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

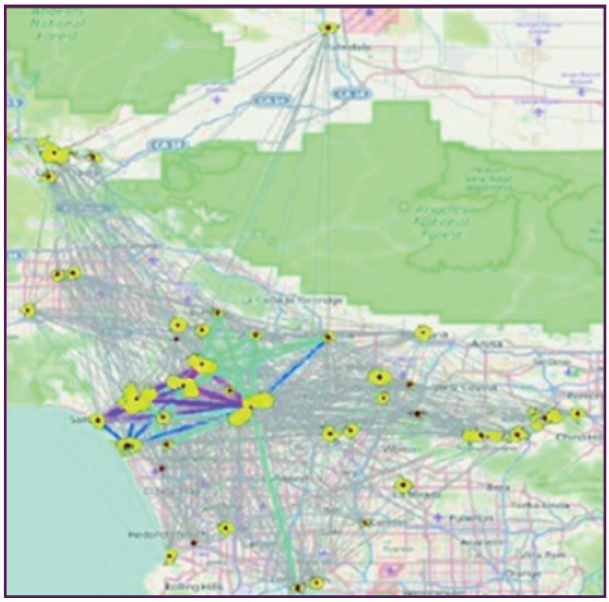
estudos, pesquisas, construção de modelos e análises para entender padrões de comportamento, opiniões públicas e dinâmicas sociais em escala global. Dessa forma são apresentados, sem a pretensão de serem descritas de forma exaustiva, trabalhos de pesquisadores brasileiros e internacionais envolvendo 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.

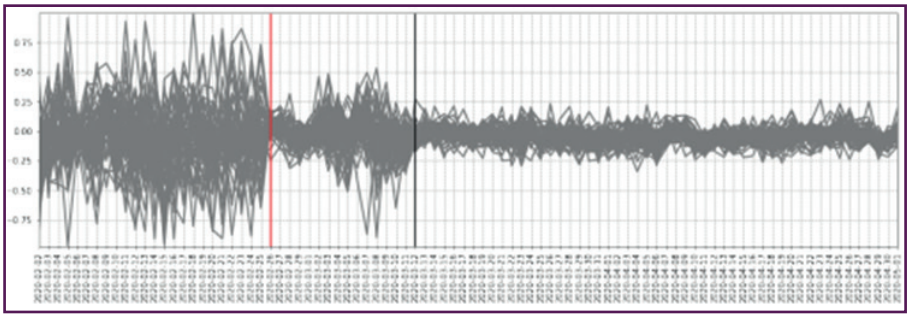
47 *Site* de *X*: https://twitter.com/X?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor. Acesso em: 18 set. 2023.

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.