

# **Relatório – Sistema de Recomendação Utilizando Similaridade do Cosseno**

## **Introdução**

Este relatório apresenta o desenvolvimento de um sistema de recomendação textual fundamentado na similaridade do cosseno, um método amplamente explorado nas áreas de Álgebra Linear e Processamento de Linguagem Natural (PLN). A técnica consiste em transformar textos em vetores numéricos e calcular o cosseno do ângulo entre eles, permitindo medir sua proximidade semântica. Diferente de métodos que analisam apenas a frequência de palavras comuns, a similaridade do cosseno permite identificar afinidades mesmo quando diferentes termos são utilizados para expressar ideias semelhantes.

Ao longo deste projeto, foi utilizado um conjunto de dados contendo artigos com seus respectivos títulos e conteúdos, em língua inglesa. A implementação incluiu etapas fundamentais como pré-processamento textual, vetorização com TF-IDF e cálculo da similaridade entre os textos, possibilitando ao usuário receber recomendações com base no conteúdo de um artigo previamente escolhido.

## **Objetivo**

O propósito deste projeto foi desenvolver um sistema de recomendação textual com base na similaridade do cosseno, um conceito estudado em disciplinas de Álgebra Linear e amplamente aplicado no Processamento de Linguagem Natural (PLN). A técnica permite mensurar o quanto dois textos se assemelham em termos de conteúdo, transformando-os em vetores e comparando seus ângulos.

Essa abordagem foca na direção dos vetores, ignorando sua magnitude, o que a torna ideal para análise de textos com tamanhos diferentes. A fórmula matemática utilizada foi:

$$\text{Similaridade (A, B)} = (\mathbf{A} \cdot \mathbf{B}) / (||\mathbf{A}|| \times ||\mathbf{B}||)$$

Valores próximos de 1 indicam alta semelhança entre os textos, enquanto valores próximos de 0 sugerem pouca ou nenhuma relação semântica. Esse cálculo é especialmente útil para comparar frases, parágrafos ou documentos, mesmo quando não há repetição literal de palavras.

## **Descrição do Dataset**

O dataset analisado foi composto por artigos com seus respectivos títulos e conteúdos textuais. Como o foco da análise estava no idioma inglês, foi necessário realizar uma etapa de pré-processamento nos dados antes de aplicar a técnica de recomendação. Esse tratamento consistiu em:

- Conversão do texto para letras minúsculas;
- Remoção de pontuação e caracteres especiais;
- Tokenização das frases (quebra em palavras);
- Eliminação de *stopwords*, ou seja, palavras comuns do inglês que não agregam valor semântico (como “de”, “e”, “a”, etc.).

Após esse processo, os textos foram limpos e prontos para serem transformados em vetores com o uso do TF-IDF Vectorizer, uma técnica que calcula a importância de cada palavra no contexto do documento.

## **Desenvolvimento do Algoritmo**

O algoritmo foi construído para receber como entrada o título de um artigo. A partir dele, o sistema localiza o texto correspondente no dataset e, usando o modelo vetorial de TF-IDF, compara essa descrição com as demais, aplicando a fórmula da similaridade do cosseno.

Primeiro é necessário que as bibliotecas necessárias estejam instaladas, como o NLTK para processamento de linguagem natural e o Scikit-learn para análise de dados e machine learning. Depois, importa outras ferramentas importantes, como Pandas, NumPy e funções para transformar textos em números e medir similaridade entre eles.

Alguns recursos do NLTK, como punkt e stopwords, também são baixados para ajudar na limpeza dos textos, removendo palavras irrelevantes e separando as frases corretamente.

Depois precisa fazer o upload do arquivo CSV com artigos, e o código usa o Pandas para ler esse conteúdo. Em seguida, uma função chamada preprocess limpa os textos: tudo vira minúsculo, pontuação e palavras comuns são removidas. Linhas com campos vazios são descartadas, e o resultado dessa limpeza é salvo em uma nova coluna.

No final, o usuário digita o título de um artigo e o sistema retorna sugestões de leitura parecidas.

A aplicação do código retorna uma lista dos artigos mais semelhantes, porém e mostrado para o usuário final os 5 mais semelhantes em relação ao que foi digitado, levando em conta que o código não considera o valor digitado a ser assimilado, assim fazendo com que não ocorra o cenário de analisar similaridade com ele mesmo.

