Ferramenta para Demonstrar Funcionamento de alguns Algoritmos de Modelagem de Tópicos

Isabella C. S. de Araújo [1,2], Marli F.G. Hernandez[1], Marbilia P Sergio[2]
isabellacs_tec@outlook.com, marli@ft.unicamp.br, marbilia.sergio@cti.gov.br

[1]**Faculdade de Tecnologia (FT)**
UNICAMP - da Universidade Estadual de Campinas

[2]**Nu´cleo de qualificação de Software (NQS)**
CTI/MCTI Renato Archer – Campinas/SP

Abstract

In a world where the amount of data exceeds human analytical capacity, techniques are developed to support manual work. Natural Language Processing (NLP), a field within artificial intelligence, aims to enhance communication between humans and computers using natural language. Various methods have been created to address specific problems. This work focused on Topic Modeling methods to classify similar texts through term treatment and usage analysis. Topic modeling offers many benefits but also presents challenges, as it requires knowledge of the model and coding expertise. As a result, this research developed an environment where some topic modeling algorithms can be tested in an intuitive and straightforward manner.

Resumo

Em um mundo com mais dados do que a capacidade humana de analisar, desenvolvem-se técnicas para apoiar o trabalho manual. O Processamento de Linguagem Natural (PNL), uma área da inteligência artificial, busca melhorar a comunicação entre homens e computadores usando linguagem natural. Diversos métodos foram criados para resolver problemas específicos. Este trabalho focou em métodos de Modelagem de Tópicos para classificar textos semelhantes através do tratamento de termos e análise de uso. A modelagem de tópicos oferece muitos benefícios, mas também apresenta desafios, pois é preciso conhecer o modelo e saber codificá-lo. Como resultado, esta pesquisa criou um ambiente onde alguns algoritmos de modelagem de tópicos podem ser testados de maneira intuitiva e simples.

Palavras chave: processamento de linguagem natural; modelagem de tópicos; LSA; LDA; Word2Vec.

1 Introdução

Na era digital, a grande massa de informação disponível traz mudanças dramáticas para milhões de pessoas em termos de como elas coletam, organizam, disseminam, e acessam, assim como a tarefa mais desafiadora, como analisam as informações. Um exemplo claro desse desafio está em empresas que lidam com "*big data"*. De acordo com uma pesquisa publicada na *Harvard Business Review* (BARTON; COURT, 2012), a capacidade de se analisar big data é crucial para impulsionar a forma como as empresas fazem negócios. Entretanto, ainda existem muitos desafios no aproveitamento dessas massas de dados. No devido de análises avançadas onde grandes volumes de dados torna-se comum. Para KIRON, PRENTICE e FERGUSON (2014), o desempenho dessas analises frequentemente não atingem os resultados esperados.

Textos em grande escala representam uma dificuldade para análise humana. E se torna, dentro desse contexto de pesquisa, ainda mais desafiador para pessoas menos familiarizadas com tecnologia de TI. Tecnologias e métodos vêm sendo estudados para auxiliar este trabalho de mineração de conteúdo que pode se tornar inviável devido o tempo e falta de recursos de TI. A área de conhecimento Processamento Natural da Linguagem (PNL) surge na busca de encontrar soluções para o processamento de grandes dados textuais. Como campo de estudo científico, a PNL assimila a ciência da computação, a linguística e a matemática com o objetivo principal de traduzir a linguagem humana (ou natural) em comandos que podem ser executados por computadores (KANG ZHAO CAI; LIU, 2020). Entre os diversos métodos existentes para cada expectativa de resultados gerados por uma análise com essas ferramentas, destaca-se a modelagem de tópicos.

Os diversos algoritmos de modelagem de tópicos fornecem técnicas de múltiplas perspectivas para encontrar semânticas ocultas em coleções de documentos e agrupá-los em temas como tópicos. Estes recursos são importantes facilitadores da análise textual, entretanto, sua utilização ainda necessita de um solido conhecimento em programação e entendimento das ferramentas e do funcionamento dos algoritmos envolvidos dando consistência nos resultados a serem obtidos (KHERWA; BANSAL, 2019).

O trabalho desta pesquisa está focado na criação de uma ferramenta para qual seja possível demonstrar o uso de alguns algoritmos de modelagem de tópicos. Assim, será estudado uma forma de disponibilizar uma plataforma web que permita que qualquer usuário insira um conjunto de textos e escolha processar entre os algoritmos implementados. Esse documento segue, com as seções de fundamentação teórica, descreve o ambiente de desenvolvimento, os resultados obtidos e a conclusão deste trabalho.

