Visualização textual automatizada com nuvem de palavras: uma análise da página da wikipedia sobre inteligência artificial

Vitor Amadeu Souza[1]; 0009-00-02-1857-6799

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
vitor.amadeu@foa.org.br

Resumo: Este estudo apresenta uma análise de conteúdo web utilizando técnicas de visualização textual através de nuvem de palavras, aplicada especificamente à página da Wikipedia sobre inteligência artificial. O objetivo principal foi desenvolver e implementar um sistema automatizado para extração, processamento e visualização de conteúdo textual web, demonstrando a eficácia das nuvens de palavras como ferramenta de análise exploratória de dados textuais. A metodologia empregou web scraping com Python utilizando as bibliotecas requests e BeautifulSoup para extração de conteúdo, NLTK para processamento de linguagem natural incluindo remoção de stopwords, e WordCloud para geração da visualização final. Os resultados obtidos através da análise da página revelaram a predominância de termos técnicos e conceituais relacionados ao domínio da inteligência artificial. A nuvem de palavras gerada evidenciou "inteligência", "artificial", "sistema", "máquina", "código", "humano", "ciência", "computador" e "pesquisa" como os termos de maior frequência, refletindo adequadamente o conteúdo temático da fonte analisada. O estudo demonstra que a visualização através de nuvens de palavras constitui uma ferramenta eficaz para análise exploratória rápida de conteúdo web, permitindo identificação imediata dos principais tópicos e conceitos abordados em textos extensos.

Palavras-chave: Nuvem de palavras. Análise de conteúdo web. Visualização textual. Web scraping. Processamento de linguagem natural. Wikipedia.

INTRODUÇÃO

A crescente disponibilidade de conteúdo textual na internet demanda o desenvolvimento de técnicas eficientes para análise e compreensão automatizada de informações (Manning; Raghavan; Schütze, 2008). Neste contexto, as nuvens de palavras emergem como uma ferramenta visual poderosa para representação de frequência textual, permitindo identificação rápida dos termos mais relevantes em um corpus de dados (Viégas; Wattenberg; Feinberg, 2009).

O processamento automatizado de conteúdo web, conhecido como web scraping, tornou-se uma técnica fundamental para coleta de dados textuais em larga escala (Mitchell, 2018). Esta abordagem permite a extração sistemática de informações de páginas web, possibilitando análises subsequentes através de técnicas de processamento de linguagem natural (Bird; Klein; Loper, 2009).

A Wikipedia, como uma das maiores enciclopédias colaborativas online, representa uma fonte rica de conhecimento estruturado e representa um objeto de estudo ideal para análise de conteúdo textual (Giles, 2005). Especificamente, a página sobre inteligência artificial na Wikipedia portuguesa constitui um corpus textual abrangente que engloba conceitos fundamentais, histórico, aplicações e desenvolvimentos recentes na área.

As nuvens de palavras, introduzidas por Viégas, Wattenberg e Feinberg (2009), oferecem uma representação visual intuitiva da frequência de termos em um texto, onde o tamanho das palavras é proporcional à sua frequência de ocorrência. Esta técnica de visualização tem sido amplamente adotada em análises exploratórias de dados textuais devido à sua capacidade de fornecer insights imediatos sobre os principais tópicos abordados em um documento (Heimerl *et al*., 2014).

O objetivo deste estudo é demonstrar a aplicação prática de técnicas de web scraping e visualização textual através da análise da página da Wikipedia sobre inteligência artificial, evidenciando como a combinação dessas tecnologias pode fornecer insights valiosos sobre o conteúdo de páginas web de forma automatizada e eficiente.

MÉTODOS

O sistema desenvolvido foi estruturado em três módulos principais: extração de dados, processamento textual e visualização. A implementação utilizou a linguagem Python devido à sua robustez em processamento de dados textuais e disponibilidade de bibliotecas especializadas (Lutz, 2013). Para a extração de conteúdo web, utilizou-se a biblioteca requests para realizar requisições HTTP e BeautifulSoup4 para parsing do HTML resultante (Richardson, 2007). A função extrair_texto_da_url() implementada no código realiza uma requisição GET para a URL especificada e extrai todo o texto visível da página, removendo elementos HTML e mantendo apenas o conteúdo textual. O processo de extração segue as melhores práticas de web scraping, incluindo tratamento de exceções para garantir robustez em caso de falhas de conexão ou páginas inacessíveis (Mitchell, 2018). A validação do status HTTP através do método raise_for_status() assegura que apenas requisições bem-sucedidas sejam processadas.

