Sumarização Automática e Geração de Nuvens de Palavras:

desenvolvimento de um software para otimização do acesso a informações em textos e mídias

Vitor Santos Lopes[1]
Maria Luisa Cervi Uzun[2]

RESUMO
O presente artigo descreve o desenvolvimento de um software destinado à geração automática de resumos e nuvens de palavras a partir de textos e outras mídias de leitura. Detalha-se a estrutura do projeto, as ferramentas empregadas na criação da aplicação e o fluxo operacional do software, concebido para usuários que carecem de tempo para a leitura integral de artigos ou livros, focando apenas nos elementos essenciais do conteúdo. A metodologia adotada envolve uma abordagem qualitativa e aplicada, utilizando-se de uma pesquisa descritiva e explicativa baseada em levantamento bibliográfico. O processamento de linguagem natural, com ênfase na técnica de sumarização de textos, foi a principal ferramenta empregada, permitindo a transformação de linguagem natural em dados manipuláveis por máquinas. Conclui-se que o software fornece ao público-alvo uma capacidade aprimorada de abstração de informações, isolando apenas os dados mais relevantes.

Palavras-chave: Inteligência Artificial. Nuvem de palavras. Processamento de Linguagem Natural. Resumo.

ABSTRACT
This paper presents the development of software designed to automatically generate summaries and word clouds from texts and other reading materials. It outlines the project's structure, the tools used in the application's development, and the software's operational workflow, targeting individuals lacking time to fully read articles or books, highlighting only the essential content. The methodology employed is qualitative and applied, incorporating a descriptive and explanatory research conducted through a bibliographic survey. Natural language processing, particularly text summarization techniques, served as the key tool, enabling the conversion of natural language into machine-manipulable information. The software ultimately enhances the target audience's ability to abstract content, filtering only the most pertinent information.

Keywords: *Artificial Intelligence. Word Cloud. Natural Language Processing. Summary.*

[1] Graduado em Análise e Desenvolvimento de Sistemas pela Fatec Dr Thomaz Novelino – Franca/SP. Endereço eletrônico: vitorsantoslopes1@gmail.com
[2] Professora Associada da Fatec Dr Thomaz Novelino – Franca/SP. Endereço eletrônico: maria.uzun@fatec.sp.gov.br.

1 INTRODUÇÃO

Em uma era dominada pela produção incansável de dados através de aplicativos móveis, sistemas, Internet das Coisas (IoT), televisores e outros dispositivos, a necessidade de capturar e processar essas informações para fins de tomada de decisão tornou-se importante. A maior parte desses dados é não estruturada, incluindo imagens, vídeos, posts em redes sociais, comentários e áudios, o que exigiu o desenvolvimento de algoritmos capazes de convertê-los em informações úteis.

Desde o início dos anos 2000, a área de inteligência artificial (IA) tem crescido exponencialmente, impulsionada pela evolução tecnológica e pelo aumento na disponibilidade de dados, decorrente da expansão do acesso à internet. Conforme Gomes (2010), a IA tem sido delineada por quatro principais linhas de pensamento, desde sistemas que pensam e atuam como humanos até aqueles que operam de maneira racional.

O Processamento de Linguagem Natural (PLN) se destaca como uma subárea da IA, dedicada a fazer com que computadores entendam a linguagem humana. Embora não seja uma ciência nova, o PLN avançou significativamente com a expansão do *Big Data*. Esta tecnologia abrange desde a compreensão de textos, incluindo análises sintática, semântica, léxica e morfológica, até a sumarização e interpretação de sentimentos.

Então, surge a proposta deste artigo: utilizar o PLN para facilitar a leitura e compreensão de textos por meio da sumarização automática. Este trabalho apresenta o desenvolvimento de um *software* capaz de gerar resumos e nuvens de palavras de textos e outras fontes de leitura, visando auxiliar estudantes, professores e profissionais a captar rapidamente os pontos principais dos documentos. A metodologia adotada foi qualitativa e aplicada, baseada em pesquisa descritiva e explicativa através de levantamento bibliográfico. A principal ferramenta utilizada foi o processamento de linguagem natural com técnica de sumarização de textos, transformando a linguagem natural em informações prontamente manipuláveis por máquinas.

2 REFERENCIAL TEÓRICO

2.1 Inteligência Artificial e Grandes Volumes de Dados

A Inteligência Artificial (IA) tem suas origens nos estudos de como os seres humanos pensam e processam informações, tentando replicar esses processos em máquinas (De Alencar, et al., 2024).

Durante 2023 foi alucinante o avanço da IA, como por exemplo, IAs generativas. E a notícia é que não vai desacelerar, pois a disponibilidade de grandes volumes de dados e o avanço das capacidades computacionais estão permitindo um crescimento exponencial na aplicação da IA.

O *Big Data* transformou a maneira como analisamos e interpretamos informações, proporcionando um terreno fértil para o desenvolvimento de sistemas de IA mais sofisticados e precisos.

