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.

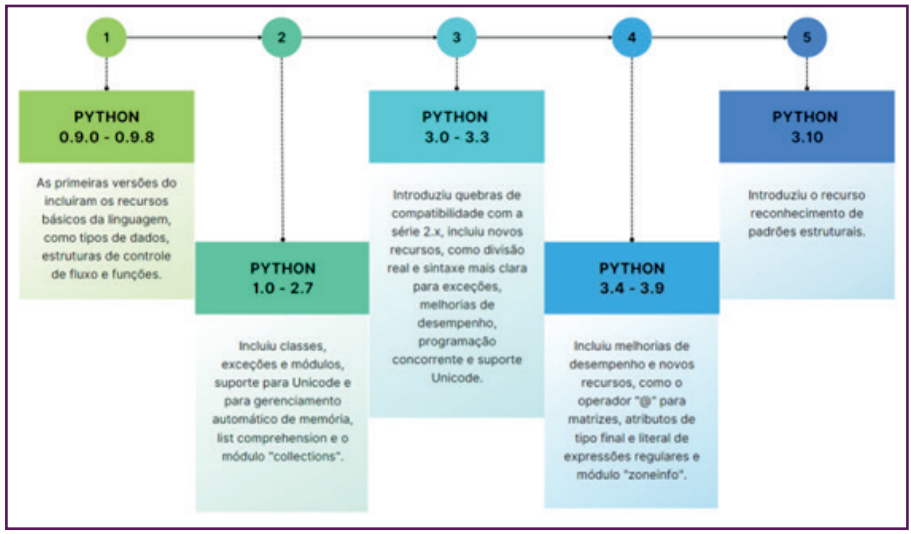
14 Python Brasil 2023: https://2023.pythonbrasil.org.br/#evento. Acesso em: 18 set. 2023.

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.

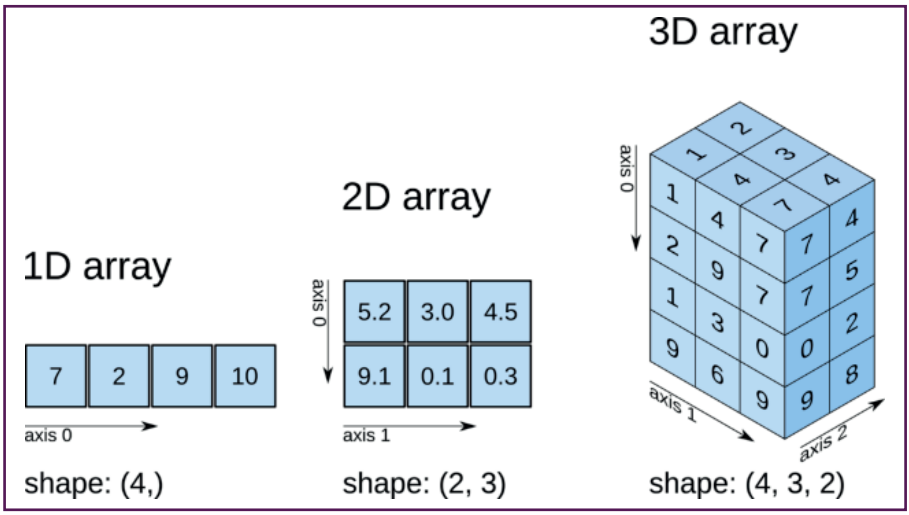
3.3 A TECNOLOGIA E A PESQUISA

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.

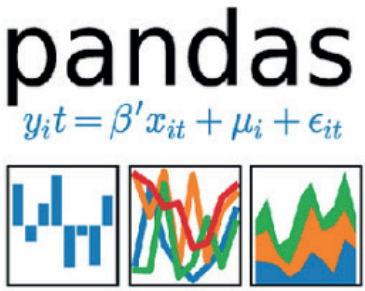
Os autores Galli *et al.* (2022) apresentam uma estrutura de referência para analisar e avaliar técnicas de aprendizado de máquina e aprendizado profundo para detectar notícias falsas, com foco na detecção precoce e na análise do conteúdo das notícias. Os autores abordam os desafios da

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.

