**Resumo**

A UFRJ mantém no Pantheon - sistema de arquivos da UFRJ que armazena toda produção científica da faculdade - um acervo de centenas de estudos sobre Inteligência Artificial. Este trabalho propõe-se a investigar esse acervo para descobrir o que uma das maiores faculdades do Brasil anda produzindo em relação a esse tópico, através da investigação e análise das dissertações, teses e monografias sobre IA. Queremos entender quais assuntos são mais relevantes e como isso mudou ao longo do tempo dentro da Instituição. Também queremos saber como o pensamento e a qualidade das publicações relacionadas à IA mudaram ao longo dos anos na faculdade e quais subtópicos foram mais estudados dentro dessa área.

### **1. Introdução**

#### **1.1 Contextualização e Importância**

A Universidade Federal do Rio de Janeiro (UFRJ), eleita a melhor universidade federal do Brasil por dez vezes consecutivas em 2023 [4], destaca-se como uma instituição de ensino de excelência no país. Dada sua reputação, é razoável supor a existência de extensas linhas de pesquisa relacionadas à Inteligência Artificial (IA) - um campo de estudo em constante destaque nos dias atuais - dentro de nossa instituição. Uma análise mais aprofundada do Pantheon, repositório digital que cataloga toda a produção acadêmica da UFRJ, revela uma ampla gama de monografias, dissertações, teses e afins publicados sobre essa temática em particular.

O presente artigo se propõe a investigar a evolução do estudo sobre Inteligência Artificial na UFRJ ao longo dos anos. Para tanto, busca-se responder às questões pertinentes: de que forma a produção acadêmica sobre Inteligência Artificial na instituição se transformou desde 2016 até o presente momento? Quais são os tópicos mais proeminentes em estudo atualmente? Como a ascensão das *Large Language Models* (LLMs) desde o lendário paper “Attention Is All You Need”[5] mudou a maneira como estudamos IA? Como o ChatGPT [6] influenciou a abordagem ao estudo da Inteligência Artificial? Essas e outras indagações serão abordadas ao longo deste trabalho.

#### **1.2 Delimitação e Justificativa do Escopo**

Inicialmente, a proposta deste estudo envolvia uma análise abrangente das pesquisas em IA realizadas globalmente ao longo do tempo. Já existem muitos estudos a respeito da evolução da IA por aí, mas a grande maioria deles analisam as obras acadêmicas em inglês [7][8]:

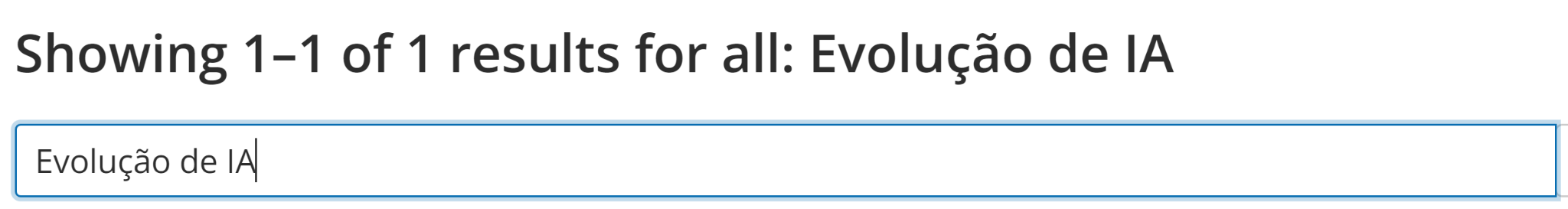


Figura 1: Quantidade de resultados retornados sobre a evolução da IA em português no Arxiv

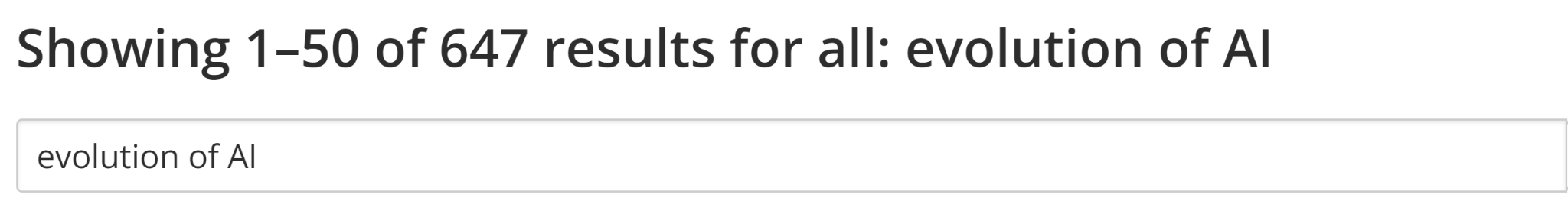


Figura 2: Quantidade de resultados retornados sobre a evolução da IA em inglês no Arxiv