À medida que dispositivos e sensores conectados à Internet aumentam, enormes quantidades de dados são geradas, juntando informações variadas como registros de atividades de usuário, transações financeiras, sinais de sensores, textos, imagens e vídeos. A capacidade de processar e analisar esses vastos conjuntos de dados usando algoritmos de IA tem revolucionado campos como a medicina, finanças, e educação, fornecendo insights profundos e permitindo decisões mais informadas e automatizadas.

A inteligência artificial utiliza esses dados para aprender padrões e tomar decisões. Modelos de aprendizado de máquina e redes neurais, por exemplo, requerem grandes conjuntos de dados para treinar e refinar suas capacidades de previsão e análise. Mayer-Schönberger e Cukier (2013) destacam como o *Big Data* não só oferece a matéria-prima para o treinamento desses modelos, mas também desafia os limites do que a IA pode alcançar, através da descoberta de correlações e tendências que seriam impossíveis de detectar com conjuntos menores de dados.

A Inteligência Artificial é poderosa com uma variedade de aplicações, oferecendo oportunidades significativas para impulsionar o crescimento econômico e social.

Alguns marcos importantes relacionados à inteligência artificial.

Quadro 1: Marcos Importantes IA

Teste de Turing	Redes Neurais Artificiais (RNAs)	Aprendizado Profundo (Deep Learning)
• Proposto pelo matemático e cientista da computação **Alan Turing** em 1950. • O teste avalia a capacidade de uma máquina de exibir comportamento inteligente indistinguível do de um ser humano. • Se uma máquina puder convencer um juiz humano de que é um humano durante uma conversa, ela passa no teste.	• Inspiradas no funcionamento do cérebro humano. • Consistem em camadas de neurônios interconectados. • **Perceptron**, desenvolvido por **Frank Rosenblatt** em 1957, foi um dos primeiros modelos de RNA. • Avanços recentes incluem redes profundas (deep learning) com muitas camadas, como **redes neurais convolucionais (CNNs)** e **redes neurais recorrentes (RNNs)**.	• Subcampo da IA que se concentra em redes neurais profundas. • **Backpropagation** (retropropagação) é um algoritmo fundamental para treinar redes profundas. • **Arquiteturas populares**: • **CNNs**: Usadas para visão computacional (reconhecimento de imagens). • **RNNs**: Úteis para processamento de sequências (como em tradução automática ou análise de texto). • **Redes Generativas Adversariais (GANs)**: Geram dados realistas, como imagens de rostos sintéticos.

Fonte: o autor (2024)

Esses marcos são essenciais para entender o desenvolvimento da IA e como ela evoluiu ao longo do tempo. O teste de Turing estabeleceu um padrão para avaliar a inteligência das máquinas, enquanto as redes neurais artificiais e o aprendizado profundo revolucionaram a capacidade de resolver problemas complexos

2.2 Processamento de Linguagem Natural (PLN)

O PLN é uma subárea da IA que foca em dar às máquinas a habilidade de entender textos escritos e falados em linguagem natural. Jurafsky e Martin (2019) explicam que o PLN combina computação e linguística para entender e manipular a linguagem humana, abordando desde a análise sintática e semântica até a interpretação contextual e a geração de respostas.

É um domínio da IA dedicado a entender e manipular a linguagem humana de forma que as máquinas possam realizar tarefas como tradução automática, geração de respostas e sumarização de textos. Segundo Jurafsky e Martin (2019), o PLN combina técnicas de computação e teorias linguísticas para desvendar a complexidade da linguagem, abordando tanto a compreensão quanto a produção de texto e fala.

Esta área enfrenta desafios únicos, como a ambiguidade da linguagem, a variedade de contextos e a necessidade de entender nuances e sentimentos. O PLN utiliza algoritmos que analisam a estrutura e o conteúdo do texto, aplicando métodos de aprendizado de máquina para interpretar a semântica e a sintaxe. Isso permite que sistemas automatizados realizem tarefas complexas como responder perguntas, interagir em linguagem natural e extrair informações relevantes de grandes volumes de texto.