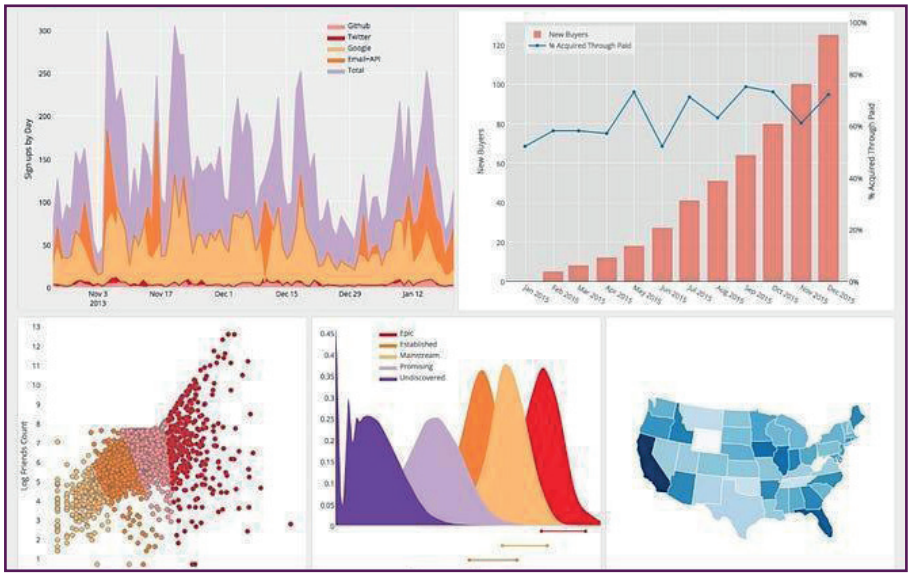
O *Pandas* é bastante utilizado nas tarefas de processamento de dados, tais como limpeza, manipulação e análise, uma vez que dispõe de vários módulos para leitura, processamento e gravação de arquivos *CSV* (Comma-separated values), *JSON* (JavaScript Object Notation) e *Excel* (Microsoft Excel). Obviamente 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.

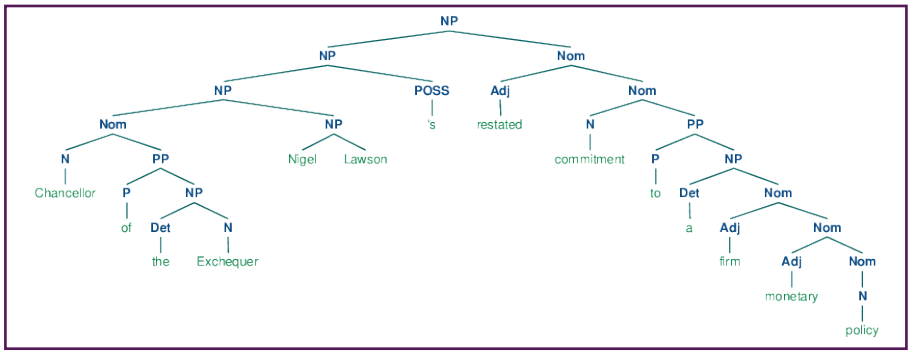
23 Disponível em: https://i.imgur.com/2AE3gch.png. Acesso em: 2 out. 2023.

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].

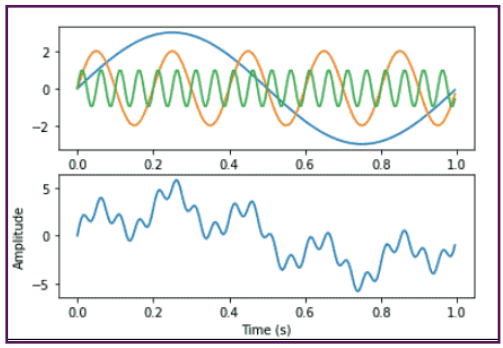
O *SciPy*[28] é um pacote para computação científica que fornece ferramentas de integração numérica, otimização, interpolação, processamento de sinais e muito mais. É útil em áreas como economia, psicologia e geografia. A Figura 6 apresenta uma visualização, usando *Matplotlib*, de um exemplo de transformada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