Em virtude disso, optou-se por restringir o escopo para uma investigação mais focada no contexto universitário nacional. Tal decisão visa evitar qualquer viés decorrente do volume massivo de publicações que ocorrem diariamente ao redor do mundo.

https://twitter.com/emollick/status/1790727871779987833?t=cdlFob3IOufRDi\_FLog1Eg&s=08

Todas as técnicas empregadas na análise dos tópicos relevantes serão detalhadas ao longo do texto, abrangendo desde a seleção dos documentos até a extração, processamento e análise dos dados textuais relacionados à Inteligência Artificial ao longo dos anos.

### **2. Metodologia**

Para a realização deste estudo, o primeiro passo consistiu na extração de texto a partir dos arquivos PDFs contendo artigos, monografias, teses e dissertações. Essa coleção de documentos, também conhecido como *corpus*, estão no formato PDF e organizados em uma estrutura de pastas onde cada diretório representa um ano específico de produção científica na UFRJ.

Utilizamos diferentes técnicas de coleta de dados, desde a criação de *scripts* automatizados para baixar os pdfs, até o *download* desses dados manualmente. O importante era que todos os arquivos pertencessem ao repositório de produções acadêmicas da UFRJ.

Após a coleta, os dados foram filtrados ( pois só queremos trabalhos relacionados a IA) e foram utilizados as técnicas de processamento de textos como o *Stemming* de Porter [9][10]. Finalmente, algumas técnicas de Topic Modeling [11] foram aplicadas, dando origem à gráficos que nos permitem visualizar a evolução dos assuntos relacionados à IA na UFRJ ao longo dos anos.

**2.1 Coleta de Dados**

**2.2 Extração de Textos:**

Todos os textos em PDF foram extraídos e agrupados por ano. Dada à quantidade de artigos disponíveis, fizemos esse processo de maneira automatizada. Depois do pré processamento dos textos, teremos como *output*, um dataframe contendo os textos processados, filtrados e com os seus respectivos anos na forma de uma coluna.

**2.3 Pré-processamento de textos**

Como dito anteriormente, esse trabalho consiste em analisar textos de artigos científicos em português, publicados na UFRJ. Em virtude disso, grande parte desse trabalho é feito antes, na hora de processar os textos dos artigos coletados.

Como visto em [12], a língua portuguesa possui muitas flexões da mesma palavra. São flexões de gênero, flexão de grau, flexões de número, sem contar a quantidade de acentuações e pontuações, que podem tornar o processamento de textos escritos em português uma atividade muito complexa. Especialmente pelo fato de que a maioria das bibliotecas de processamento de texto disponíveis terem sido criadas visando textos em inglês. As principais etapas de processamento de texto que foram feitas, foram descritas quase que integralmente em [13]:

“

1. Remoção do plural:

Consiste em remover o “s” do final das palavras. Há uma lista de exceções como a palavra “lápis” , por exemplo.

2. Remoção do feminino:

Nesta etapa as formas femininas são transformadas na correspondente masculina. Ex.: “chinesa” → “chinês”.

3. Remoção de advérbio:

Esta é a etapa mais simples, uma vez que o único sufixo que denota um advérbio é “mente”. Neste caso também há uma lista de exceções.

4. Remoção de aumentativo e diminutivo:

Remove os sufixos dos substantivos e adjetivos que podem ter aumentativo e diminutivo. Por exemplo, “gatinha” ou “menininha”.

5. Remoção de sufixos em verbos:

Os verbos da língua portuguesa possuem mais de 50 formas diferentes de conjugação (na língua inglesa existem apenas quatro). Cada uma delas possui seu conjunto de sufixos específico. Os verbos podem variar de acordo com o tempo, a pessoa, o número e o modo. A estrutura das formas verbais pode ser representada por: radical Mineração de Textos + vogal temática + tempo + pessoa. Por exemplo: “andaram” = “and + a + ra + m”. As formas verbais são reduzidas ao seu radical correspondente.

6. Remoção de acentos:

Esta atividade é necessária porque existem vários casos onde algumas variantes são acentuadas e outras não, como em “psicólogo” e “psicologia”, por exemplo. A execução deste passo por último é importante, porque a presença de acentos é significativa em algumas regras, por exemplo: “óis” para “ol” transformando “sóis” em “sol”, por exemplo. Se a regra fosse “ois” para “ol”, poderia causar erros no caso de “dois” para “dol”.

