WebScraping com R: aplicação no site Stackoverflow

Ana Beatriz Ramos da Silva* Jodavid Ferreira†

14 de setembro de 2025

Resumo

Este trabalho explora a técnica de Web Scraping no R com suporte de pacotes como *rvest, ggplt2, dplyr, stringr, reshape2, ggcorplot, tidytext* que foram aplicados para funções como extração, limpeza, organização e visualização dos dados. O site escolhido para demostrar a técnica foi o Stack Overflow, um site usado especialmente por profissionais da tecnologia como uma comunidade de suporte interno. A aplicação foi satisfatória quando a eficácia da técnica e trouxe resultados referentes a popularidade das perguntas no site e que existe uma associação entre o número de respostas e o número de visualizações para cada pergunta, por fim, recomenda-se utilizar a técnica para extração de informação das mais variadas fontes desde que respeitando as recomendações legais.

1 Introdução

Web Scraping ou raspagem de dados é o nome dado ao processo de extrair informações de uma página da internet seja esta qual for, dentre os principais tipos de raspagem feita com esta técnica está a comparação de preços e avaliação de conteúdos, ou validação de popularidade de assuntos específicos. Os conteúdos coletados por esta técnica são os mais variados desde, imagens até textos.

2 Metodologia

Para realização da coleta automatizada de dados em sites na internet chamada Web Scraping alguns passos podem ser seguidos, sendo estes os possíveis passos:

1. Seleção dos dados que serão coletados, neste trabalho serão utilizados dados do site Stack Overflow disponível no https://stackoverflow.com/questions?sort=votes, este site é uma plataforma online de perguntas e repostas direcionadas para profissionais da tecnologia como programadores, dentro deste ambiente dúvidas e questionamentos que possam ajudar programadores a resolver problemas encontrados no caminho são levantados afim de que outros possam oferecer soluções.

*UFPE, anaramos.silva@ufpe.br
†UFPE, jodavid.ferreira@ufpe.br

2. Avaliação do tipo de estrutura da página, ou seja, a depender do tipo de conteúdo o tipo de ferramenta para extração varia, ou seja, se trabalha-se com textos, tabelas, links ou imagens.

3. Extração dos dados com R (este passo pode ser realizado com outra linguagem como Python ou Java). Neste ambiente alguns pacotes serão usados, o pacote direcionado a técnica de Web Scraping será o *rvest*

4. Limpeza e organização dos dados, na maioria dos casos os dados não são extraídos de forma a estarem prontos para serem analisados, dessa forma outras técnicas de análise de dados precisam ser empregadas para estrurura seja lida adequadamente como por exemplo, texto, variável númerica, data e etc.Além da parte técnica, esta coleta deve respeitar a lei e os termos de uso dos sites de extração.

3 Resultados

Toda a execução desta ténica foi relizada no Google Colab, utilizando a interface do software R. Pacotes como, *rvest, dplyr, ggplot2, tidytext, stringr, reshape2, ggcorplot.*

Inicialmente, a url foi inserida e os dados do site foram coletados, dentro deste site foram encontrados links, títulos (que se referem as perguntas feitas pelos internautas e variáveis númericas que correspodem ao número de vizualizações, respostas e score de cada pergunta. Algumas manipulações foram realizadas e uma matriz de dados foi montada. Cada variável possui 15 observações, ou seje,a 15 informações para cada uma delas.

