1 Document Content

Busca de Assuntos Interessantes em IA:
Estrat gias de Topic Modeling

Wagner Lobo; Leandro Yamachita
24 de junho de 2024
Resumo
A UFRJ mant m no Pantheon - sistema de arquivos da UFRJ que armazena toda produ o cient fica da faculdade - um acervo de centenas de estudos sobre Intelig ncia Artificial. Este trabalho prop e-se a investigar esse acervo para descobrir o que uma das maiores faculdades do Brasil anda produzindo em rela o a esse t pico, atrav s da investiga o e an lise das disserta es, teses e monografias sobre IA. Queremos entender quais assuntos s o mais relevantes e como isso mudou ao longo do tempo dentro da Institui o. Tamb m queremos saber como o pensamento e a qualidade das publica es relacionadas IA mudaram ao longo dos anos na faculdade e quais subt picos foram mais estudados dentro dessa rea.
1. Introdu o
1.1 Contextualiza o e Import ncia
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 institui o de ensino de excel ncia no pa s. Dada sua reputa o, razo vel supor a exist ncia de extensas linhas de pesquisa relacionadas Intelig ncia Artificial (IA) - um campo de estudo em constante destaque nos dias atuais - dentro de nossa institui o. Uma an lise mais aprofundada do Pantheon, reposit rio digital que cataloga toda a produ o acad mica da UFRJ, revela uma ampla gama de monografias, disserta es, teses e afins publicados sobre essa tem tica em particular.
O presente artigo se prop e a investigar a evolu o do estudo sobre Intelig ncia Artificial na UFRJ ao longo dos anos. Para tanto, busca-se responder s quest es pertinentes: de que forma a produ o acad mica sobre Intelig ncia Artificial na institui o se transformou desde 2010 at o presente momento? Quais s o os t picos mais proeminentes em estudo atualmente? Como a ascens o das Large Language Models (LLMs) desde o lend rio paper Attention Is All You Need [5] mudou a maneira como estudamos IA? Como o ChatGPT [6] influenciou a abordagem ao estudo da Intelig ncia Artificial? Essas e outras indaga es ser o abordadas ao longo deste trabalho.
1.2 Delimita o e Justificativa do Escopo
Inicialmente, a proposta deste estudo envolvia uma an lise abrangente das pesquisas em IA realizadas globalmente ao longo do tempo. J existem muitos estudos a respeito da evolu o da IA por a , mas a grande maioria deles analisam as obras acad micas em ingl s [7][8]:
Figura 1: Quantidade de resultados retornados sobre a evolu o da IA em portugu s no Arxiv
Figura 2: Quantidade de resultados retornados sobre a evolu o da IA em ingl s no Arxiv
Em virtude disso, optou-se por restringir o escopo para uma investiga o mais

