

Sistema de Recomendação de Músicas Usando Similaridade do Cosseno

Autor: Kayky Martins

Maio de 2025

Objetivo

O objetivo deste trabalho foi desenvolver um algoritmo de recomendação baseado no conceito de similaridade do cosseno, abordado nas aulas de Álgebra Linear. Esse conceito é amplamente utilizado no processamento de texto para comparar documentos ou frases representados por vetores, como vetores de frequência de palavras, por exemplo.

A similaridade do cosseno permite medir o grau de semelhança entre dois textos, considerando apenas a direção dos vetores e desconsiderando seu tamanho (magnitude). A fórmula para calcular a similaridade do cosseno entre dois vetores A e B é:

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

onde:

- $A \cdot B$ é o produto escalar entre os vetores;
- $\|A\|$ e $\|B\|$ são as normas (magnitudes) dos vetores.

Se a similaridade do cosseno for 1, os textos são iguais ou muito semelhantes; se for próxima de 0, os textos são bastante diferentes. Quanto maior o valor (mais próximo de 1), maior a semelhança. Quanto menor (mais próximo de 0), menor a semelhança. Essa métrica permite comparar dois textos de forma matemática e objetiva, mesmo que eles não sejam idênticos palavra por palavra.

Além disso, ao calcular o arco cosseno (arccos) dessa razão, obtém-se o ângulo entre os vetores que representam os textos. É importante destacar que o valor do arccos retornado está em radianos — para convertê-lo em graus, basta multiplicar o resultado por $\frac{180}{\pi}$. Esse ângulo fornece uma outra forma intuitiva de medir a similaridade, onde ângulos menores indicam maior proximidade entre os textos.

Tratamento dos Dados

O dataset utilizado neste trabalho, intitulado **Letras de Bossa Nova** e retirado da plataforma Kaggle, reúne letras de músicas do gênero Bossa Nova, juntamente com informações como nome da música, artista, letra da música, compositores e idioma da música (português, inglês, espanhol, italiano, entre outros).

Como o foco da análise está na língua portuguesa, foi feito um tratamento inicial para remover todas as músicas que não estavam nesse idioma. O dataset original continha 6.107 músicas, mas, após a filtragem, passou a contar com 4.118 músicas em português. Esse processo foi essencial para garantir que os resultados do algoritmo de recomendação fossem mais consistentes e relevantes.

Algoritmos Desenvolvidos

Com o intuito de investigar diferentes metodologias para sistemas de recomendação, foram implementados dois algoritmos distintos, ambos na linguagem Python e desenvolvidos no Visual Studio Code:

- **Recomendação fundamentada no título da música:** nessa abordagem, os títulos das canções foram transformados em vetores e empregados como base para a mensuração da similaridade entre as músicas. Essa estratégia visa identificar faixas com títulos semelhantes, o que pode refletir proximidade temática ou estilística. O pré-processamento envolveu a limpeza dos títulos, remoção de acentuação, pontuação e stopwords, além da criação de uma matriz binária bag-of-words. Em seguida, vetores de entrada são comparados com os vetores das músicas no dataset usando a similaridade do cosseno.
- **Recomendação fundamentada na letra da música:** neste segundo método, utilizam-se as letras completas das músicas para o cálculo da similaridade. Tal procedimento tende a gerar resultados mais consistentes, pois incorpora o conteúdo lírico integral, permitindo recomendações mais precisas em relação a temas, sentimentos e vocabulário presentes nas composições.

Ambos os algoritmos aplicam técnicas de vetorização textual para converter os dados textuais em representações numéricas, possibilitando a aplicação do cálculo da similaridade do cosseno entre os diferentes elementos.

Resultados

Para validar os algoritmos implementados, foram realizados testes práticos com a inserção de trechos reais, permitindo observar o comportamento de cada modelo. As duas imagens

a seguir ilustram os resultados produzidos por ambas as abordagens de recomendação: uma baseada no nome da música e outra fundamentada no conteúdo lírico.

```
Digite o nome da música para recomendar: girassol
O Girassol - Vinicius de Moraes
Catavento e Girassol - Sérgio Mendes
Tomara - Alaide Costa
Tempo Calado - Alaide Costa
Quem Sou Eu? - Alaide Costa
Diariamente - Alaide Costa
Coração - Alaide Costa
Absinto - Alaide Costa
Amigo Amado - Alaide Costa
Onde Está Você - Alaide Costa
```

Figura 1: Resultados do algoritmo baseado no título das músicas(Executado no google colab)

```
Digite a letra ou parte da letra da música para recomendar: girassol
Músicas recomendadas:
Título: O Girassol
Artista: Vinicius de Moraes
Letra (trecho):
    Sempre que o sol Pinta de anil Todo o céu O girassol Fica um gentil Carrossel Roda, roda, roda Carrossel Roda, rod
a, rod...

Título: Catavento e Girassol
Artista: Sérgio Mendes
Letra (trecho):
    Meu catavento tem dentro O que há do lado de fora do teu girassol Entre o escancarado e o contido Eu te pedi sustenido E você
riu ...

Título: Cântico
Artista: Vinicius de Moraes
Letra (trecho):
    Não, tu não és um sonho, és a existência Tens carne, tens fadiga e tens pudor No calmo peito teu. Tu és a estrela Sem nome,
és a...

Título: Absinto
Artista: Alaide Costa
Letra (trecho):
    Hoje a coisa mais firme que creio É que eu te odeio E é esse o ódio que eu mais choro Eu te adoro Eu te adoro Eu bebo essas
água...

Título: Tempo Calado
Artista: Alaide Costa
Letra (trecho):
    Ah! Se o esquecer me for negado Esta paixão posta de lado Irá me perseguir E tanto faz Que o tempo passe Ou que distantes N
ossos...
```

Figura 2: Resultados do algoritmo baseado na letra das músicas(Executado no google colab)

Na Figura 1, visualizamos os retornos obtidos pelo modelo que opera a partir do título das músicas, recomendando os 10 títulos mais próximos aos inseridos.

Já na Figura 2, observa-se o funcionamento do algoritmo que analisa as letras completas das músicas. Por considerar o texto integral, essa abordagem consegue entregar recomendações mais ricas e coerentes com o conteúdo emocional, temático ou linguístico da entrada fornecida.

Conclusão

Com a realização deste projeto, foi possível aplicar, de maneira prática, alguns dos conceitos fundamentais da Ciência de Dados, como o pré-processamento de dados, a vetorização

de textos e a medição de similaridade entre informações. Utilizando a técnica de similaridade do cosseno, foram desenvolvidos dois algoritmos de recomendação: um baseado no título da música e outro no conteúdo da letra. Essa experiência mostrou como diferentes formas de representação textual impactam os resultados, reforçando a importância da escolha adequada das variáveis na análise de dados.

Além disso, o projeto ajudou a compreender o fluxo de um processo típico em Ciência de Dados, desde a coleta e tratamento do dataset até a análise dos resultados. Mesmo sendo um trabalho inicial e com limitações, os resultados obtidos reforçaram o potencial do uso de técnicas de Processamento de Linguagem Natural (PLN) para gerar valor a partir de dados textuais. Assim, esta atividade contribuiu para o desenvolvimento das habilidades analíticas e técnicas essenciais para quem está começando na área de Ciência de Dados.