2 Fundamentação Teórica

Nessa seção, descrevemos algumas das bases teóricas relevantes para o entendimento deste trabalho.

2.1 Modelagem de Tópicos

A modelagem de tópicos é composta por algoritmos para tratar um grande volume de dados obtendo-se uma dimensão reduzida, revelando conceitos ocultos, características proeminentes e variáveis latentes nos dados fornecidos. Seus algoritmos de múltiplas perspectivas visam encontrar semânticas ocultas na coleção de documentos e agrupá-los em tópicos (SAXTON, 2018).

A redução de dimensionalidade foi inicialmente vista através de uma perspectiva algébrica, mas com o tempo começou a usar métodos probabilísticos. Pôde-se então classificar os modelos de modelagem de tópicos em duas categorias (KHERWA; BANSAL, 2019):

- Modelo probabilístico
- Modelo de tópico não probabilístico (modelo algébrico)

2.2 Gensim Python

Dentre os pacotes de software disponíveis para geração de modelos de tópicos, um dos mais utilizados é a biblioteca de código aberto Gensim. Devido a sua popularidade e algoritmos disponíveis, ela se destaca pela disponibilização de algoritmos de modelagem no projeto (SAXTON, 2018). Os algoritmos de modelagem de tópico probabilístico são classificados em dois grupos: os supervisionados e os não supervisionados, ou seja, este último trata de algoritmos com padrões semânticos encontrados estatisticamente em textos sem quaisquer metadados prévio, só precisando de um corpus de documentos de texto simples (REHUREK, 2024). Alguns algoritmos disponíveis no GENSIM e implementados nesta pesquisa são: *Latent Dirichlet Allocation* (LDA) um modelo probabilístico, *Latent Semantic Indexing* (LSI ou LSA) um modelo não probabilístico e o *Word2Vec* que não é um modelo de tópicos, mas sim um método que transforma palavras em vetores numéricos passível de ser utilizado junto com modelos de tópicos melhorando a precisão de seus resultados. (JOHNSON; MURTY; NAVAKANTH, 2024).

3 Desenvolvimento do ambiente

Nessa seção, será apresentado o ambiente utilizado para o desenvolvimento da plataforma. Foram utilizas linguagens de programação web, além de Python e SQL. E a plataforma, em sua primeira versão, já permite testar três dos algoritmos. Esses três algoritmos trazem resultados distintos possíveis no processamento de texto. A tela inicial do site traz instruções de uso e brevemente explicação sobre os três algoritmos apoiando na compreensão dos algoritmos e dos resultados fornecidos.

A figura 1 - Tela de inserção de base de textos exibe a tela onde o usuário pode fazer o upload do arquivo que deseja submeter aos algoritmos.

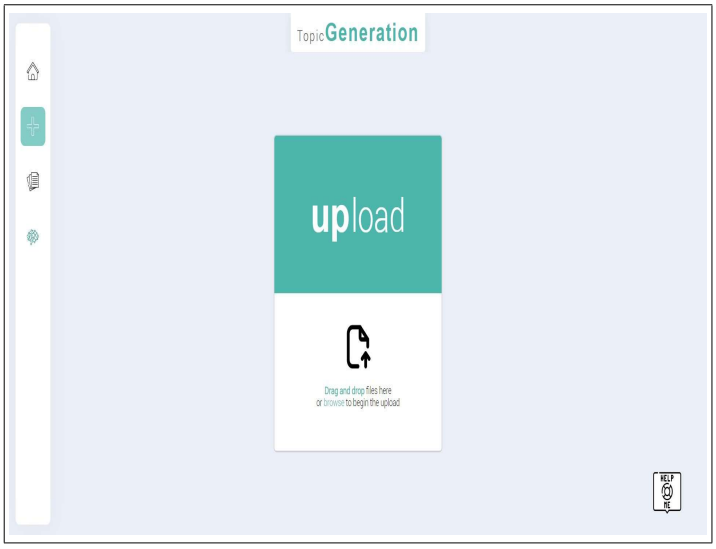

Figura 1: Tela de inserção de base de textos.

Figura 2: Tela dos modelos apresenta os três modelos disponíveis em formato de cartão. Ao clicar em um deles, a tela na qual é possível inserir os parâmetros necessários para executar o respectivo modelo é exibida.

Figura 2: Tela dos modelos.

Na figura 3 apresenta algumas distinções de parâmetros para cada algoritmo. Para o LSA e LDA, o usuário tem a opção de executar a função "*Clean Corpus",* etapa de pré-processamento[1] onde são remover dos textos símbolos e caracteres especiais, são retirar palavras com menos de duas letras e ocorre a "*Lemmatize Corpus*"[2], transforma palavras em sua forma de radial, igualando termos similares.