focada no contexto universit rio nacional. Tal decis o visa evitar qualquer vi s decorrente do volume massivo de publica es que ocorrem diariamente ao redor do mundo.
Todas as t cnicas empregadas na an lise dos t picos relevantes ser o detalhadas ao longo do texto, abrangendo desde a sele o dos documentos at a extra o, processamento e an lise dos dados textuais relacionados Intelig ncia Artificial ao longo dos anos.
2. Metodologia
Para a realiza o deste estudo, o primeiro passo consistiu na extra o de texto a partir dos arquivos PDFs contendo artigos, monografias, teses e disserta es. Essa cole o de documentos, tamb m conhecido como corpus, est o no formato PDF e organizados em uma estrutura de pastas onde cada diret rio representa um ano espec fico de produ o cient fica na UFRJ.
Utilizamos diferentes t cnicas de coleta de dados, desde a cria o de scripts automatizados para baixar os pdfs, at o download desses dados manualmente. O importante era que todos os arquivos pertencessem ao reposit rio de produ es acad micas da UFRJ.
Ap s a coleta, os dados foram filtrados (pois s queremos trabalhos relacionados a IA) e foram utilizados as t cnicas de processamento de textos como o Stemming de Porter [9][10]. Finalmente, algumas t cnicas de Topic Modeling [11] foram aplicadas, dando origem gr ficos que nos permitem visualizar a evolu o dos assuntos relacionados IA na UFRJ ao longo dos anos.
2.1 Coleta de Dados
Os dados para este estudo foram obtidos exclusivamente da plataforma Pantheon. Foram selecionados trabalhos categorizados como Teses e Disserta es (T&Ds) e Trabalhos de Conclus o de Curso (TCCs). Todos os estudos dentro dessas categorias, incluindo metadados e o arquivo PDF de cada trabalho, foram coletados de maneira automatizada atrav s de scripts desenvolvidos para esse fim. Ao todo, foram reunidos 3050 T&Ds e 16490 TCCs de todos os cursos e programas oferecidos pela UFRJ.
2.1.1 Filtragem dos trabalhos
Inicialmente, realizou-se uma filtragem para incluir apenas trabalhos a partir do ano de 2010, resultando em 1622 T&Ds e 14264 TCCs. Em seguida, uma filtragem espec fica foi aplicada para selecionar apenas os trabalhos relacionados Intelig ncia Artificial (IA). Esta etapa utilizou o ChatGPT, acessado via API oferecida pela OpenAI. Cada trabalho foi submetido ao ChatGPT atrav s do prompt descrito na Figura 3, utilizando apenas o t tulo, resumo e palavras-chave para determinar sua relev ncia para o estudo. Os trabalhos foram classificados com base nas respostas "sim" ou "n o" obtidas do ChatGPT. Foi utilizado a vers o gpt-3.5-turbo-0125 do ChatGPT. Ap s esta etapa, foram selecionados 91 T&Ds e 110 TCCs relevantes para o tema de IA.
Figura 3: Prompt para coleta de dados
A distribui o dos trabalhos por ano e categoria mostrada na Figura 4. Nota-se que, para T&Ds relacionadas IA, foram encontradas publica es nos anos de 2017, 2018, 2019 e 2020, com poucos registros em 2020. No caso dos TCCs, observa-se um aumento significativo de trabalhos a partir de 2017. A Figura 5

apresenta a distribui o geral dos trabalhos inclu dos neste estudo. importante ressaltar que os padr es encontrados podem refletir tanto a produ o real quanto a disponibilidade dos trabalhos na plataforma Pantheon.
Figura 4: Distribui o dos Trabalhos por Ano e Categoria
Figura 5: Distribui o Geral dos Trabalhos
2.2 Extra o de Textos:
Todos os textos em PDF foram extra dos e agrupados por ano. Dada quantidade de artigos dispon veis, fizemos esse processo de maneira automatizada. Depois do pr 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 Pr -processamento de textos
Como dito anteriormente, esse trabalho consiste em analisar textos de artigos cient ficos em portugu s, publicados na UFRJ. Em virtude disso, grande parte desse trabalho feito antes, na hora de processar os textos dos artigos coletados. Como visto em [12], a l ngua portuguesa possui muitas flex es da mesma palavra. S o flex es de g nero, flex o de grau, flex es de n mero, sem contar a quantidade de acentua es e pontua es, que podem tornar o processamento de textos escritos em portugu s uma atividade muito complexa. Especialmente pelo fato de que a maioria das bibliotecas de processamento de texto dispon veis terem sido criadas visando textos em ingl s. As principais etapas de processamento de texto que foram feitas, foram descritas quase que integralmente em [13]:

1. Remo o do plural:
Consiste em remover o s do final das palavras. H uma lista de exce es como a palavra l pis , por exemplo.
2. Remo o do feminino:
Nesta etapa as formas femininas s o transformadas na correspondente masculina. Ex.: chinesa chin s .
3. Remo o de adv rbio:
Esta a etapa mais simples, uma vez que o nico sufixo que denota um adv rbio mente . Neste caso tamb m h uma lista de exce es.
4. Remo o de aumentativo e diminutivo:
Remove os sufixos dos substantivos e adjetivos que podem ter aumentativo e diminutivo. Por exemplo, gatinha ou menininha .
5. Remo o de sufixos em verbos:
Os verbos da l ngua portuguesa possuem mais de 50 formas diferentes de conjuga o (na l ngua inglesa existem apenas quatro). Cada uma delas possui seu conjunto de sufixos espec fico. Os verbos podem variar de acordo com o tempo, a pessoa, o n mero e o modo. A estrutura das formas verbais pode ser representada por: radical Minera o de Textos + vogal tem tica + tempo + pessoa. Por exemplo: andaram = and + a + ra + m . As formas verbais s o reduzidas ao seu radical correspondente.
6. Remo o de acentos:
Esta atividade necess ria porque existem v rios casos onde algumas variantes s o acentuadas e outras n o, como em psic logo e psicologia , por exemplo. A execu

o deste passo por ltimo importante, porque a presen a de acentos significativa em algumas regras, por exemplo: is para ol transformando s is 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 um conjunto de tokens. De acordo com [14], tokens s o as unidades b sicas de processamento na Linguagem Natural (NLP), que podem ser palavras, express es fixas, idioms ou compostos que n o precisam ser decompostos em etapas subsequentes. Em suma, tokens s o a parte significante de palavras contidas em um documento e que podem ser contadas.
Outra etapa importante do processamento de texto a remo o de stop words. Existem algumas bibliotecas como o NLTK e o Spacy, que cont m m dulos contendo muitas stop words em portugu s. Mas como nosso objetivo encontrar palavras que sejam relevantes para a Intelig ncia Artificial, essas ferramentas n o 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 n o contribuir em nada para nossa pesquisa. E a maneira como analisamos qual palavra contribui e qual n o contribui totalmente emp rica: Verificamos os resultados e analisamos se aqueles resultados s o relevantes para demonstrar aquilo que queremos demonstrar. Caso as palavras encontradas n o sejam relevantes e n o tenham sido deletadas previamente com o filtro das stop words em portugu s, n s adicionamos essas palavras em uma lista de palavras que ser o ignoradas.
Todo esse processamento de texto ser importante para a aplica o da t cnica de Topic Modeling.
3. Topic Modeling
Extrair contexto e conte do relevante de textos pode ser uma tarefa rdua e demorada. Especialmente se o texto for muito longo e a quantidade de textos for muito grande. Alguns cen rios, onde a compreens o de um determinado assunto/problema exige que voc debruce-se por longos per odos analisando textos e mais textos incluem a an lise de muitos documentos em longos processos judiciais, escolha de um candidato dentre muitos curr culos enviados em um processo seletivo para empresas e etc. Em virtude disso, fez-se necess rio a cria o de ferramentas que pudessem auxiliar nesse processo. Nesse contexto surgiu o Topic Modeling.
Como explicitado em [15], Topic Modeling s o estrat gias que utilizam-se de m todos estat sticos 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 tem tica principal de uma cole o de documentos n o estruturados. importante salientar que essas estrat gias n o exigem conhecimento pr vio do que est descrito nos textos e nem que o texto esteja segmentado previamente em t picos. Assume-se que nada sabemos sobre os textos quando utilizamos Topic Modeling. A ideia que todo o conhecimento seja extra do diretamente do texto e sup e-se que n o sabemos nada sobre eles.
Dentre as in meras estrat gias existentes para aplicar o Topic Modeling, foram escolhidas 4: Bag of Words, TF-IDF, LDA (Latent Dirichlet Allocation) e

BERTopic.