O processamento textual foi implementado através da função limpar_texto(), que executa uma sequência de operações de normalização textual baseadas em técnicas estabelecidas de processamento de linguagem natural (Manning; Raghavan; Schütze, 2008). As etapas de processamento incluem conversão para minúsculas para padronização, remoção de pontuação utilizando a biblioteca string do Python, tokenização por espaços, filtragem de stopwords em português e inglês utilizando o corpus do NLTK (Bird; Klein; Loper, 2009), e filtragem de palavras com menos de três caracteres para reduzir ruído. A utilização do Natural Language Toolkit (NLTK) para remoção de stopwords é fundamentada na necessidade de eliminar palavras funcionais que não contribuem significativamente para o conteúdo semântico do texto (Wilbur; Sirotkin, 1992).

A visualização foi implementada através da biblioteca WordCloud, que gera representações visuais onde o tamanho de cada palavra é proporcional à sua frequência no texto (Mueller, 2018). A função gerar_nuvem_de_palavras() configura parâmetros específicos incluindo dimensões (800x400 pixels), cor de fundo branca e interpolação bilinear para suavização visual. A integração com matplotlib permite a exibição da nuvem de palavras com formatação adequada, incluindo remoção de eixos coordenados e adição de título descritivo (Hunter, 2007). O estudo utilizou como fonte de dados a página da Wikipedia em português sobre

inteligência artificial (https://pt.wikipedia.org/wiki/Intelig%C3%AAncia_artificial), selecionada por representar um conteúdo técnico abrangente e constantemente atualizado sobre o domínio temático de interesse.

O código-fonte está disponível para download através do link: https://github.com/vitor-souza-ime/np.

RESULTADOS E DISCUSSÃO

A execução do código resultou na extração do conteúdo textual da página da Wikipedia sobre inteligência artificial. O processo de extração capturou 10.726 palavras do conteúdo principal da página, incluindo definições, histórico, aplicações, métodos e referências relacionadas ao tema. O processamento de linguagem natural aplicado ao texto extraído resultou na identificação de 3.043 palavras únicas após a remoção de stopwords e filtragem de termos com menos de três caracteres. Esta redução significativa demonstra a eficácia das técnicas de normalização textual em concentrar a análise nos termos semanticamente relevantes (Manning; Raghavan; Schütze, 2008).

A nuvem de palavras gerada apresenta uma distribuição visual clara dos termos mais frequentes no conteúdo analisado. Os resultados evidenciam a predominância de conceitos fundamentais relacionados à inteligência artificial, confirmando a relevância temática da fonte de dados selecionada. A Figura 1 representa a visualização obtida através da execução do código, demonstrando a distribuição de frequência dos termos extraídos da página informada.

A análise visual da nuvem de palavras revela que os termos com maior destaque incluem "inteligência", "artificial", "sistema", "máquina", "código", "humano", "ciência", "computador", "pesquisa" e "fonte". Esta distribuição reflete adequadamente o conteúdo da página analisada, evidenciando os conceitos centrais abordados no artigo da Wikipedia. O termo "inteligência" aparece com maior proeminência, seguido por "artificial", o que é consistente com o tema principal da página. A presença significativa de "sistema", "máquina" e "código" indica a ênfase dada aos aspectos técnicos e computacionais da inteligência artificial na fonte analisada (Russell; Norvig, 2016). A aparição de termos como "humano", "ciência" e "pesquisa" sugere que o conteúdo aborda não apenas os aspectos técnicos, mas também

as implicações sociais, científicas e de pesquisa relacionadas à inteligência artificial, demonstrando a abrangência do artigo da Wikipedia como fonte de informação (Nilsson, 2009).

Figura 1 - Nuvem de palavras



Fonte: O autor.

A validação dos resultados foi realizada através da comparação manual dos termos identificados na nuvem de palavras com o conteúdo original da página da Wikipedia. Esta análise confirmou que os termos de maior frequência correspondem efetivamente aos conceitos centrais abordados no texto, incluindo definições de inteligência artificial, descrições de sistemas computacionais, discussões sobre capacidades humanas versus artificiais, e referências a pesquisas e desenvolvimentos científicos na área. A presença de termos técnicos específicos como "algoritmo", "robótica", "máquina" e "computacional" na visualização confirma que o processamento capturou adequadamente o vocabulário especializado característico do domínio da inteligência artificial (Goodfellow; Bengio; Courville, 2016).

