ANÁLISE VETORIAL SEMÂNTICA DE TEXTO: UMA APLICAÇÃO DA ÁLGEBRA LINEAR E TF-IDF NA MENSURAÇÃO DE SIMILARIDADE DE NOTÍCIAS

Henrique Rodrigues de Freitas

Ciência de Dados, Fatec Rubens Lara

henrique.freitas01@fatec.sp.gov.br

Resumo

O presente trabalho demonstra a aplicação de conceitos de Álgebra Linear, especificamente a Similaridade de Cosseno, na área de Processamento de Linguagem Natural (PLN) para quantificar o grau de semelhança temática entre documentos textuais. A metodologia empregou o algoritmo Term Frequency-Inverse Document Frequency (TF-IDF) em Python para transformar um corpus de notícias jornalísticas em um vetor numérico em um espaço de alta dimensionalidade. Os resultados validam que pares de documentos com altíssima similaridade (cosseno de 0.9608, ângulo de 16.09°) compartilham um vocabulário temático intenso e vetores quase colineares. Em contraste, o par menos semelhante apresentou baixa colinearidade ($\cos(\theta) = 0.2667$ ou $\theta = 74.53°$), refletindo a divergência de temas. Este estudo confirma a eficácia da Similaridade de Cosseno como uma métrica robusta para análise semântica em grande escala, demonstrando a relevância da geometria vetorial para a recuperação e classificação de informação.

Palavras-chave: álgebra linear; similaridade de cosseno; TF-IDF; análise vetorial; PLN.

ABSTRACT This article aims to demonstrate the practical application of core Linear

Algebra concepts, specifically Cosine Similarity, in the field of Natural Language Processing (NLP) to quantify the thematic similarity between text documents. The methodology employed the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm in Python to transform a corpus of news articles into numerical vectors in a high-dimensional space. Results validate that document pairs with very high similarity (cosine of 0.9608, angle of 16.09°) share an intense thematic vocabulary and are nearly collinear. In contrast, the least similar pair showed low collinearity ($\cos(\theta) = 0.2667$ or $\theta = 74.53°$), reflecting thematic divergence. This study confirms the effectiveness of Cosine Similarity as a robust metric for semantic analysis, demonstrating the relevance of vector geometry for large-scale information retrieval and classification.

Keywords: linear algebra; cosine similarity; TF-IDF; vector analysis; NLP.

1 Introdução

A gestão e análise de dados textuais são desafios centrais na Ciência da Computação moderna. A Álgebra Linear oferece uma solução robusta por meio do *Vector Space Model* (VSM), onde a linguagem natural é traduzida em uma representação geométrica. Neste modelo, cada documento é tratado como um vetor em um espaço euclidiano M-dimensional.

O propósito central deste trabalho é demonstrar o rigor e a eficácia dessa conversão, focando no cálculo do Cosseno do Ângulo (θ) entre os vetores de documentos. Esta métrica, conhecida como Similaridade de Cosseno, permite avaliar o quão semelhantes são as orientações dos vetores no espaço, fornecendo uma quantificação direta da semelhança temática entre as notícias.

2 Referencial Teórico

2.1 Representação Vetorial e TF-IDF

O algoritmo TF-IDF (Term Frequency-Inverse Document Frequency) é o vetorizador escolhido para este estudo. Ele constrói a Matriz Documento-Termo onde cada dimensão do vetor de um documento representa a importância de uma palavra (termo).

A ponderação TF-IDF é calculada como o produto da frequência do termo no documento (TF) e o inverso da frequência do documento no corpus (IDF):

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

Onde o IDF(t) penaliza termos comuns que ocorrem em muitos documentos do corpus (como 'e', 'de', 'a'), garantindo que apenas termos tematicamente relevantes contribuam significativamente para a direção do vetor.

2.2 Análise da Similaridade de Cosseno

A Similaridade de Cosseno é uma métrica geométrica que mede o ângulo entre dois vetores. Sua principal vantagem é a **independência da magnitude** (comprimento) do vetor, sendo ideal para comparar documentos de tamanhos desiguais.

O cosseno do ângulo (θ) entre dois vetores $\vec{A}$ (Documento D_i) e $\vec{B}$ (Documento D_j) é dado pelo Produto Escalar normalizado:

$$\text{Similaridade}(\vec{A}, \vec{B}) = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|}$$

O resultado varia de 0 (vetores ortogonais, temas distintos) a 1 (vetores colineares, temas idênticos). A conversão para o ângulo em graus fornece uma métrica intuitiva da separação temática.