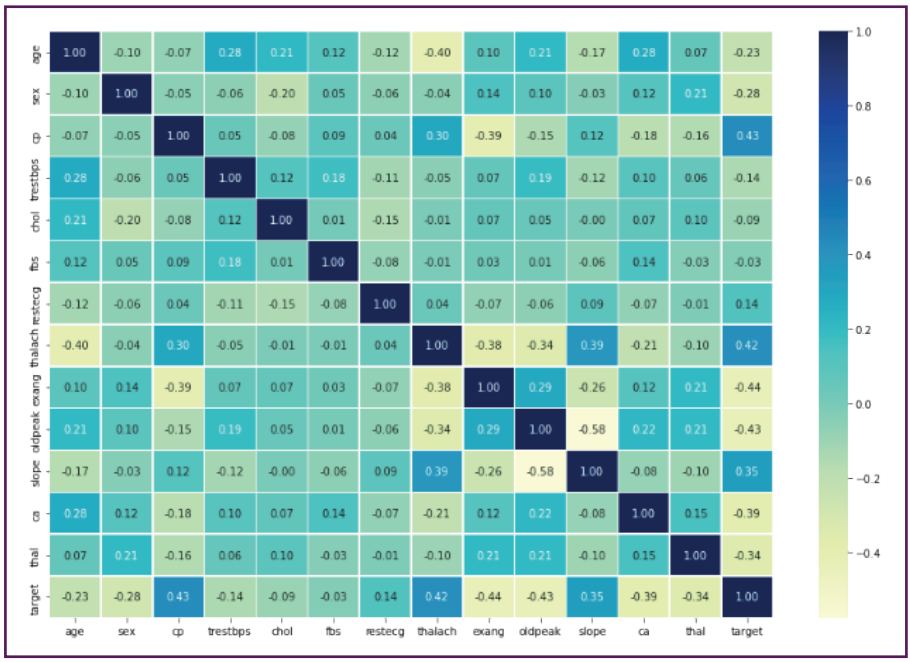
O estudo de Aparício, Romão e Costa (2022) propôs um modelo preditivo para a oscilação de valores de *Bitcoin*. Em uma das etapas, os autores precisaram construir um modelo de estimativa baseado em regressão linear pelo método dos mínimos quadrados utilizando recursos combinados do *Pandas* e do *Statsmodels*.

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

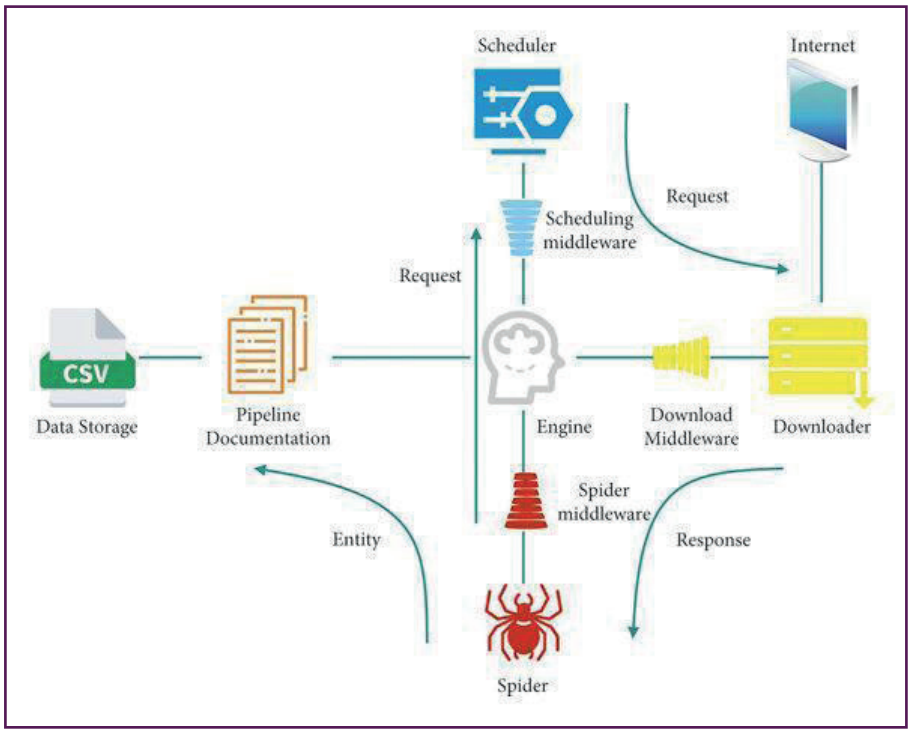
33 *INFOSLACK*. Disponível em: https://infoslack.pro/ml-book/contents/ml-sklearn.html. Acesso em: 3 out. 2023.