3.1 Bag of Words:

Segundo [17], o m todo Bag of Words (BoW) utilizado em larga escala, tanto na vis o computacional e classifica o de textos, como tamb m na classifica o de imagens, v deos e localiza o de rob s. Este m todo uma das estrat gias de Topic Modeling mais simples para categoriza o de texto e objetos. Ele consiste em contar a frequ ncia de ocorr ncia de palavras em documentos ou caracter sticas em imagens, ignorando a ordem, a gram tica e o contexto em que os tokens aparecem. Na classifica o de textos, BoW cria um vetor com a contagem de palavras; na classifica o visual, usa descritores locais agrupados em clusters, representados por histogramas.

importante salientar que existem outras maneiras de demonstrar a import ncia de tokens usando uma Bag of Words que n o seja simplesmente a contagem de palavras. Mas para nossas pesquisas, optamos por utilizar essa m trica.

Apesar de sua simplicidade, Bag of Words, exige um robusto processamento de textos, pois caso contr rio, os resultados n o ser o satisfat rios. Como dito anteriormente, em nossas pesquisas, tivemos que filtrar muitas palavras manualmente, que n o estavam demonstrando aquilo que quer amos explicar apropriadamente. Apesar desses desafios, BoW mostrou-se uma t cnica eficiente para a classifica o de documentos e um bom m todo para demonstrar o qu o quente est o os t picos relacionados IA na UFRJ.

Figura 6: Representa o vetorial de uma Bag of Words extra dos de [18]

Figura 7: Representa o visual de uma Bag of Words extra dos de [19]

3.2 TF-IDF

Embora a frequ ncia de um token em um documento possa ser um bom indicativo de sua import ncia, isso n o 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 espec fico, como bem explicado em [20][21], a relev ncia desse token pode ser diminu da se ele tamb m aparece frequentemente em outros documentos do mesmo corpus. Portanto, a alta frequ ncia de um token em um documento n o garante sua signific ncia geral. Em resumo, o que foi dito que um token relevante quando ele muito frequente em um nico documento, mas que raro para outros documentos, dentro do mesmo corpus.

Conforme mencionado anteriormente, TF (Term Frequency) refere-se frequ ncia de termos, enquanto IDF (Inverse Document Frequency) refere-se frequ ncia inversa de documentos. A estrat gia TF-IDF considerada mais robusta em compara o com a abordagem Bag of Words, pois esta ltima est incorporada na primeira. Em outras palavras, a abordagem Bag of Words corresponde parte TF da estrat gia TF-IDF.

A Inverse Document Frequency (IDF) uma m trica 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 preposi es. Isso ajuda a identificar termos que s o mais importantes para a distin o dos documentos no corpus. O IDF calculado pela f rmula abaixo:

IDF = log e (N/n), onde N o n mero total de documentos e n o n mero de

documentos que cont m o termo.
Finalmente, TF-IDF dado pelo valor da frequ ncia do token, multiplicado pelo c lculo de seu IDF.
3.2.1 Problemas com a abordagem TF-IDF
O principal problema com a estrat gia de TF-IDF que se o processamento pr vio de textos for ruim, TF n o apresentar um bom resultado. Como dito em [21], o principal problema de TF-IDF que ele n o consegue reconhecer palavras com varia es de tempo, tratando formas como "go" e "goes" ou "play" e "playing" como palavras distintas, o que pode levar a inconsist ncias. Por isso, um bom processamento de textos e uso de estrat gias como Stemming de Porter s o t o necess rios.
3.3 Latent Dirichlet allocation (LDA)
Segundo [16] Latent Dirichlet Allocation (LDA) um m todo de Topic Modeling que parte da hip tese em que cada documento uma mistura de v rios t picos e que cada t pico uma distribui o de palavras. A intui o por tr s do LDA que documentos n o s o homog neos, mas sim compostos por uma mistura de m ltiplos t picos. Em nosso caso espec fico, queremos encontrar v rios t picos relacionados Intelig ncia Artificial.
Partindo do pressuposto que um documento uma mistura de v rios t picos e cada t pico, por sua vez, uma cole o de palavras que frequentemente aparecem juntas, usando o exemplo utilizado em [16], em um artigo sobre gen tica e biologia, pode-se encontrar palavras como "genes", "DNA", "evolu o" e "organismos". LDA identifica esses grupos de palavras e os associa a t picos espec ficos.
O funcionamento do LDA pode ser entendido em duas etapas principais:
Defini o de T picos e Correla o Entre T picos: Al m de clusterizar termos em t picos, LDA tamb m identifica a distribui o de t picos em cada documento. Ou seja, cada documento uma mistura de v rios t picos, e cada palavra no documento atribu da a um desses t picos com base em uma distribui o de probabilidade.
Correla o Entre T picos: Os t picos podem compartilhar alguns termos, e h uma distribui o de t picos em cada documento. Ent o, enquanto palavras dentro de um t pico s o fortemente correlacionadas, t picos diferentes podem ainda compartilhar algumas palavras. A correla o entre t picos pode existir, mas os t picos s o definidos para maximizar a coes o interna (termos fortemente relacionados dentro de um t pico) e a separa o externa (t picos distintos s o compostos por termos diferentes).
Por exemplo, se um artigo sobre 70% gen tica e 30% biologia evolutiva, LDA usa essas propor es para associar palavras aos t picos relevantes. No final, o que se tem uma vis o de quais t picos est o presentes em cada documento e em que medida, sem precisar ler todos os textos. Isso facilita a organiza o e a an lise de grandes volumes de texto, ajudando a identificar padr es e tend ncias tem ticas que seriam dif ceis de perceber manualmente.
3.3.1. Robustez de Latent Dirichlet allocation (LDA)
LDA uma estrat gia de Topic Modeling bem mais robusta que o Bag of Words e o TF-IDF. Isso porque sua execu o n o depende exclusivamente da frequ ncia de palavras como nas duas outras estrat gias. LDA consegue identificar diferentes

t picos, agrupando diferentes termos, baseado em probabilidade e n o apenas na contagem de palavras.Al m do mais, TF-IDF e Bag of Words s o eficazes para algumas tarefas mais espec ficas como recupera o de informa o e classifica o de textos, mas s o bem limitados quando se trata de aplica es que requerem uma maior compreens o dos t picos.

3.4 BERTopic

O BERTopic utiliza modelos baseados em Transformers para criar representa es (embeddings) de documentos. Esses embeddings s o ent o utilizados para agrupar automaticamente documentos similares em diferentes t picos, eliminando a necessidade de pr -determinar o n mero de t picos. Os clusters resultantes s o interpretados como t picos, e para cada cluster s o identificadas palavras que melhor representam o conte do de cada t pico. comum empregar o m todo TF-IDF baseado em classe (c-TF-IDF) para selecionar as palavras representativas de cada t pico. Nesta abordagem, o TF-IDF calculado considerando cada cluster como um documento nico, obtido pela concatena o dos documentos pertencentes ao mesmo cluster. Ap s a extra o das palavras chave de cada t pico, poss vel utilizar um modelo de linguagem para resumir a descri o do t pico em uma nica senten a, com base nestas palavras.

Abaixo s o descrito os passos que foram desenvolvidos neste trabalho para a gera o de t picos atrav s do BERTopic:

3.4.1 - Prepara o dos textos:

Os documentos estudados foram representados apenas por seus t tulos, resumos e palavras-chave. Todos os trabalhos, incluindo teses de doutorado (T&Ds) e trabalhos de conclus o de curso (TCCs), foram agrupados sem distin o de tipo. Cada texto passou por um pr -processamento simples, que incluiu remo o de pontua es, stopwords, sequ ncias de escape, al m da convers o para letras min sculas.

3.4.2 - Embedding

Todos os documentos foram representados por vetores de embeddings de 768 dimens es gerados pelo BERTimbau, um modelo pr -treinado em dados em portugu s.

