administração de Empresas com ênfase em Gestão estratégica da informação (Fundação Pedro Leopoldo/MG). Especialista em Banco de Dados (UNI-BH). Bacharel em Ciência da Computação (PUC-MG). Possui mais de 15 anos de experiência em Gestão de Dados atuando com administração de banco de dados, arquitetura e administração de dados e implantação de governança de dados em grandes empresas como Unisys Brasil, Cedro Têxtil, Prodemge. Atua há cerca de 15 anos como docente em cursos de graduação e pós-graduação em renomadas instituições de ensino no estado de Minas Gerais como PUC-MG, IEC, Fundação Pedro Leopoldo, Universidade de Itaúna, Faculdade Cotemig, Unipac e IGTI. Possui as certificações CDMP, CBIP, CDP e OCP.

https://orcid.org/0000-0003-2338-8872
fernanda.farinelli@unb.br

Como referenciar o capítulo 10:

FARINELLI, Fernanda. Revolucionando a pesquisa científica com a plataforma KNIME Analytics. *In*: SHINTAKU, Milton; MACÊDO, Diego José; MARIN, Luciano Heitor Gallegos (org.). **Tecnologias utilizadas em pesquisas acadêmicas em Ciências Sociais Aplicadas.** Brasília, DF: Ibict, 2023. cap. 10. p. 275-326. ISBN 978-65-89167-94-5. DOI: http://doi.org/10.22477/9786589167938cap10.



Brasília
2023

O uso de tecnologias livres na metodologia de pesquisa tem se tornado comum em todas as etapas da pesquisa, desde a coleta até o apoio à análise dos dados. Com isso, torna-se essencial uma obra que busque apresentar algumas ferramentas a serem utilizadas na pesquisa das ciências sociais aplicadas, com vistas a apoiar estudos voltados às necessidades de processamento de dados. Assim, este trabalho, resultado de ações do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), colabora com a comunidade de pesquisadores da área de ciências sociais aplicadas, na qual está contida a Ciência da Informação, e se alinha à missão da instituição, que é promover a competência, o desenvolvimento de recursos e a infraestrutura de informação em ciência e tecnologia para a produção, socialização e integração do conhecimento científico e tecnológico.

Tiago Emmanuel Nunes Braga
Diretor do Instituto Brasileiro de Informação em Ciência e Tecnologia

Editora Ibict
ibict Instituto Brasileiro de Informação em Ciência e Tecnologia
UNIDADE DE PESQUISA DO MCTI
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

ISBN 978-65-89167-93-8
ISBN: 978-65-89167-93-8
CBL
9 786589 167938