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



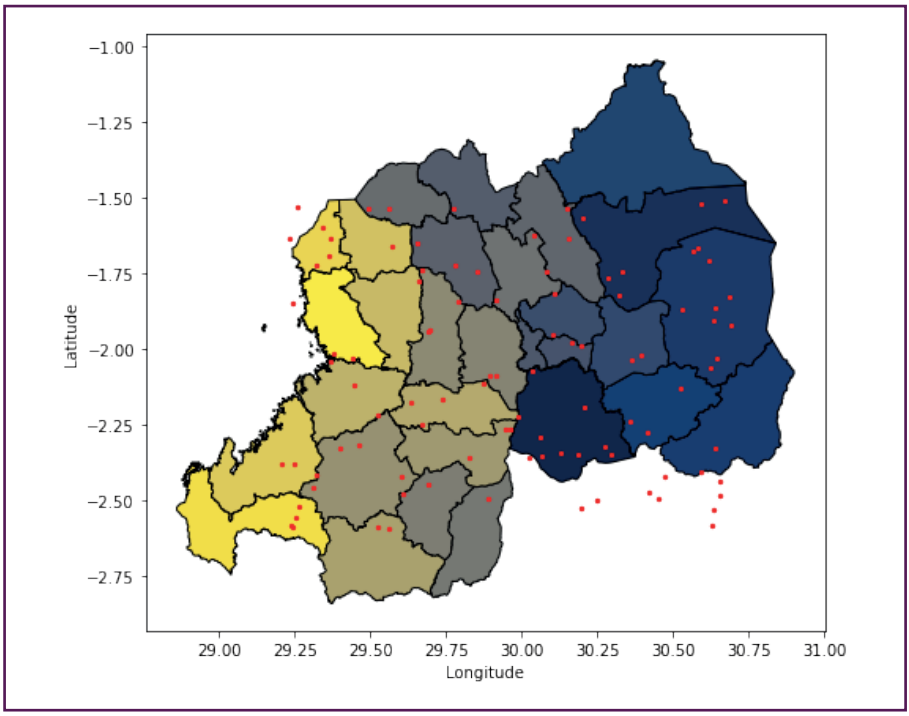
Fonte: Wang *et al.* (2022).

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).

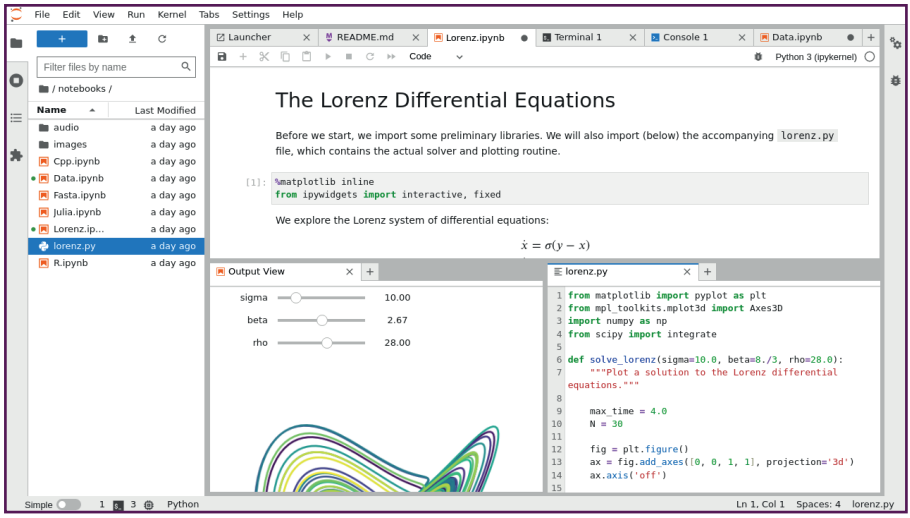
37 Disponível em: https://pypi.org/project/nlpnet/. Acesso em: 21 set. 2023.

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.