3.4.3 - Redu o de dimensionalidade

Em seguida, aplicou-se o algoritmo UMAP para reduzir a dimensionalidade dos embeddings para 4 dimens es. Essa etapa crucial para otimizar a clusteriza o posterior, pois embeddings de alta dimens o podem capturar varia es irrelevantes nos dados. A redu o de dimensionalidade concentra-se nas varia es mais significativas e interpret veis.

3.4.4 - Clusteriza o

Os embeddings reduzidos foram agrupados em clusters utilizando o algoritmo HDBSCAN. Cada cluster representa um t pico distinto, com um m nimo de 3 documentos por cluster.

3.4.5 - Palavras-chave dos t picos

Para identificar as palavras mais representativas de cada cluster, utilizou-se o m todo c-TF-IDF. Este m todo calcula o TF-IDF das palavras considerando cada cluster como um documento nico.

4 Resultados

O resultado esperado com esse artigo era demonstrar a produ o acad mica na UFRJ ao longo dos anos. Para que consegu ssemos obter resultados minimamente coerentes com a proposta deste artigo, muita pr processamento e pr filtragem de texto precisou ser feita. Especialmente quando utilizando-se das estrat gias de Bag of Words e TF-IDF. Como utilizamos uma biblioteca de processamento de textos que n o muito adequada para o portugu s (spacy.load("pt_core_news_sm")) , tivemos que criar um longo dicion rio de dados, contendo palavras que precisavam ser exclu das dos gr ficos. Precisamos deletar palavras como naqueles , baixo , por , anos , tamb m e etc.

4.1 Resultados de Bag of Words e TF-IDF

Como dito anteriormente, Bag of Words e TF-IDF baseiam-se fortemente na contagem de palavras. E por causa disso, a qualidade de seus outputs depende fortemente da quantidade de documentos dispon veis. Como n o temos uma quantidade t o significativa de textos, os resultados n o foram muito satisfat rios. Isso especialmente vis vel quando voc est trabalhando com um ano em que n o temos muitos pdfs contendo a produ o acad mica da UFRJ, dado a problemas t cnicos na extra o de dados como os anos de 2016 e 2023

Figura 8: Representa o da Bag of Words ao longo dos anos

Figuras 9 e 10: Representa o gr fica do score TF-IDF de 2016 at 2023

Com a nuvem de palavras e os gr ficos de TF-IDF representando os assuntos mais estudados relacionados IA, retirados diretamente de produ es acad micas da UFRJ, podemos fazer as seguintes considera es, que podem ser observadas nas 2 estrat gias de Topic Modeling:

Ao longo dos anos, os t picos discutidos sobre Intelig ncia Artificial na UFRJ evolu ram bastante. Inicialmente, em 2016, o foco estava na organiza o e pr tica de pesquisa, destacando termos como "pesquisa", "base" e "pr tica". Com o passar do tempo, houve uma mudan a para a an lise de dados e t cnicas aplicadas, refletindo um interesse crescente em redes neurais e aprendizado de m quina, como era de se esperar. Termos como "analysis", "applications" e "technique" se tornaram mais frequentes, indicando um aprofundamento nas metodologias e na aplica o de t cnicas avan adas de IA.

A partir de 2020, observamos um destaque maior em predi o e simula o, com nfase em "neuron" e "prediction". Isso aponta um foco mais intenso no aprendizado de m quina e suas aplica es pr ticas. Nos anos seguintes, a discuss o se expandiu para incluir uma quantidade maior de reas, incluindo geologia e outras ci ncias, demonstrando a aplica o abrangente das t cnicas de IA.

vis vel a quantidade de termos que n o contribuem em nada para a IA, nos dois tipos de gr ficos. Isso porque, como dito anteriormente, Nuvem de Palavras e TF-IDF dependem fortemente da frequ ncia das palavras. Al m do mais, temos uma amostra muito pequena para poder inferir qualquer coisa mais avan ada usando as duas estrat gias.