Embora a nuvem de palavras forneça uma visão geral útil do conteúdo, é importante reconhecer suas limitações. A técnica não preserva relações contextuais entre termos nem hierarquias conceituais presentes no texto original (Heimerl *et al.*, 2014). Além disso, a frequência de um termo não necessariamente reflete sua importância semântica no contexto do documento. A abordagem de remoção de stopwords, embora eficaz na redução de ruído, pode ocasionalmente remover termos que possuem significado específico no contexto da inteligência artificial. Futuras implementações poderiam beneficiar-se de listas de stopwords customizadas para o domínio específico.

CONCLUSÕES

Este estudo demonstrou a eficácia da combinação de técnicas de web scraping, processamento de linguagem natural e visualização através de nuvens de palavras para análise automatizada de conteúdo web. A implementação desenvolvida conseguiu extrair, processar e visualizar com sucesso o conteúdo da página da Wikipedia sobre inteligência artificial, gerando insights valiosos sobre os principais tópicos abordados.

Os resultados obtidos confirmam que a nuvem de palavras constitui uma ferramenta eficaz para análise exploratória rápida de conteúdo textual, permitindo identificação imediata dos conceitos mais relevantes em um documento. A visualização gerada refletiu adequadamente o conteúdo temático da fonte analisada, evidenciando termos centrais como "inteligência", "artificial", "sistema", "máquina" e "código".

A metodologia proposta apresenta aplicabilidade ampla para análise de diferentes tipos de conteúdo web, podendo ser adaptada para diversos domínios temáticos através de ajustes nos parâmetros de processamento e filtragem. A integração de bibliotecas Python especializadas demonstrou ser uma abordagem robusta e eficiente para implementação de sistemas de análise textual automatizada.

Para trabalhos futuros, recomenda-se a exploração de técnicas mais avançadas de processamento de linguagem natural, incluindo análise de sentimentos, identificação de entidades nomeadas e modelagem de tópicos, que poderiam complementar a análise visual fornecida pelas nuvens de palavras. Adicionalmente, a implementação de interfaces interativas permitiria análises mais dinâmicas e personalizadas do conteúdo extraído.

A contribuição deste estudo reside na demonstração de como técnicas de processamento de dados podem fornecer insights significativos sobre conteúdo web, oferecendo uma base sólida para desenvolvimento de sistemas mais sofisticados de análise de informação digital.

REFERÊNCIAS

BIRD, S.; KLEIN, E.; LOPER, E. Natural language processing with Python: analyzing text with the natural language toolkit. Beijing: O'Reilly, 2009.

GILES, J. Internet encyclopaedias go head to head. Nature, v. 438, n. 7070, p. 900-901, 2005.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning. Cambridge: MIT Press, 2016.

HEIMERL, F. et al. Word cloud explorer: text analytics based on word clouds. In: 47th Hawaii International Conference on System Sciences. Proceedings... IEEE, 2014. p. 1833-1842.

HUNTER, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007.

LUTZ, M. Learning Python. 5th ed. Beijing: O'Reilly Media, 2013.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008.

MITCHELL, R. Web scraping with Python: collecting more data from the modern web. 2nd ed. Beijing: O'Reilly Media, 2018.

MUELLER, A. WordCloud for Python documentation. Disponível em: https://amueller.github.io/word_cloud/. Acesso em: 31 jul. 2025.

NILSSON, N. J. The quest for artificial intelligence: a history of ideas and achievements. Cambridge: Cambridge University Press, 2009.

RICHARDSON, L. Beautiful Soup documentation. Disponível em: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Acesso em: 31 jul. 2025.

RUSSELL, S. J.; NORVIG, P. Artificial intelligence: a modern approach. 3rd ed. Boston: Pearson, 2016.

VIÉGAS, F. B.; WATTENBERG, M.; FEINBERG, J. Participatory visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics, v. 15, n. 6, p. 1137-1144, 2009.

WILBUR, W. J.; SIROTKIN, K. The automatic identification of stop words. Journal of Information Science, v. 18, n. 1, p. 45-55, 1992.