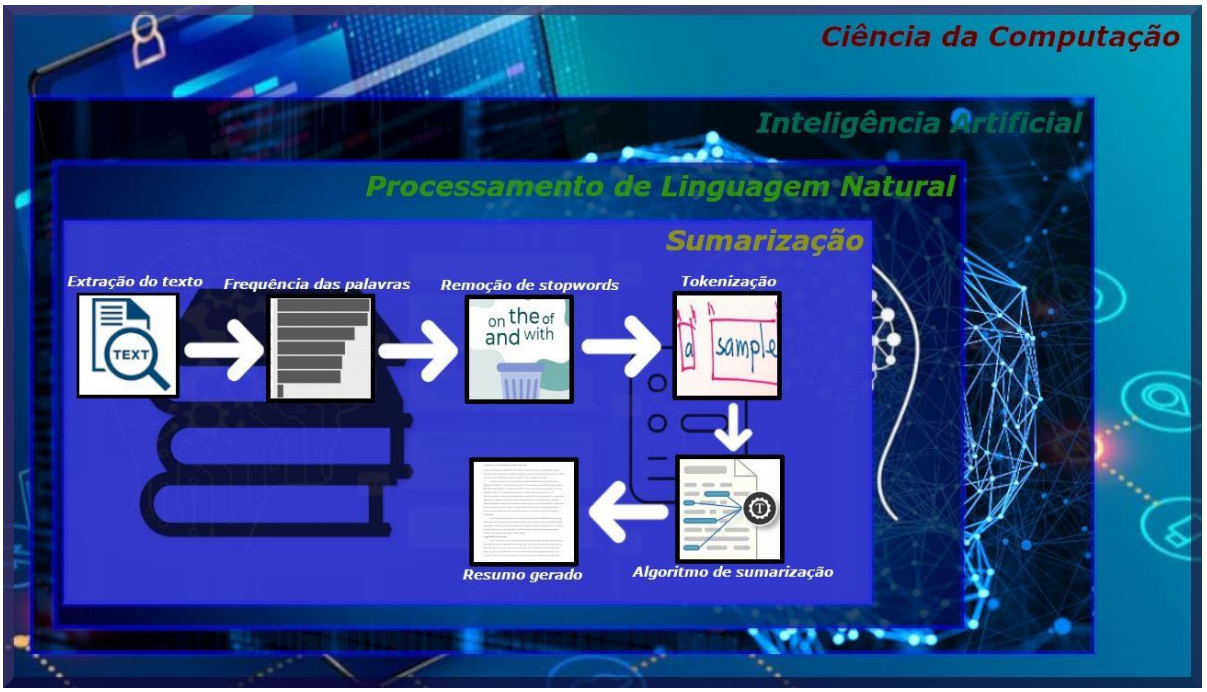
2.3 Técnicas de Sumarização de Textos e Remoção de *Stopwords* e *Tokenização.*

A sumarização automática de textos é uma das aplicações práticas do PLN. Liddy (2001) afirma que a sumarização pode ser realizada através de métodos como a extração de palavras-chave, onde a frequência e a relevância das palavras dentro de um contexto são calculadas para identificar as sentenças mais informativas. Mani e Maybury (1999) complementam que técnicas avançadas de sumarização consideram não apenas a frequência das palavras, mas também a sua posição estrutural no texto, a relevância temática e a inter-relação com outros segmentos do texto.

A Figura 1 ilustra as etapas do algoritmo de sumarização, desde a extração do texto até a seleção e apresentação das sentenças mais relevantes.

ISSN: 2595-5942

Figura 1: Etapas de sumarização



Fonte: o autor (2024)

A sumarização é uma técnica aplicada no processamento de linguagem natural objetiva identificar e extrair os elementos mais relevantes de um texto. O processo inicia-se com a extração do texto alvo, que pode ser um artigo, livro ou notícia. Segue-se a análise da frequência das palavras no texto, atribuindo-se pesos a estas para indicar sua importância relativa. A etapa subsequente envolve a remoção de *stopwords*, ou seja, palavras que, apesar de frequentes, possuem pouco valor informativo para a análise (Silge e Robinson, 2016).

Após esta filtragem, procede-se à *tokenização*, que envolve segmentar o texto em unidades menores como palavras ou frases, é crucial para a análise estrutural subsequente (Bird, Klein, & Loper, 2009). Finalmente, um algoritmo é empregado para determinar quais sentenças são mais significativas, culminando na geração do resumo que destaca os pontos cruciais do texto. Esta abordagem sistemática permite sintetizar o conteúdo de maneira eficiente e relevante.

Existem várias abordagens para realizar a sumarização de textos, cada uma com suas vantagens e desvantagens, dependendo do contexto e dos requisitos específicos da aplicação, são elas: extração de Frases, extração de palavras-chave, abstração, grafo de sentenças, redes neurais e combinação de técnicas. A escolha da

técnica mais apropriada depende das características do texto, do contexto da aplicação e dos objetivos específicos do processo de sumarização.

Ambos os processos são essenciais para reduzir a complexidade e aumentar a eficácia dos sistemas de PLN, preparando os dados de uma forma que maximiza a capacidade dos modelos de extrair *insights* e realizar tarefas específicas de forma eficiente.

3 FERRAMENTAS E MÉTODOS

3.1 Ferramentas

A implementação desta aplicação baseou-se principalmente na linguagem de programação *Python,* concebida por Guido Van Rossum. De acordo com McKinney (2022, p. 68):

> Python é uma linguagem de programação versátil e poderosa, conhecida por sua sintaxe simples e elegante, que facilita a expressão de ideias complexas de forma clara e concisa. Com uma vasta gama de bibliotecas especializadas, como Pandas, NumPy e Matplotlib, Python se tornou a escolha preferida para análise de dados, aprendizado de máquina e desenvolvimento de aplicativos web, proporcionando aos usuários uma combinação única de facilidade de uso e capacidades avançadas.