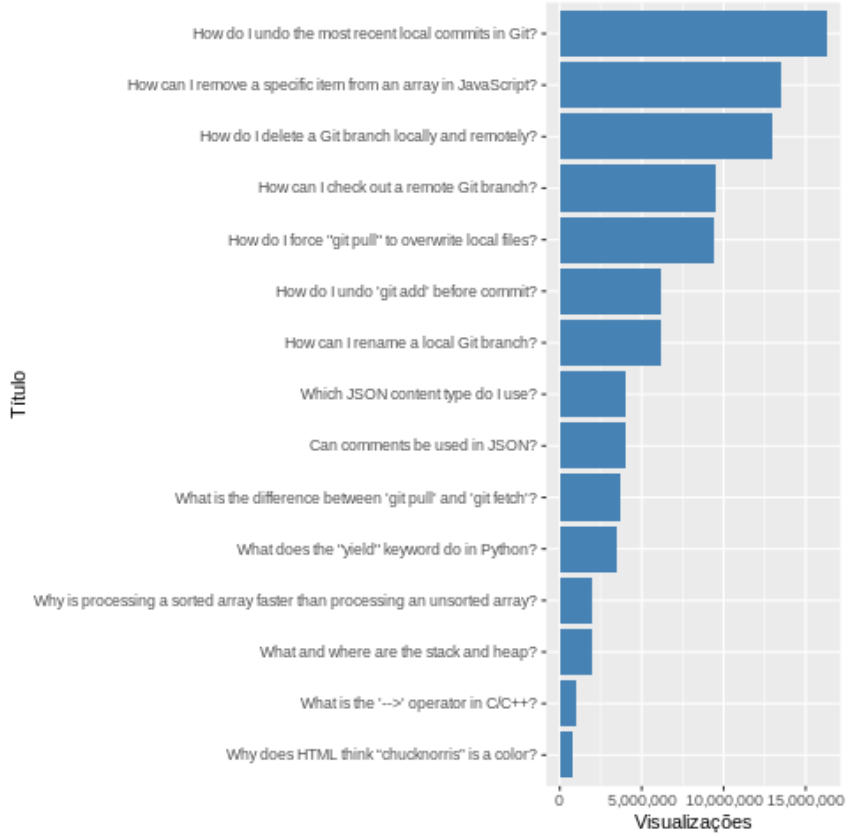
Para melhor visualização dos dados gráficos foram gerados, os quais serão exibidos a seguir.

A pergunta que mais obteve vizaulizações foi "How do I undo the most recent local commits in Git?"em português significa "Como desfaço os commits locais mais recentes no Git?". A segunda investigação se refere ao número de respostas para cada pergunta.

Novamente a pergunta que lidera o ranking é a referente aos commits no Git. Nem todas as perguntas seguem o mesmo lugar nos dois rankings.

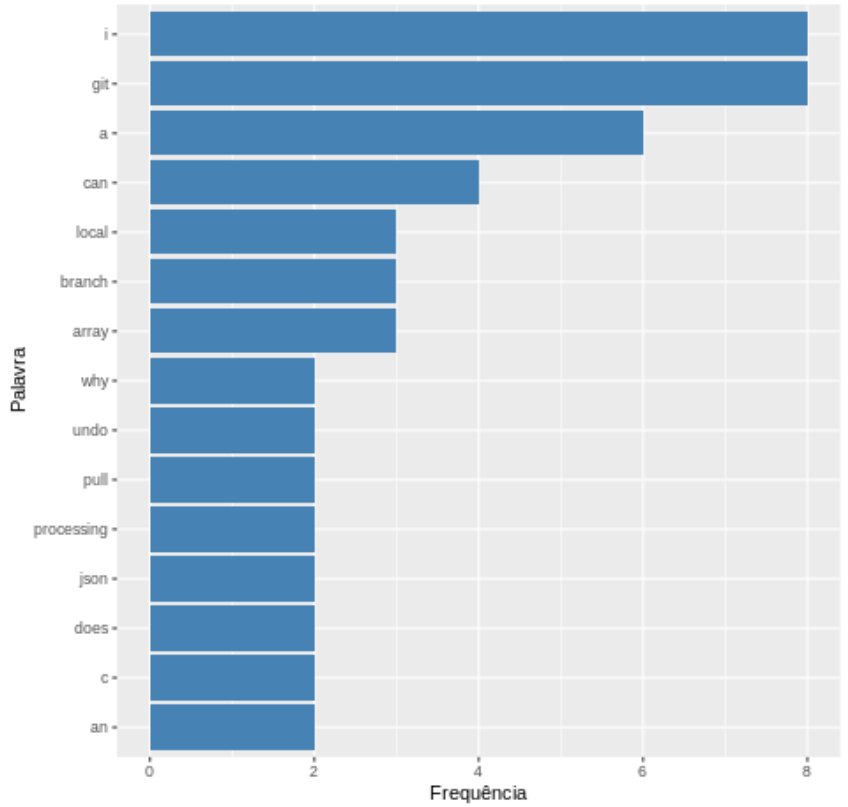
Afim de continuar com a análise exploratória de dados, os textos das perguntas foram separados em palavras individuais e uma ordenação para avaliação das palavras que mais aparecem foi realizada. Para melhor visualização as stepwords mais comuns como "the", "is", "do", "how", "what", "and", "in", "of", "to"foram retiradas.