”

No final, o que teremos é um conjunto de tokens. De acordo com [14], tokens são as unidades básicas de processamento na Linguagem Natural (NLP), que podem ser palavras, expressões fixas, idioms ou compostos que não precisam ser decompostos em etapas subsequentes. Em suma, tokens são a parte significante de palavras contidas em um documento e que podem ser contadas.

Outra etapa importante do processamento de texto é a remoção de *stop words*. Existem algumas bibliotecas como o NLTK e o Spacy, que contém módulos contendo muitas *stop words* em português. Mas como nosso objetivo é encontrar palavras que sejam relevantes para a Inteligência Artificial, essas ferramentas não cobrem toda a vasta gama de palavras irrelevantes que encontramos ao longo de muitos textos sobre o tema. Em virtude desse problema, precisamos tirar manualmente cada palavra que julgamos não contribuir em nada para nossa pesquisa. E a maneira como analisamos qual palavra contribui e qual não contribui é totalmente empírica: Verificamos os resultados e analisamos se aqueles resultados são relevantes para demonstrar aquilo que queremos demonstrar. Caso as palavras encontradas não sejam relevantes e não tenham sido deletadas previamente com o filtro das stopwords em português, nós adicionamos essas palavras em uma lista de palavras que serão ignoradas.

Todo esse processamento de texto será importante para a aplicação da técnica de Topic Modeling

### **3. Topic Modeling**

Extrair contexto e conteúdo relevante de textos pode ser uma tarefa árdua e demorada. Especialmente se o texto for muito longo e a quantidade de textos for muito grande. Alguns cenários, onde a compreensão de um determinado assunto/problema exige que você debruce-se por longos períodos analisando textos e mais textos incluem a análise de muitos documentos em longos processos judiciais, escolha de um candidato dentre muitos currículos enviados em um processo seletivo para empresas e etc. Em virtude disso, fez-se necessário a criação de ferramentas que pudessem auxiliar nesse processo. Nesse contexto surgiu o Topic Modeling.

Como explicitado em [15], Topic Modeling são estratégias que utilizam-se de métodos estatísticos capazes de extrair significados de uma grande quantidade de documentos de maneira automatizada. Ou como melhor explicado por David M. Blei em [16], Topic Modeling permite descobrir a temática principal de uma coleção de documentos não estruturados. É importante salientar que essas estratégias não exigem conhecimento prévio do que está descrito nos textos e nem que o texto esteja segmentado previamente em tópicos. Assume-se que nada sabemos sobre os textos quando utilizamos Topic Modeling. A ideia é que todo o conhecimento seja extraído diretamente do texto e supõe-se que não sabemos nada sobre eles.

Dentre as inúmeras estratégias existentes para aplicar o Topic Modeling, foram escolhidas 4: Bag of Words, TF-IDF, LDA (Latent Dirichlet Allocation) e BERT.

**3.1 Bag of Words:**

Segundo [17], o método Bag of Words (BoW) é utilizado em larga escala, tanto na visão computacional e classificação de textos, como também na classificação de imagens, vídeos e localização de robôs. Este método é uma das estratégias de Topic Modeling mais simples para categorização de texto e objetos. Ele consiste em contar a frequência de ocorrência de palavras em documentos ou características em imagens, ignorando a ordem, a gramática e o contexto em que os tokens aparecem. Na classificação de textos, BoW cria um vetor com a contagem de palavras; na classificação visual, usa descritores locais agrupados em clusters, representados por histogramas.

É importante salientar que existem outras maneiras de demonstrar a importância de tokens usando uma Bag of Words que não seja simplesmente a contagem de palavras. Mas para nossas pesquisas, optamos por utilizar essa métrica.

Apesar de sua simplicidade, Bag of Words, exige um robusto processamento de textos, pois caso contrário, os resultados não serão satisfatórios. Como dito anteriormente, em nossas pesquisas, tivemos que filtrar muitas palavras manualmente, que não estavam demonstrando aquilo que queríamos explicar apropriadamente. Apesar desses desafios, BoW mostrou-se uma técnica eficiente para a classificação de documentos e um bom método para demonstrar o quão quente estão os tópicos relacionados à IA na UFRJ.