Para o desenvolvimento do *front-end*, foram empregadas tecnologias como *HTML, CSS* e *JavaScrip*t, em conjunto com o *framework Bootstrap*, visando criar uma interface simples e responsiva para o usuário.

Para o controle de versão do projeto, foram utilizados o *Git* e o *GitHub*. O *Git* é um sistema de controle de versão distribuído que registra o histórico de alterações realizadas no código-fonte, possibilitando o rastreamento e a reversão para versões anteriores. Já o *GitHub* é uma plataforma de hospedagem de repositórios *Git*, facilitando a colaboração entre os membros da equipe, uma vez que permite o armazenamento remoto dos repositórios e o gerenciamento eficiente das contribuições (GitHub, 2022).

No que diz respeito ao processamento de linguagem natural (PLN), foram empregadas as bibliotecas *Sumy* e *NLTK* (*Natural Language Toolkit*), integradas ao ambiente *Python*. O *NLTK* é amplamente reconhecido como uma das bibliotecas mais poderosas para PLN, oferecendo uma ampla gama de ferramentas e recursos para trabalhar com dados de linguagem humana (*NLTK,* 2022). Além disso, a aplicação foi

hospedada em um servidor *web* especializado em hospedar aplicações *Python*, garantindo a disponibilidade e o desempenho adequados do sistema.

3.2 Metodologia

A metodologia adotada para o desenvolvimento desta aplicação foi de natureza qualitativa e aplicada, com o objetivo de atender aos requisitos estabelecidos e proporcionar uma experiência satisfatória ao usuário. Para alcançar esses objetivos, foi realizada uma pesquisa descritiva e explicativa por meio de um levantamento bibliográfico, buscando embasar teoricamente as decisões e escolhas técnicas. O processo de desenvolvimento seguiu uma sequência lógica de etapas, iniciando-se com o levantamento dos requisitos e a elaboração da documentação, onde foram detalhadas todas as características e funcionalidades da aplicação. Em seguida, foi elaborado um cronograma de trabalho, estabelecendo as etapas e prazos para a construção do sistema, de forma a estruturar o fluxo de desenvolvimento de maneira organizada e eficiente.

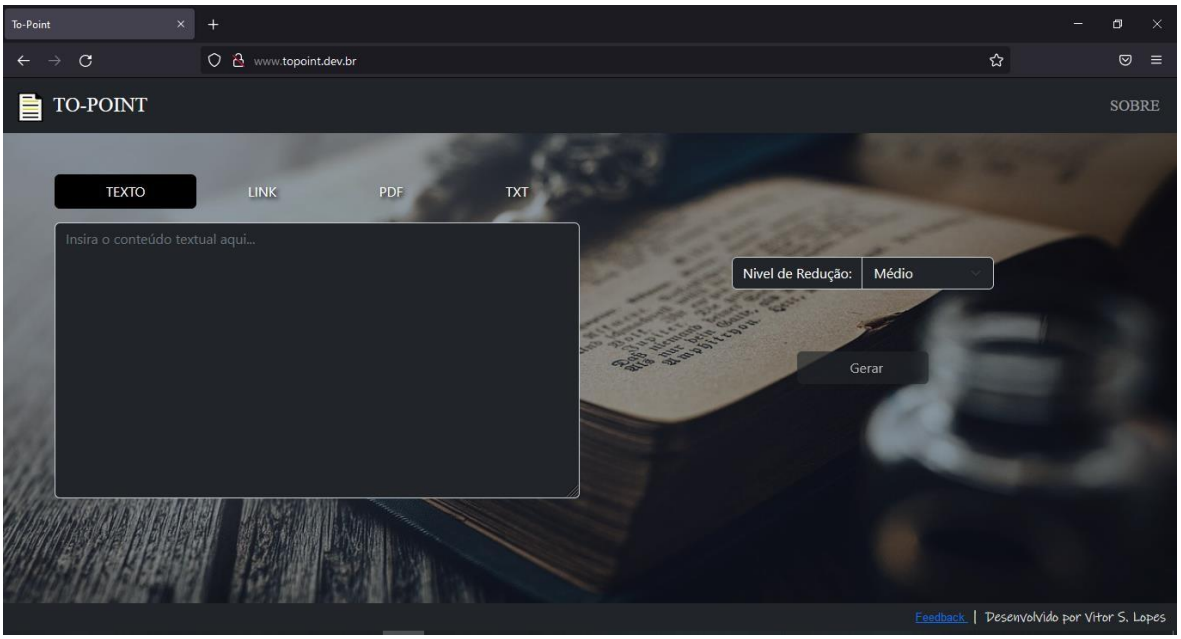
Após a definição das ferramentas a serem utilizadas, deu-se início à implementação do algoritmo de sumarização de textos. Para isso, foram empregados os algoritmos *LexRank* e *TextRank*, inspirados no funcionamento da *World Wide Web*, que buscam identificar as conexões entre as frases e destacar aquelas relacionadas com as palavras mais significativas do texto.