3 Metodologia

3.1 Descrição do Corpus e Pré-processamento

O *corpus* utilizado é o dataset de notícias brasileiras disponibilizado na plataforma Kaggle (`diogocaliman/notcias-publicadas-no-brasil`). O *dataset* contém notícias classificadas por assunto, como 'esportes', 'economia' e 'política'.

O pré-processamento incluiu a **remoção de stop words** da língua portuguesa para assegurar que o foco vetorial estivesse nos termos temáticos.

3.2 Amostragem e Mitigação de Vício (Data Viciado)

O problema do vício de dados na amostragem foi identificado durante a análise exploratória: a aplicação de uma amostragem simples ('head(40)'), sobre o dataset se-

quencialmente ordenado, resultou em uma amostra inicial desbalanceada, na qual predominavam as notícias de 'economia' e 'esportes' e poucas de 'política', comprometendo a prova geométrica.

Para solucionar este vício e garantir que o PCA (Análise de Componentes Principais) e a análise de ortogonalidade fossem estatisticamente válidos e visualmente demonstráveis, foi aplicada a técnica de **Amostragem Aleatória Estratificada**. Esta técnica selecionou um número igual e aleatório ($N = 14$) de documentos de cada uma das três categorias de interesse ('economia', 'esportes', 'política'), garantindo que os vetores representassem uniformemente todos os domínios temáticos.

3.3 Vetorização e Cálculo

O processamento em Python (utilizando 'pandas' e 'scikit-learn') seguiu as etapas:

1. **Vetorização:** Geração da matriz
2. **Similaridade:** Cálculo da matriz de Similaridade de Cosseno entre todos os pares de vetores.
3. **Ângulos:** Conversão dos valores de $\cos(\theta)$ para θ em graus.

4 Resultado e Discussão

A análise de similaridade identificou os pares com o ângulo mais agudo (máxima similaridade) e o ângulo mais obtuso (mínima similaridade), comprovando o modelo vetorial.

4.1 Análise da Colinearidade: O Ângulo de $16.09°$

O par de documentos **D6 e D31** apresentou a maior similaridade no *corpus*:

- **Similaridade de Cosseno ($\cos(\theta)$):** 0.9608
- **Ângulo Vetorial (θ):** $16.09°$

O ângulo de $16.09°$ demonstra que os vetores são **quase colineares**, indicando que eles abordam o mesmo domínio temático de forma quase idêntica. A razão direta para este alto alinhamento vetorial é a maximização do Produto Escalar ($\vec{A} \cdot \vec{B}$), causada pela sobreposição de termos de alto peso TF-IDF.

Tabela 1: Termos de Alto Peso Comuns aos Documentos D6 e D31

Termo Comum	**Peso D6**	**Peso D31**	**Justificativa Temática**
processos	0.2766	0.2592	Alta Frequência em ambos, tema central.
planos	0.2700	0.2530	Alta Frequência em ambos, tema central.
stf	0.1800	0.2108	Termo específico do Judiciário e Política.
poupadores	0.1955	0.1833	Termo específico do domínio Econômico/Legal.
tribunais	0.1800	0.1687	Domínio Legal.
fux	0.1467	0.1374	Domínio Político/Judiciário (nome de Ministro).

Conforme a Tabela 1, a forte presença de termos como 'processos', 'planos' e 'stf' confirma que ambos os documentos tratam da mesma pauta (a judicialização de planos econômicos). A contribuição destes termos com pesos elevados nos vetores maximiza o Produto Escalar e, consequentemente, alinha a direção dos vetores, validando a Similaridade de Cosseno como um indicador semântico preciso.

4.2 Análise da Ortogonalidade: Separação Vetorial

Para a amostra utilizada, o par de documentos **D0 e D13** foi identificado como o par com a **menor colinearidade**:

- **Similaridade de Cosseno ($\cos(\theta)$):** 0.2667
- **Ângulo Vetorial (θ):** 74.53°

O ângulo de 74.53°, próximo da ortogonalidade (90°), demonstra que a intersecção de termos de alto peso entre os vetores é baixa. Isso ocorre quando os documentos abordam temas significativamente diferentes, resultando em um Produto Escalar baixo, o que valida a capacidade da Similaridade de Cosseno em medir a separação temática.