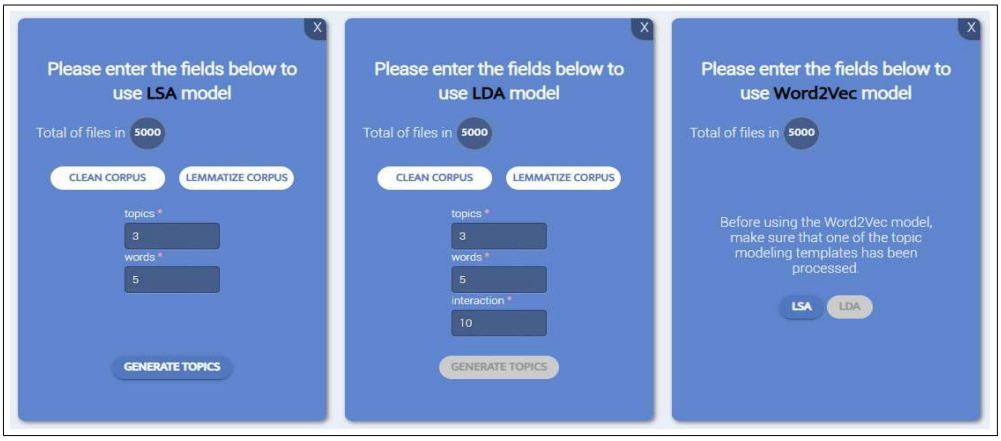

Figura 3: Execução algoritmos.

Os modelos LSA e LDA oferecem técnicas diferentes de tratamento dos termos dando resultados distintos utilizando uma mesma base de dados inserida. Os modelos pedem dois parâmetros obrigatórios: a quantidade de tópicos e a quantidade de palavras que cada tópico conterá. O LDA possui um parâmetro obrigatório adicional, que é o número de iterações que o algoritmo realizará - o número de vezes que o algoritmo passa pelo conjunto de dados para ajustar o modelo (VAYANSKY; KUMAR, 2020).

Já o Word2Vec é um algoritmo que possui vários usos, mais aqui, para contribuir com a proposta da plataforma, ele compara à similaridade entre as palavras dos tópicos gerados pelos modelos de tópicos, permitindo que o usuário tenha uma

[1] Conjunto de técnicas para melhorar a qualidade dos dados para serem analisados e modelados.

[2] Representar as palavras através do infinitivo dos verbos e masculino singular dos substantivos e adjetivos, a fim de simplificar a palavra.

percepção mais técnica se as palavras selecionadas para um tópico fazem ou não sentido ao estarem agrupadas.

Os resultados fornecidos pelos modelos são disponibilizados numa planilha, hoje extensão do *Excel,* exportada automaticamente para a máquina do usuário.

4 Resultados

Os resultados obtidos no uso dos três algoritmos implementados na plataforma são apresentados a seguir. Para sua execução, foi utilizado o conjunto de dados chamado "*Global News Articles*" adquirido por meio do *Kaggle*[1]. De acordo com o colaborador que disponibilizou os dados, o conjunto de dados contém uma seleção com curadoria de artigos de notícias cobrindo uma ampla gama de tópicos, incluindo política, negócios, tecnologia, saúde e muito mais. Para a análise, foram selecionadas 5000 linhas do conjunto total de dados.

Na Tabela 1, apresenta o resultado de um dos tópicos gerados pelos modelos LSA e LDA. Observa-se a diferente combinação de palavras escolhidas representado m tópico. Os tópicos permitem uma percepção distinta sobre a mesma base de dados. O peso associado a cada palavra também possui significados distintos para cada modelo. No modelo LSA, o peso indica a contribuição da palavra para o tópico. Já para o LDA, o peso representa a probabilidade daquela palavra estar associada ao seu tópico. Em ambos os modelos, quanto mais próximo o peso for de 1, mais próximo do "ideal" a palavra está do tópico que ela está contida.

Tabela 1: Comparação de palavras e pesos 4ntre os dois modelos de tópico

LSA		LDA	
Palavra	Peso	Palavra	Peso
char	0,921396818	char	0,046576679
year	0,141222843	new	0,023544982
get	0,070319993	apple	0,018656868
new	0,068306284	app	0,010947695
say	0,06533182	macbook	0,008873228

A tabela 2 apresenta o grau de similaridade entre as palavras quando o algoritmo Word2Vec é utilizado antes dos algoritmos LDA e LSA. Pode-se observar que o grau de similaridade entre os termos de um tópico torna-se maiores apontando para a recomendação de sua utilização no processamento do texto. O Word2Vec oferece uma percepção adicional ao usuário ao permitir a comparação de similaridade entre as palavras dos tópicos, mostrando quão semelhantes essas palavras são dentro do conjunto de textos treinados.