Posteriormente, foi realizado o desenvolvimento do *layout* da aplicação, seguindo as especificações previamente documentadas, e, por fim, a interface foi integrada aos algoritmos de sumarização e às demais funcionalidades da aplicação web, assegurando um desempenho eficaz e proporcionando uma solução satisfatória aos usuários.

4 RESULTADOS E DISCUSSÃO

Nesta etapa, são analisados os resultados da utilização da ferramenta, explorando sua eficácia na geração de resumos a partir de alguns contextos discutindo as implicações práticas, bem como são abordadas as limitações encontradas durante o processo de desenvolvimento e utilização.

Figura 2: Tela inicial



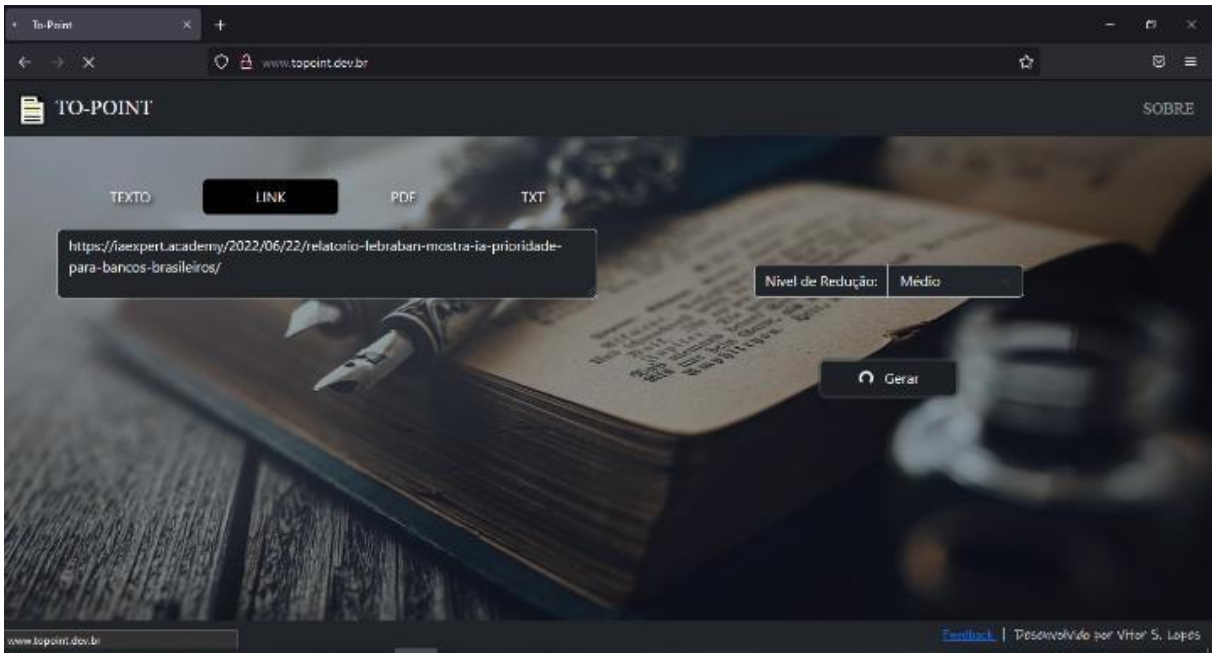
Fonte: o autor (2024)

A Figura 2 apresenta a tela inicial da aplicação, exibindo as opções disponíveis para o usuário. Inicialmente, o usuário tem a possibilidade de inserir um *link* de uma página web ou selecionar um arquivo nos formatos *PDF* ou *TXT* contendo o conteúdo textual a ser processado. Em seguida, são apresentados os níveis de redução disponíveis para aplicação no texto, permitindo ao usuário escolher o grau de sumarização desejado.

Além disso, na parte inferior da tela, encontra-se a opção de fornecer *feedback* sobre a aplicação, permitindo aos usuários compartilhar suas impressões e sugestões para melhoria. O botão "Sobre" oferece ao usuário informações adicionais sobre o projeto, possibilitando uma compreensão mais abrangente de sua finalidade e funcionalidades.

Por fim, ao clicar no botão "Gerar", inicia-se o processo de redução do texto selecionado, onde os algoritmos de sumarização são aplicados conforme as configurações definidas pelo usuário. Essa etapa marca o início da geração do resumo, que será apresentado ao usuário após a conclusão do processamento.

Figura 3: O sistema está processando o conteúdo



Fonte: o autor (2024)

Na Figura 3, observa-se o processo no qual o usuário insere um link de uma página *web* na interface da aplicação. Posteriormente, o usuário tem a opção de selecionar o nível de redução desejado para o conteúdo textual. Neste caso, foi escolhido o nível médio, que é o padrão pré-definido na aplicação.