4.3 Representação Geométrica e Agrupamento (Clustering)

A Figura 1 demonstra a visualização dos vetores após a redução de dimensionalidade para duas componentes principais (PCA). Esta técnica, fundamental da Álgebra Linear, projeta o espaço vetorial de alta dimensionalidade em um plano 2D, preservando a máxima variância dos dados.

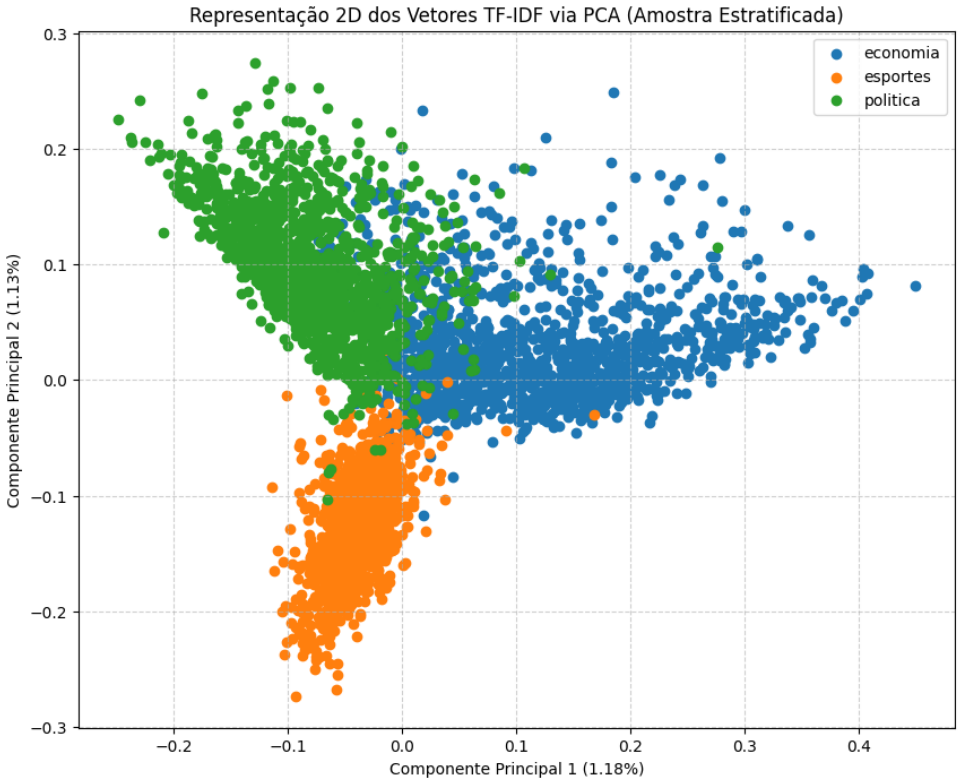

Figura 1: Representação 2D dos Vetores TF-IDF via PCA (Análise por Componentes Principais com Amostra Estratificada).

O gráfico PCA exibe a formação de *clusters* visivelmente distintos (Economia, Esportes e Política), corroborando o poder da métrica de cosseno. Os vetores de conteúdos temáticos idênticos se agrupam, reforçando que a direção vetorial é a chave para a similaridade, enquanto os grupos distintos (ortogonais) se distanciam no plano, comprovando a eficácia da transformação do texto em geometria vetorial para fins de classificação.

5 Considerações Finais

O trabalho atingiu seu objetivo de demonstrar a aplicação direta de conceitos de Álgebra Linear, em particular o Produto Escalar normalizado (Similaridade de Cosseno), para a

quantificação da semelhança semântica de textos. A análise detalhada da colinearidade entre D6 e D31, justificada pela sobreposição de termos TF-IDF de alto peso, provou que a direção vetorial no espaço $\mathbb{R}^M$ é um substituto robusto para a essência temática do texto. O resultado do ângulo de 74.53° para o par menos colinear demonstra o rigor da métrica na medição da divergência. Este método é essencial para as bases de algoritmos de sistemas de recomendação, motores de busca e *clustering* automático de documentos.

6 Referências

1. CALIMAN, D. *Notícias publicadas no Brasil*. Dataset disponível em: `https://www.kaggle.com/datasets/diogocaliman/notcias-publicadas-no-brasil`. Acesso em: 21/10/2025.

2. JURAFSKY, Daniel; MARTIN, James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 3. ed. Pearson, 2021.

3. LAY, David C.; LAY, Steven R.; MCDONALD, Judi J. *Álgebra Linear e suas Aplicações*. 5. ed. LTC, 2018.

4. PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825-2830, 2011.