Figura 1: Comparação do título da pergunta e o número de visualizações.



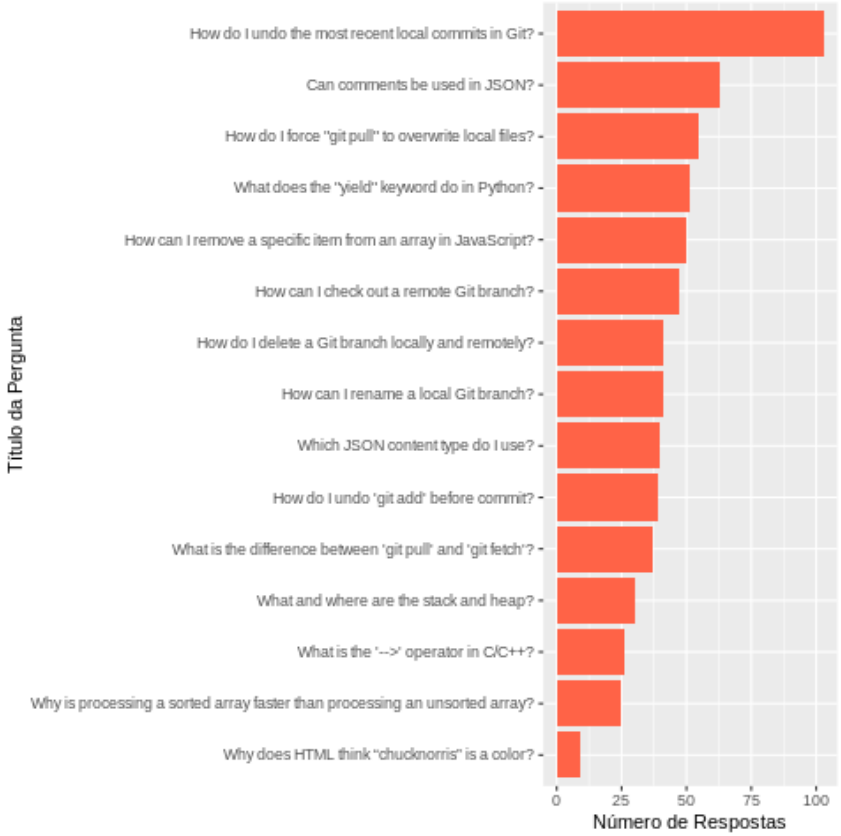
Fonte: Elaborada pela autora, (2025).

Figura 3: Frequência das palavras que mais aparecem.



Fonte: Elaborada pela autora, (2025).