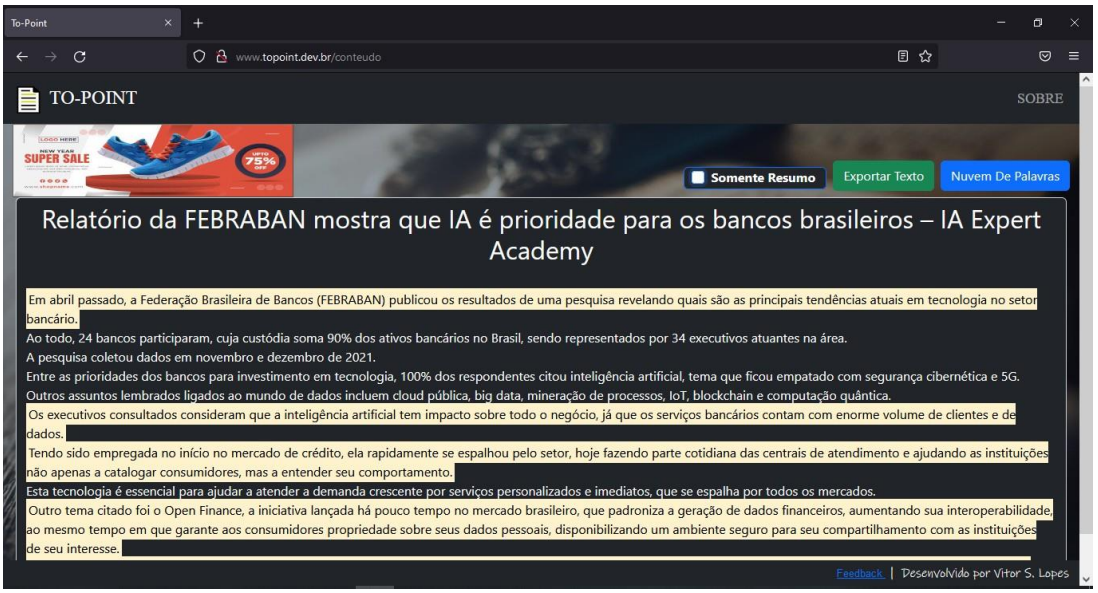
Após selecionar o nível de redução desejado, o usuário clica no botão "Gerar", dando início ao processo de geração do resumo. Durante esse processo, é exibido um ícone de carregamento, indicando que o conteúdo está sendo validado pelo sistema e que o processo de sumarização está em andamento. Esse ícone de carregamento é um indicativo visual para o usuário de que a aplicação está processando o texto e gerando o resumo conforme as configurações selecionadas.

Figura 4: Resumo Exibido

Fonte: o autor (2024)

A Figura 4 destaca a área onde o resumo será exibido, indicando assim a opção "Somente Resumo" marcada. Essa configuração permite que apenas a parte do texto considerada importante pelo algoritmo de processamento natural seja apresentada ao usuário. Além disso, é perceptível a presença de um anúncio na parte superior da tela, logo abaixo do logotipo da aplicação. Esses anúncios são gerados aleatoriamente, visando proporcionar alguma forma de receita para a aplicação.