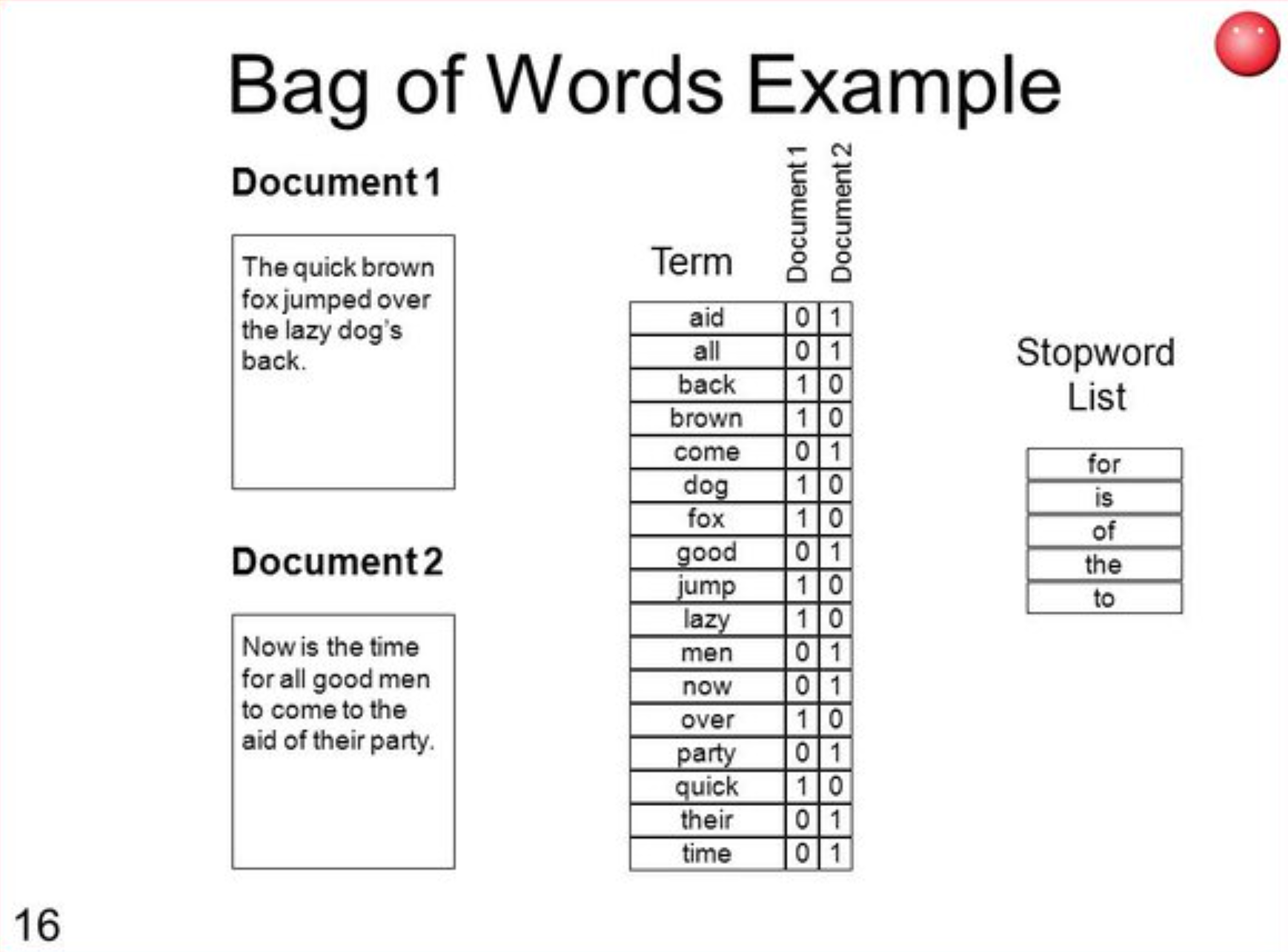


Figura 3: Representação vetorial de uma Bag of Words extraídos de [18]



Figura 4: Representação visual de uma Bag of Words extraídos de [19]

**3.1 TF-IDF**

Embora a frequência de um token em um documento possa ser um bom indicativo de sua importância, isso não implica necessariamente que o token seja relevante para o corpus como um todo. Em outras palavras, mesmo que a contagem de um token, ou sua Term Frequency (TF), seja alta em um documento específico, como bem explicado em [20][21], a relevância desse token pode ser diminuída se ele também aparece frequentemente em outros documentos do mesmo corpus. Portanto, a alta frequência de um token em um documento não garante sua significância geral. Em resumo, o que foi dito é que um token é relevante quando ele é muito frequente em um único documento, mas que é raro para outros documentos, dentro do mesmo corpus.

Conforme mencionado anteriormente, TF (Term Frequency) refere-se à frequência de termos, enquanto IDF (Inverse Document Frequency) refere-se à frequência inversa de documentos. A estratégia TF-IDF é considerada mais robusta em comparação com a abordagem Bag of Words, pois esta última está incorporada na primeira. Em outras palavras, a abordagem Bag of Words corresponde à parte TF da estratégia TF-IDF.