O sistema retorna uma lista dos artigos mais semelhantes com base no conteúdo, permitindo assim a geração de recomendações com fundamento textual.

A estrutura geral envolveu:

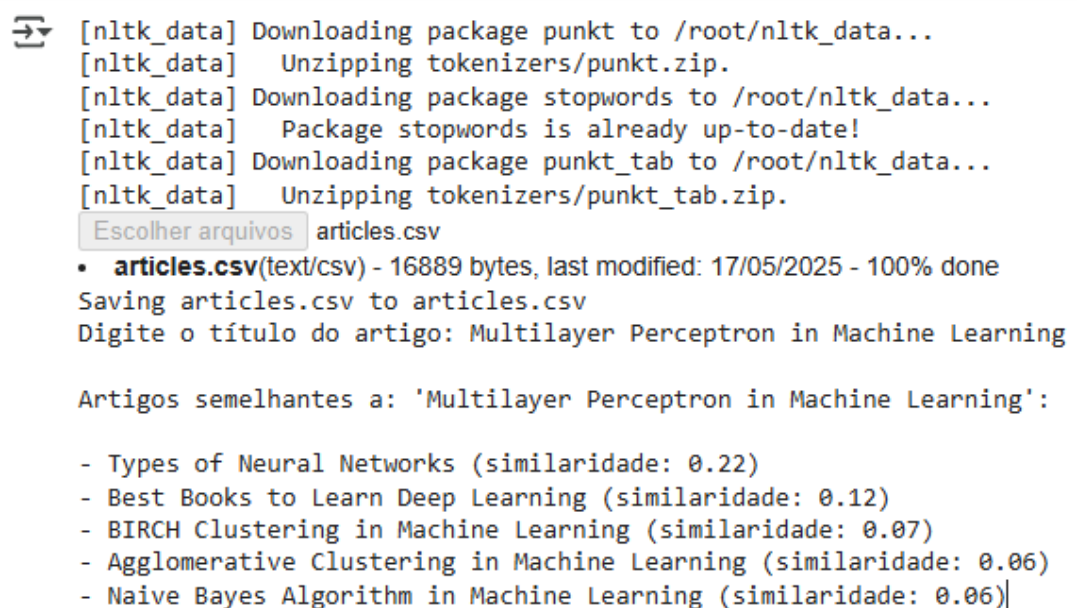
- Leitura do CSV com os artigos;
- Aplicação de pré-processamento textual com NLTK;
- Vetorização do conteúdo;
- Cálculo da similaridade;
- Geração das recomendações com base nos maiores índices de similaridade.

## Resultados

Os testes realizados com o sistema demonstraram que a abordagem baseada no conteúdo do artigo entrega recomendações coerentes, e não apenas a ocorrência de palavras idênticas.

Mesmo que dois artigos tenham vocabulários diferentes, se abordarem temas semelhantes, o modelo consegue identificar essa relação e sugerir-los como relevantes.

por exemplo, o nome de um artigo como *“Multilayer Perceptron in Machine Learning”*, o sistema retornava sugestões com artigos semelhantes:



```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt_tab.zip.
Escolher arquivos articles.csv
• articles.csv(text/csv) - 16889 bytes, last modified: 17/05/2025 - 100% done
Saving articles.csv to articles.csv
Digite o título do artigo: Multilayer Perceptron in Machine Learning

Artigos semelhantes a: 'Multilayer Perceptron in Machine Learning':

- Types of Neural Networks (similaridade: 0.22)
- Best Books to Learn Deep Learning (similaridade: 0.12)
- BIRCH Clustering in Machine Learning (similaridade: 0.07)
- Agglomerative Clustering in Machine Learning (similaridade: 0.06)
- Naive Bayes Algorithm in Machine Learning (similaridade: 0.06)
```

## **Conclusão**

Este trabalho proporcionou uma oportunidade prática de aplicar conceitos importantes da área de Ciência de Dados, como o pré-processamento textual, vetorização e cálculo de similaridade do cosseno. A utilização da similaridade do cosseno mostrou-se eficaz na construção de um sistema de recomendação textual simples e funcional.

Além disso, a experiência reforçou a importância da preparação dos dados e da escolha das representações adequadas para que os algoritmos produzam resultados significativos. Mesmo sendo um modelo inicial, ele abre espaço para aprimoramentos futuros.