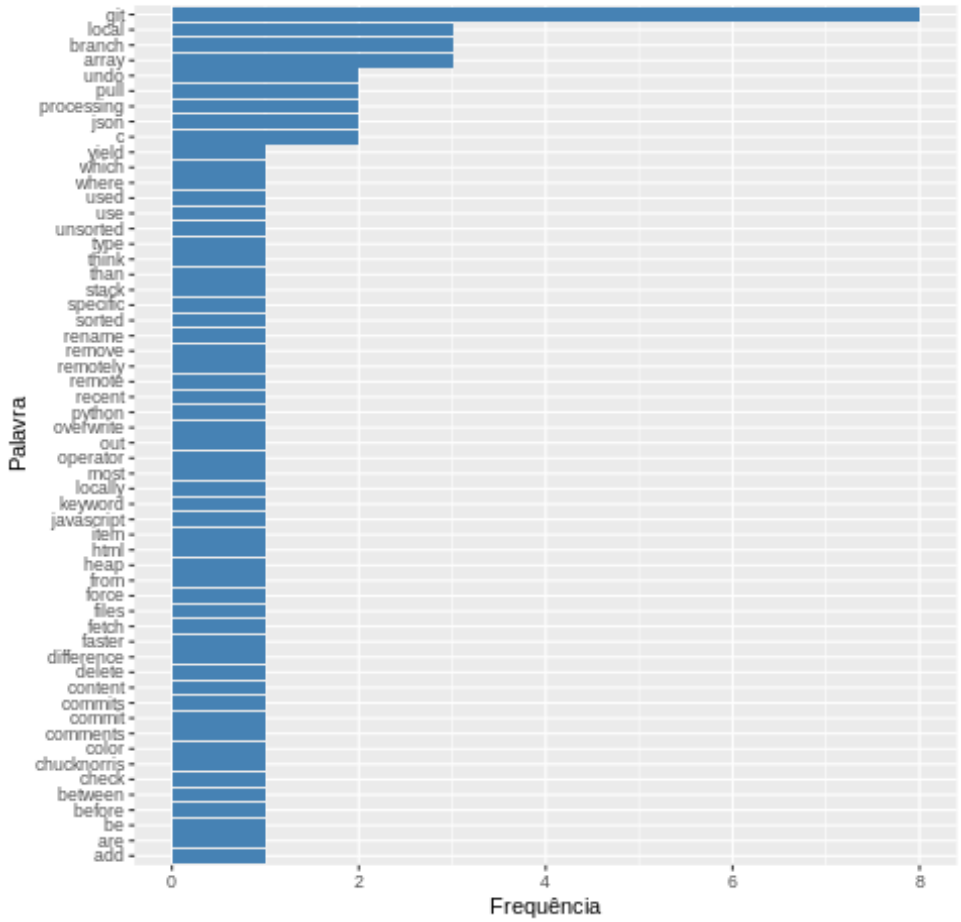
Figura 2: Comparação do título da pergunta e o número de respostas.



Fonte: Elaborada pela autora, (2025).

Mesmo após a remoção das stepwords algumas palavras são genéricas para qualquer tipo de texto e não refletem o conteúdo buscado, elas também serão removidas para que possa ser visto as palavras mais relevantes para conteúdo.