4.2 Resultados de Latent Dirichlet Allocation (LDA)

Por mais inesperado que isso seja, os resultados dessa an lise usando como estrat gia o LDA, tiveram resultados piores do que as estrat gias de Nuvem de Palavras e TF-IDF. Isso provavelmente ocorreu devido a pouca quantidade de amostra, especialmente em alguns anos espec ficos. O LDA n o baseia sua m

trica em cima da contagem de palavras, mas sim em clusteriza o atrav s de probabilidades. Como temos uma quantidade de amostra menor para alguns anos, o resultado conter muitas palavras aleat rias que nada tem a ver com IA, conforme visto nas imagens abaixo:
Figuras 11 e 12 de gr ficos de LDA sobre IA.
Algumas considera es:
Com uma amostra pequena, o modelo LDA pode n o ser capaz de capturar t picos importantes de forma adequada. Modelos de t picos geralmente funcionam melhor com grandes volumes de dados, onde a diversidade de termos e a estrutura dos t picos s o mais evidentes.
A qualidade dos dados utilizados para a an lise pode n o ser boa o suficiente para aplicar o LDA. Se a base de dados contiver muitos termos irrelevantes ou ru dos, o modelo LDA pode acabar identificando esses termos como t picos.
A quantidade de t picos escolhida (3) foi aleat ria. Talvez uma escolha mais adequada da quantidade de t pico leve a resultados mais eficientes
Em suma, mesmo a t cnica LDA sendo muito mais robusta que Nuvem de Palavras ou TF-IDF, para que essa superioridade fique evidente, precisamos de amostras maiores.
4.3 Resultados de BERT
Para definir cada t pico em uma breve senten a, utilizou-se o ChatGPT. O prompt incluiu os 4 textos mais relevantes de cada t pico, determinados pela similaridade com a representa o do t pico usando os valores do c-TF-IDF. Al m disso, foram inclu das as 10 palavras-chave mais importantes calculadas para cada t pico atrav s do c-TF-IDF.
Os t picos obtidos atrav s deste processo podem ser vistos na tabela abaixo:
Refer ncias:
[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: polit cnica 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 portugu s. Lisboa: Universidade Aberta, 2008.
[13] MORAIS, Edison Andrade Martins; AMBR SIO, Ana Paula L. Minera o de textos. Relat rio T cnico Instituto de Inform tica (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.

T picos	N de Trabalhos
Aprendizado de M quina em Processos Operacionais Industriais.	42
Detec o de pl gio e not cias falsas em portugu s.	19
Previs o e monitoramento de energia sustent vel e eficiente.	18
Redes Neurais em Combust o e L gica Fuzzy	11
Estima o e propaga o ac stica utilizando redes neurais	8
Tecnologias emergentes na auditoria e contabilidade empresarial	8
Processamento de udio e ac stica em diversos contextos	7
Intelig ncia Artificial em Esportes e Jogos Digitais	7
Detec o de resist ncia e tropismo viral em HIV	7
Detec o e classifica o de part culas em experimentos	7
Previs o e an lise de varia es clim ticas e falhas	6
An lise de minera o de dados no mercado financeiro	6
Explora o geol gica e petr leo atrav s de dados	6
Aprendizado de m quina em rob tica e jogos competitivos	6
Sistema de Previs o e Suporte para Usinas Nucleares	5
Previs o e otimiza o em geotecnia e engenharia civil	5
Interface C rebro-M quina para Controle de Movimentos de Membros	5
Detec o avan ada de amea as em redes de seguran a	5
Detec o de Fake News usando Classificadores de Texto	5
Controle e previs o em processos e s ries temporais	4
Detec o e classifica o ac stica de submarinos e navios	4
Classifica o espectral de estrelas de alta massa	4
Aprendizado por Refor o em Intelig ncia Artificial com Deep Learning	3
Reconhecimento de imagens e modifica es em arquiteturas	3