A Inverse Document Frequency (IDF) é uma métrica utilizada no algoritmo TF-IDF para atribuir maior peso a palavras menos frequentes em um conjunto de documentos e menor peso a palavras comuns, como artigos e preposições. Isso ajuda a identificar termos que são mais importantes para a distinção dos documentos no corpus. O IDF é calculado pela fórmula abaixo:

IDF = log e​(N/n​), onde N é o número total de documentos e n é o número de documentos que contêm o termo.

Finalmente, TF-IDF é dado pelo valor da frequência do token, multiplicado pelo cálculo de seu IDF.

**3.1.1 Problemas com a abordagem TF-IDF**

O principal problema com a estratégia de TF-IDF é que se o processamento prévio de textos for ruim, TF não apresentará um bom resultado. Como dito em [21], o principal problema de TF-IDF é que ele não consegue reconhecer palavras com variações de tempo, tratando formas como "go" e "goes" ou "play" e "playing" como palavras distintas, o que pode levar a inconsistências. Por isso, um bom processamento de textos e uso de estratégias como Stemming de Porter são tão necessários.

Referências:

[1]

* url: <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html>
* acessado em: 05/05/24
* autor: gov.br

[2]

* url: <http://members.unine.ch/jacques.savoy/clef/>
* acessado em: 05/05/24
* autor: gov.br

[3]

* url: <http://repositorio.poli.ufrj.br/rep-relat-projetocursoano.php>
* acessado em: 05/05/24
* autor: politécnica ufrj

[4]

* url: <https://www.parque.ufrj.br/ufrj-e-eleita-melhor-universidade-federal-do-brasil-pela-10a-vez-consecutiva-em-ranking-internacional/>
* acessado em: 05/05/24
* autor: <https://www.parque.ufrj.br/>

[5] VASWANI, Ashish et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

[6] LIANG, Weixin et al. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. **arXiv preprint arXiv:2403.07183**, 2024.

[7] SPECTOR, Lee. Evolution of artificial intelligence. **Artificial Intelligence**, v. 170, n. 18, p. 1251-1253, 2006.

[8] VASHISHTH, Tarun Kumar et al. The Evolution of AI and Its Transformative Effects on Computing: A Comparative Analysis. In: **Intelligent Engineering Applications and Applied Sciences for Sustainability**. IGI Global, 2023. p. 425-442.

[9] PORTER, Martin F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p. 130-137, 1980.

[10] ORENGO, Viviane Moreira; HUYCK, Christian R. A Stemming Algorithmm for the Portuguese Language. In: **spire**. 2001. p. 186-193.

[11] ALGHAMDI, Rubayyi; ALFALQI, Khalid. A survey of topic modeling in text mining. **Int. J. Adv. Comput. Sci. Appl.(IJACSA)**, v. 6, n. 1, 2015.

[12] VILLALVA, Alina; MATEUS, Maria Helena Mira. **Morfologia do português**. Lisboa: Universidade Aberta, 2008.

[13] MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

[14] WEBSTER, Jonathan J.; KIT, Chunyu. Tokenization as the initial phase in NLP. In: **COLING 1992 volume 4: The 14th international conference on computational linguistics**. 1992.

[15] KHERWA, Pooja; BANSAL, Poonam. Topic modeling: a comprehensive review. **EAI Endorsed transactions on scalable information systems**, v. 7, n. 24, 2019.

[16] BLEI, David M. Probabilistic topic models. **Communications of the ACM**, v. 55, n. 4, p. 77-84, 2012.

[17] QADER, Wisam A.; AMEEN, Musa M.; AHMED, Bilal I. An overview of bag of words; importance, implementation, applications, and challenges. In: **2019 international engineering conference (IEC)**. IEEE, 2019. p. 200-204.

[18]

* url: https://www.quora.com/What-is-the-difference-between-the-Bag-of-Words-model-and-the-Continuous-Bag-of-Words-model
* acessado em: 04/06/24
* autor: <https://pt.quora.com/>

[19]

* url: https://es.mathworks.com/discovery/bag-of-words.html
* acessado em: 04/06/24
* autor: <https://es.mathworks.com/>

[20] RAMOS, Juan et al. Using tf-idf to determine word relevance in document queries. In: **Proceedings of the first instructional conference on machine learning**. 2003. p. 29-48.

[21] QAISER, Shahzad; ALI, Ramsha. Text mining: use of TF-IDF to examine the relevance of words to documents. **International Journal of Computer Applications**, v. 181, n. 1, p. 25-29, 2018.