

**FATEC BAIXADA SANTISTA - FACULDADE DE
TECNOLOGIA - RUBENS LARA**

**SIMILARIDADE DE COSSENO APLICADA A
MÚSICAS**

Recomendação de Músicas com Base na Similaridade de Letras
e Temas

Ana Luiza Bispo Aguiar
Curso de Ciência de Dados

Santos, SP

Conteúdo

Resumo	2
1 Introdução	3
2 Metodologia	3
2.1 Base de Dados (Dataset)	3
2.2 Ferramentas e Processamento	3
3 Fundamentação Matemática	4
4 Resultados e Discussão	4
4.1 Análise Crítica dos Dados	4
5 Conclusão	5

Resumo

O presente trabalho investiga a aplicabilidade da técnica de **Similaridade de Cosseno** no domínio de sistemas de recomendação musical. O objetivo central é quantificar o grau de semelhança textual entre letras de músicas de gêneros e temas variados, utilizando o Processamento de Linguagem Natural (PLN). Para tal, foi construído um *dataset* em formato CSV contendo 20 músicas de artistas internacionais, com atributos como Título, Artista, Gênero, Tema e um Trecho da Letra. O processamento foi realizado em Python, utilizando o modelo **TF-IDF** (Term Frequency-Inverse Document Frequency) para vetorializar as letras e calcular a similaridade vetorial. Os resultados obtidos demonstraram que a Similaridade de Cosseno é altamente eficaz na identificação de músicas com conteúdo lírico semelhante, agrupando canções de temas emocionais e românticos com alta correlação (≈ 0.62), e isolando aquelas com vocabulário e estilo distintos, como o Rock enérgico (≈ 0.10). A técnica confirma sua valia como um pilar para a construção de *engines* de recomendação musical personalizados, indo além das categorizações tradicionais baseadas unicamente no gênero ou histórico de escuta.

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	Letra (Trecho)
1	Ariana Grande	we can't be friends	Pop	Amor e Saudade	"I'll pretend I'm fine..."
2	Adele	Someone Like You	Soul/Pop	Coração Partido	"Never mind, I'll find..."
3	Harry Styles	Fine Line	Pop/Indie	Superação e Amor	"We'll be a fine line..."
4	AC/DC	Back in Black	Rock	Renascimento	"Back in black, I hit..."
5	S. Carpenter	Sue Me	Pop	Autoestima	"So sue me for looking..."

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:

- **Vectorização TF-IDF:** As *strings* das letras foram transformadas em vetores numéricos usando o modelo TF-IDF (Term Frequency-Inverse Document Frequency).

Este modelo atribui pesos maiores a palavras que são distintas e relevantes dentro do conjunto de músicas, após a remoção de *stop-words* (palavras comuns) na língua inglesa.

- **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

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 1:

$$\text{Sim}(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} \quad (1)$$

O valor da similaridade varia entre 0 e 1. Um valor próximo de 1 indica alta similaridade (ângulo 0°), significando que os vetores (e, portanto, os textos) apontam na mesma direção. Um valor próximo de 0 indica baixa similaridade (ângulo 90° ou maior).

4 Resultados e Discussão

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 Carpenter
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

4.1 Análise Crítica dos Dados

- **Alta Correlação (0.62):** A maior similaridade foi observada entre as músicas de Ariana Grande e Adele. Ambas as letras abordam temas de relacionamento, amor e saudade, compartilhando um vocabulário emocional que o modelo TF-IDF capturou com sucesso.
- **Divergência Extrema (0.08 - 0.12):** A música do AC/DC apresentou consistentemente a menor similaridade com todas as outras. Seu gênero Rock, tema de renascimento e energia, e o vocabulário direto e distinto, resultaram em um vetor ortogonal (diferente) dos temas emocionais das demais músicas, demonstrando a

capacidade do algoritmo de diferenciar textos de domínios semânticos totalmente distintos.

- **Relação Pop/Sentimental (0.54):** Harry Styles manteve uma similaridade moderada com as músicas de teor romântico, indicando que seu vocabulário, apesar de focar em "Superação", ainda possui uma base comum com o Pop sentimental.

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.

5 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, que empregou a vetorização TF-IDF para ponderar a importância das palavras, 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.