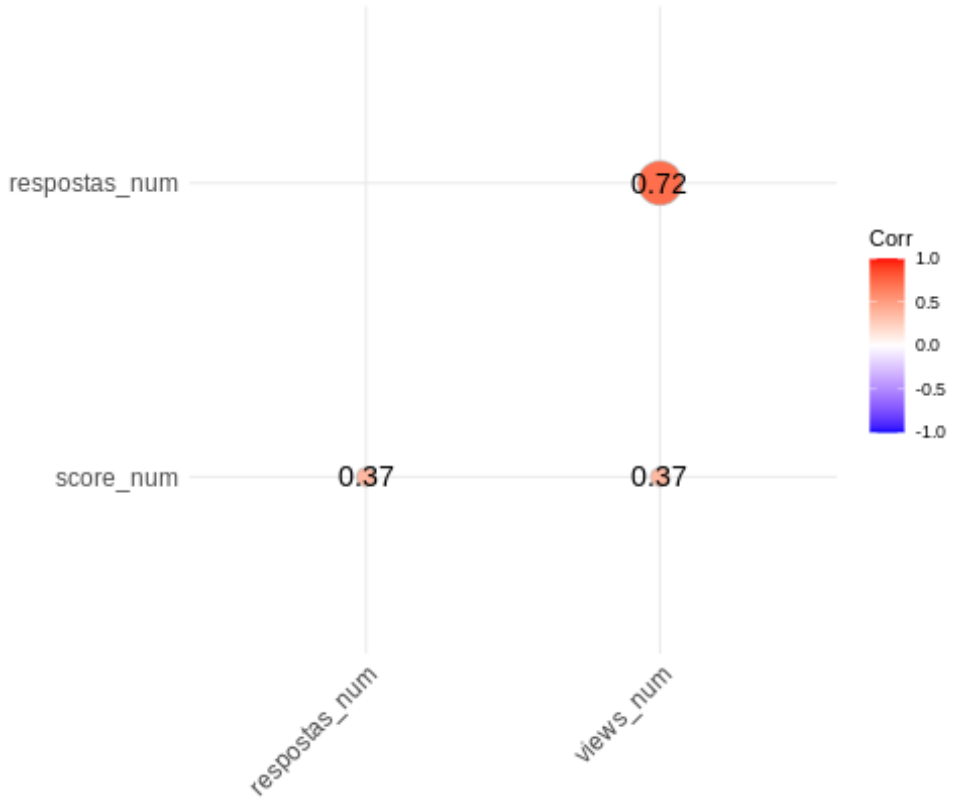
Figura 4: Frequência das palavras que mais aparecem com remoção de palavras comuns.



Fonte: Elaborada pela autora, (2025).

Por fim, para a alivação da correlação entre as variáveis númericas, elas foram transformadas para garantir que o formato seja lido corretamente e uma matriz de correlação possa ser construida.

Figura 5: Frequência das palavras que mais aparecem com remoção de palavras comuns.



Fonte: Elaborada pela autora, (2025).

A maior correlação encontrada foi entre o número de visualizações e o número de respostas para cada pergunta. Como esperado, correlações negativas.

4 Discussão

Observando os resultados obtidos pela técnica de Web Scraping no site Stack Overflow mostram uma pequena parte do potencial desta técnica para coleta de dados e posteriormente uma análise. A análise envidência que dentro da plataforma questionamentos em relação ao Git são mais explorados e alcançam uma comunidade maior. Com a matriz de correlação foi observado uma associação positiva entre o número de visualização e de respostas para cada pergunta, oque é de se esperar intuitivamente. Vale salientar que o tipo de raspagem varia de acordo com o formato dos dados em cada site e o interesse do utilizador da ferramenta.

5 Conclusão

A técninca de Web Scraping se mostrou eficaz para coleta automatizada de dados para diversos setores que buscam extrair informações da web. O uso da linguagem R, aliado

a pacotes específicos como *rvest* e *ggplot2*, mostrou-se adequado para realizar tanto a extração quanto o tratamento e visualização das informações.

Como perspectivas futuras, outras bases podem ser usadas, incluindo períodos mais longos e outras plataformas de discussão para que diferentes públicos possam ser atingidos, além de aplicar técnicas de aprendizado de máquina para classificação e previsão de tendências. Dessa forma, o Web Scraping pode ser considerado como um recurso valioso para investigações acadêmicas e empresariais considerando a execução em R ou outra linguagem de programação, desde que conduzido com responsabilidade ética e respeito às normas legais.

Referências

[1] https://blog.casadodesenvolvedor.com.br/webscraping/

[2] Bradley, Alex e Richard JE James. "Web scraping usando R."Avanços em Métodos e Práticas em Ciências Psicológicas 2.3 (2019): 264-270.

[3] MUNZERT, Simon et al. Coleta automatizada de dados com R: Um guia prático para web scraping e mineração de texto . West Sussex: Wiley, 2015.

[4] KHALIL, Salim; FAKIR, Mohamed. RCrawler: Um pacote R para rastreamento e raspagem paralelos na web. SoftwareX , v. 6, p. 98-106, 2017.

[5] https://www.r-project.org/