[1] Plataforma que abriga competições de aprendizado de máquina, hosts de datasets e fornece um ambiente para a criação de notebooks.

Tabela 2: Comparação de similaridade palavras do tópico com Word2Vec

LSA			LDA		
Tipo Palavra	Palavra similar	Similaridade	Tipo Palavra	Palavra similar	Similaridade
char	year	0,974728346	char	new	0,997596025
char	get	0,927948236	char	apple	0,995978892
char	new	0,997944534	char	app	0,995330393
char	say	0,998720109	char	macbook	0,995838583

Na comparação da primeira palavra entre as outras de seu tópico, nos dois modelos, tem-se que valores mais próximos de 1 indicam uma alta similaridade entre as palavras, enquanto valores próximos de -1 indicam quão opostas as palavras são.

5 Conclusão

Este trabalho gerou o desenvolvimento de uma plataforma que, de maneira simples, permite a execução de três algoritmos voltados para a modelagem de tópicos. O usuário pode inserir um conjunto de texto de sua escolha e aplicar sobre eles um ou mais dos códigos disponíveis: LSA, LDA e Word2Vec. Os resultados dos algoritmos são disponibilizados em planilha, exportados automaticamente para a máquina do usuário. O resultado é um conjunto de tópicos composto por um conjunto de palavras baseados nos textos processados pela plataforma. A analise destes tópicos pode contribuir para que o usuário compreenda do que se trata a modelagem de tópicos pode proporcionar. Além disso, a plataforma permite observar a proximidade ou distância entre as palavras de um mesmo tópico utilizando o algoritmo Word2Vec.

O objetivo do trabalho é atendido ao permitir que, de forma simples, o usuário possa testar o funcionamento dos dois principais algoritmos de modelagem de tópico.

Como trabalhos futuros pretende-se incrementar a plataforma disponibilizando novos algoritmos. Por exemplo, modelos de classificação e de análise de sentimento enriquecendo o entendimento dos algoritmos PNL e implementar métricas de avaliação na plataforma que possibilitem que o usuário mensure a qualidade dos resultados dos modelos. Um exemplo dessa ampliação poderá ser pela inclusão de modelos que medem as características do texto (legibilidade ou subjetividade), como a classificação da escrita por Tom na análise de sentimentos (KANG ZHAO CAI; LIU, 2020).

Referências

BARTON, D.; COURT, D. Making advanced analytics work for you. *Harvard business review*, v. 90, n. 10, p. 78–83, 2012.

JOHNSON, S. J.; MURTY, M. R.; NAVAKANTH, I. A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia, Tools and Applications*, v. 83, n. 13, p. 37979 – 38007, 2024. Cited by: Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2. 0-

85173083987&doi=10.1007\%2fs11042-023-17007-z&partnerID=40&md5=c676adad061423bf90be9eab42a95135⟩.

KANG ZHAO CAI, C.-W. T. Q. H. Y.; LIU, H. Natural language processing (nlp) in management research: A literature review. *Journal of Management Analytics*, Taylor & Francis, v. 7, n. 2, p. 139–172, 2020. Disponível em: ⟨https://doi.org/10.1080/23270012.2020.1756939⟩.

KHERWA, P.; BANSAL, P. Topic modeling: A comprehensive review. EAI Endorsed Transactions on Scalable Information Systems, EAI, v. 7, n. 24, 7 2019.

KIRON, D.; PRENTICE, P. K.; FERGUSON, R. B. The analytics mandate. MIT. Sloan Management Review, v. 55, n. 4, p. 1–25, Summer 2014 2014. Copyright © Massachusetts Institute of Technology, 2014. All rights reserved; Artigo principal do documento -; última atualização em - 2023-11-29; CODEN´

SMRVAO; SubjectsTermNotLitGenreText - United States–US. Disponível em: ⟨https://www.proquest.com/scholarly-journals/analytics-mandate/docview/1543709856/se-2⟩.

REHUREK, R. Gensim: Topic modelling for humans. 2024. Acessado em: 2024-06-26. Disponível em: ⟨https://radimrehurek.com/gensim/⟩.

SAXTON, M. A gentle introduction to topic modeling using python. Theological Librarianship, v. 11, p. 18–27, 04 2018.

VAYANSKY, I.; KUMAR, S. A. A review of topic modeling methods. Information Systems, v. 94, p. 101582, 2020. ISSN 0306-4379. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0306437920300703⟩.