Figura 5: Tela de visualizar o resumo com o conteúdo original

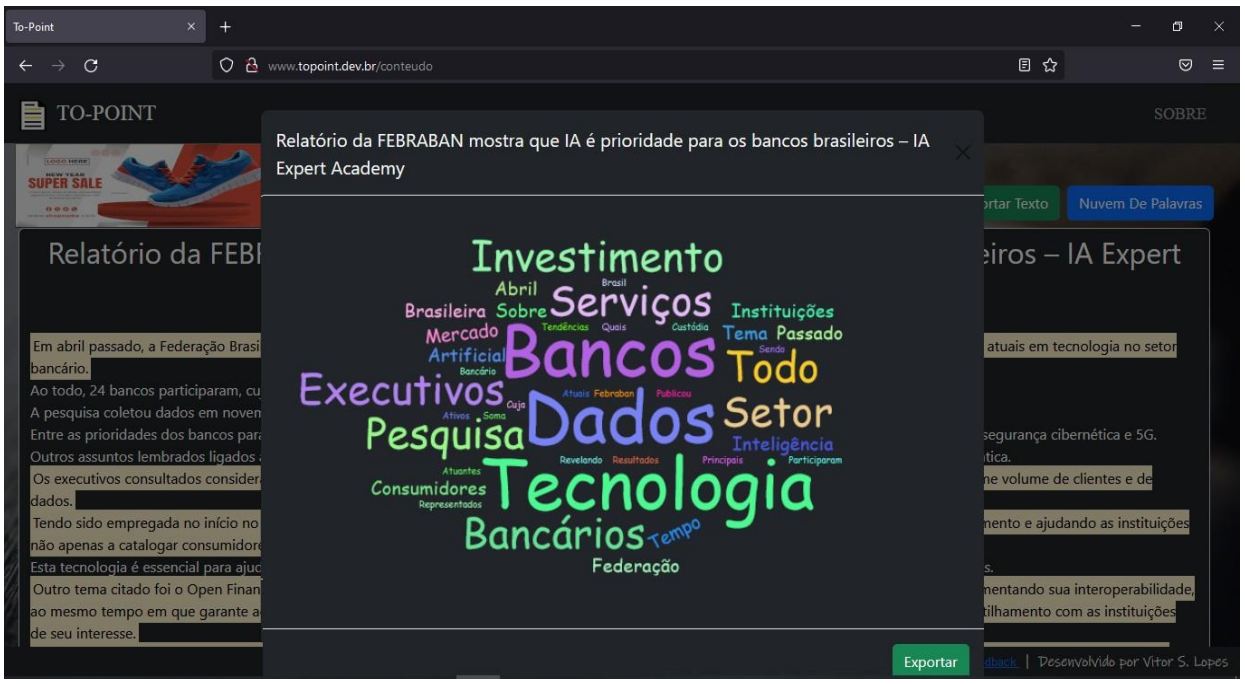


Fonte: o autor (2024)

Na Figura 5, é apresentada novamente a tela de visualização do resumo, semelhante à figura anterior. A distinção reside no fato de que a opção "Somente Resumo" está desmarcada, indicando que o sistema irá exibir o conteúdo original extraído do arquivo ou site. Além disso, em destaque, serão mostradas as principais sentenças selecionadas pelo algoritmo de processamento de linguagem natural.

Destaca-se a presença de um botão denominado "Exportar Texto" localizado acima do texto. A função desse botão é permitir que o usuário exporte o conteúdo presente na caixa de texto para o formato PDF. Essa funcionalidade oferece ao usuário a possibilidade de salvar o resumo gerado para futuras referências ou compartilhamento com outras pessoas.

Figura 6: Tela de visualizar a nuvem de palavras



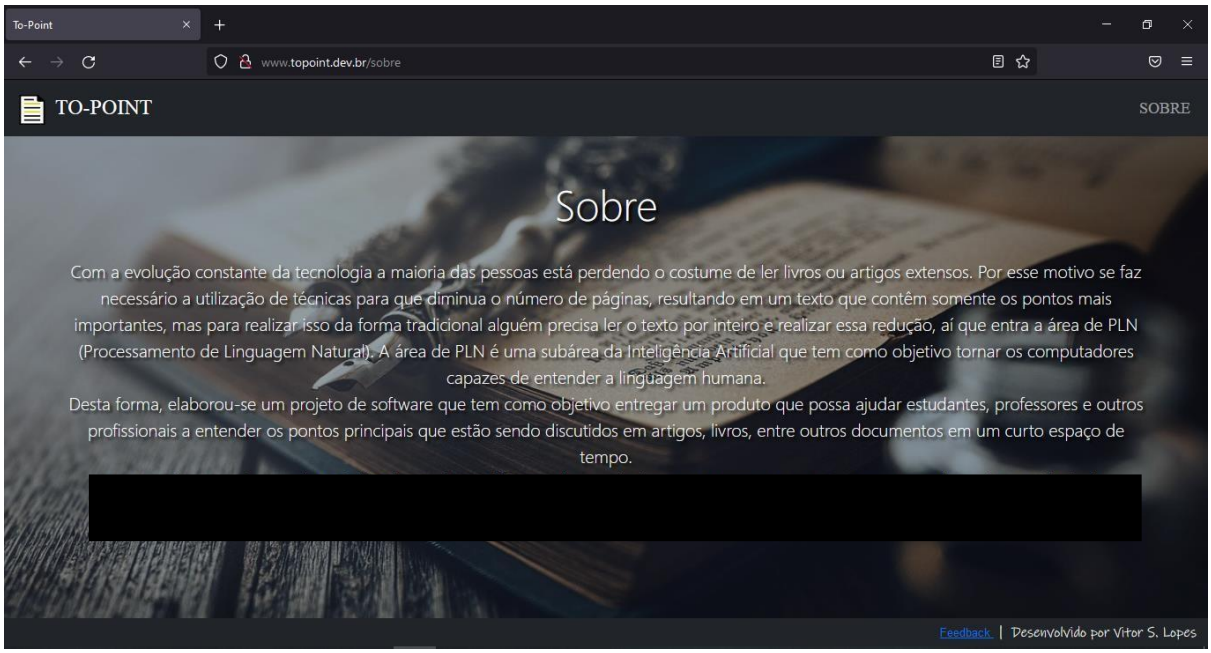
Fonte: o autor (2024)

Na Figura 6, é exibida a tela destinada à visualização da nuvem de palavras. Esta janela é acionada quando o usuário clica no botão "Nuvem de Palavras". A partir dessa ação, são apresentadas ao usuário as palavras-chave do conteúdo escrito, sendo que o tamanho de cada palavra na nuvem reflete sua importância no texto.

Na parte inferior da janela, encontra-se um botão intitulado "Exportar". A função deste botão é possibilitar ao usuário exportar a nuvem de palavras no formato PNG.

