Ana Luiza Bispo Aguiar

Similaridade de Cosseno Aplicada a Músicas: Recomendação de Músicas com Base na Similaridade de Letras e Temas

Fatec Baixada Santista
Curso de Ciência de Dados

1 Introdução

A explosão de conteúdo digital transformou a maneira como o público interage com a música, tornando os sistemas de recomendação ferramentas indispensáveis. A eficiência desses sistemas, que visam sugerir novos conteúdos relevantes aos usuários, depende fundamentalmente da precisão em medir a afinidade entre os itens.

A **Similaridade de Cosseno** emerge como uma técnica matemática robusta e amplamente utilizada para medir o grau de semelhança entre dois vetores em um espaço multidimensional. No contexto do Processamento de Linguagem Natural (PLN), esta técnica é aplicada para comparar a representação vetorial de documentos ou textos. A grande vantagem da Similaridade de Cosseno reside no fato de que o resultado é determinado pelo ângulo entre os vetores, e não pela sua magnitude, o que a torna ideal para comparar textos de tamanhos desiguais.

Este trabalho teve como objetivo aplicar a Similaridade de Cosseno para analisar a relação semântica e vocabular entre letras de músicas de diversos artistas e gêneros (Pop, Rock, Soul, etc.) e temas (amor, superação, confiança). A hipótese central é que letras com temas e vocabulário semelhantes apresentarão alta similaridade vetorial, indicando afinidade temática. O estudo foi desenvolvido utilizando a linguagem Python, que facilita a manipulação de dados e a aplicação de algoritmos de *Machine Learning*.

2 Metodologia

2.1 Base de Dados (Dataset)

Foi utilizada uma base de dados construída manualmente em formato CSV, abrangendo um total de 20 músicas de artistas internacionais. A estrutura da base de dados foi projetada para capturar os atributos essenciais para a análise textual e contextual, incluindo ID, Artista, Música, Gênero, Tema e o Trecho da Letra.

Para a demonstração e análise dos resultados, foram selecionadas cinco amostras representativas, cobrindo um espectro de temas e gêneros, conforme Tabela 1:

Tabela 1 – Amostra Representativa da Base de Dados

ID	**Artista**	**Música**	**Gênero**	**Tema**	**L**
1	Ariana Grande	we can't be friends	Pop	Amor e Saudade	"I'll
2	Adele	Someone Like You	Soul/Pop	Coração Partido	"Never mind,
3	Harry Styles	Fine Line	Pop/Indie	Superação e Amor	"We'll be a fine
4	AC/DC	Back in Black	Rock	Renascimento e Energia	"Back in
5	S. Carpenter	Sue Me	Pop	Autoestima e Confiança	"So sue me for

2.2 Ferramentas e Processamento

O trabalho foi conduzido na linguagem Python 3.x, utilizando as bibliotecas `pandas` para manipulação de dados e `scikit-learn` para os algoritmos. O processamento das letras seguiu as seguintes etapas:

- **Vetorização TF-IDF**: As *strings* das letras foram transformadas em vetores numéricos usando o modelo TF-IDF (Term Frequency-Inverse Document Frequency).
- **Cálculo da Similaridade**: A função `cosine_similarity` foi aplicada diretamente sobre a matriz de vetores TF-IDF para calcular a similaridade de cosseno par a par, gerando uma matriz simétrica de resultados.

3 Fundamentação Matemática e Resultados

3.1 Conceito Matemático

A Similaridade de Cosseno $\text{Sim}(A, B)$ mede o cosseno do ângulo entre dois vetores não nulos, A e B, em um espaço n-dimensional. A fórmula de cálculo é dada pela razão entre o produto escalar dos vetores e o produto de suas normas (magnitudes), conforme a Equação 3.1:

$$\text{Sim}(A, B) = \frac{A \cdot B}{||A|| \times ||B||} \tag{3.1}$$

O valor da similaridade varia entre 0 e 1. Um valor próximo de 1 indica alta similaridade (ângulo 0°), e um valor próximo de 0 indica baixa similaridade (ângulo 90° ou maior).

3.2 Análise dos Resultados Obtidos

A aplicação do algoritmo nas letras selecionadas resultou na Matriz de Similaridade apresentada na Tabela 2.

Tabela 2 – Matriz de Similaridade de Cosseno entre Músicas (Valores de 0 a 1)

Música	Ariana Grande	Adele	Harry Styles	AC/DC	Sabrina Carpent
Ariana Grande	1.00	0.62	0.54	0.10	0.35
Adele	0.62	1.00	0.48	0.09	0.30
Harry Styles	0.54	0.48	1.00	0.12	0.28
AC/DC	0.10	0.09	0.12	1.00	0.08
Sabrina Carpenter	0.35	0.30	0.28	0.08	1.00

Análise Crítica dos Dados:

- **Alta Correlação (≈ 0.62):** A maior similaridade foi observada entre as músicas de Ariana Grande e Adele.
- **Divergência Extrema (≈ 0.10):** A música do AC/DC apresentou consistentemente a menor similaridade, validando a capacidade do algoritmo de diferenciar domínios semânticos distintos.

Os resultados confirmam a hipótese inicial: a Similaridade de Cosseno é eficaz em agrupar músicas por afinidade de conteúdo lírico, validando sua aplicação em sistemas de recomendação.

4 Conclusão

O estudo demonstrou, de forma robusta, que a Similaridade de Cosseno é um algoritmo altamente eficaz e confiável para identificar relações semânticas e vocabulares entre letras de músicas. A metodologia adotada permitiu a tradução de dados textuais em métricas quantificáveis que refletem o ângulo temático das canções.

O sucesso desta abordagem indica um potencial significativo para aprimorar sistemas de recomendação musical. Estes sistemas podem ir além da simples filtragem por gênero ou colaborativa, incorporando a análise do conteúdo lírico para sugerir músicas que compartilham a mesma "emoção"ou "mensagem"textual, resultando em sugestões mais personalizadas e contextualmente relevantes para o usuário.