Essa funcionalidade oferece ao usuário a oportunidade de salvar a nuvem de palavras gerada para uso posterior ou compartilhamento com outros indivíduos.

Figura 7: Tela onde é falado sobre o projeto



Fonte: o autor (2024)

Na Figura 7, é apresentada uma introdução que descreve a função do software e os objetivos do projeto. Essa seção contextualiza o usuário sobre a finalidade da aplicação e o que se espera alcançar com sua utilização.

Logo abaixo dessa introdução, ao lado do nome do desenvolvedor responsável pelo projeto do software, encontra-se a opção "Feedback". Esta funcionalidade permite que os usuários do sistema deixem críticas e sugestões sobre a aplicação. Todo feedback enviado por meio dessa opção é encaminhado diretamente para o e-mail do desenvolvedor do software. Essa abordagem visa promover a participação dos usuários no processo de melhoria contínua da aplicação, garantindo que suas necessidades e expectativas sejam consideradas no desenvolvimento futuro.

CONSIDERAÇÕES FINAIS

Com base no desenvolvimento deste projeto, os objetivos iniciais estabelecidos foram alcançados com sucesso. A aplicação de algoritmos de aprendizado de máquina em um software funcional representou um desafio para aprimorar habilidades

em engenharia de aprendizado de máquina e desenvolvimento de aplicações orientadas a dados.

Ao longo do processo de desenvolvimento, diversos desafios e adversidades foram enfrentados, fazendo assim adquirir novos conhecimentos e habilidades. Um dos desafios significativos foi lidar com o processamento de linguagem natural de forma eficiente, especialmente ao trabalhar com diferentes tipos de textos e idiomas. Além disso, ao integrar a funcionalidade de resumo e nuvem de palavras feitas em *Python* à aplicação web, então foi importante garantir uma coesão com o *frontend* (HTML/CSS). A interface do usuário também deve ser intuitiva, permitindo que os resultados dos resumos e nuvens de palavras sejam apresentados de forma clara e interativa. Revisões constantes do design e da usabilidade da aplicação são essenciais para oferecer uma experiência sólida e eficaz aos usuários.

O resultado alcançado até o momento é motivo de satisfação, pois demonstra o potencial da aplicação em facilitar a compreensão e análise de textos de maneira eficiente e rápida. No entanto, ainda há espaço para melhorias e aprimoramentos. Entre as futuras melhorias planejadas estão a inclusão de novas funcionalidades, como a seleção de diferentes algoritmos de sumarização, personalização de tamanho e cor da fonte, e otimizações de processamento para tornar a aplicação ainda mais eficaz e responsiva.

Este projeto não apenas consolidou nossos conhecimentos e habilidades em áreas-chave da ciência de dados e inteligência artificial, mas também nos preparou para enfrentar novos desafios e oportunidades que possam surgir no futuro. Estamos confiantes de que esta aplicação terá um impacto significativo na forma como textos são compreendidos e analisados, contribuindo para avanços contínuos na área da linguagem natural e processamento de texto.

REFERÊNCIAS

BIRD, S., KLEIN, E., & LOPER, E. **Natural Language Processing with Python**. O'Reilly Media, Inc, 2009

DE ALENCAR, V. C.; JÚNIOR, S. F. A. X.; DE SALES, G. P. S**. Inteligência artificial: Histórico, Conceitos e Aplicações.** Editora CRV, 2024.

GOMES, D. dos S. **Inteligência Artificial: conceitos e aplicações**. Olhar Científico. v1, n. 2, p. 234-246, 2010.

JURAFSKY, D., & MARTIN, J.H. **Speech and Language Processing** (3rd ed.). Prentice Hall, 2019.

LIDDY, E.D. **Natural Language Processing**. In Encyclopedia of Library and Information Science (2nd ed.). Marcel Dekker, 2001.

MANI, I., & MAYBURY, M.T. **Advances in Automatic Text Summarization**. The MIT Press, 1999.
MAYER-SCHÖNBERGER, V., & CUKIER, K. **Big Data: A Revolution That Will Transform How We Live, Work, and Think.** Houghton Mifflin Harcourt, 2013.

MCKINNEY, Wes. **Python para análise de dados.** "O'Reilly Media, Inc.", 2022.

NLTK, **NLTK.** Disponível em: < https://www.nltk.org/>. Acesso em: 27.março, 2022.
RODRIGUES, Jéssica. **O que é o Processamento de Linguagem Natural?.** Disponível em: < https://medium.com/botsbrasil/o-que-%C3%A9-o-processamento-de-linguagem-natural-49ece9371cff>. Acesso em: 26.agosto, 2022.

SILGE, J., & ROBINSON, D. **Text Mining with R: A Tidy Approach**. O'Reilly Media, Inc, 2016.

WANG, Y. et al. **Cloud Computing for Large-Scale Resource Computation and Storage in Machine Learning.** Journal of Theory and Practice of Engineering Science, v. 4, n. 03, p. 163-171, 2024.