FATEC BAIXADA SANTISTA - FACULDADE DE TECNOLOGIA - RUBENS LARA

SISTEMA DE RECOMENDAÇÃO DE FILMES

Aplicação das Métricas TF-IDF e Similaridade de Cosseno

Flávia Barbosa
Ciclo 2 - Ciência de Dados

Santos, SP
Novembro de 2025

Conteúdo

1 Resumo

Este projeto implementa um Sistema de Recomendação Baseado em Conteúdo utilizando a métrica de **Similaridade de Cosseno** e a técnica de vetorização **TF-IDF**. O objetivo é mapear o perfil de preferência textual do usuário (*Query*) em um espaço vetorial de alta dimensão e calcular sua proximidade angular com uma base filtrada de filmes do TMDB (*The Movie Database*). Após um rigoroso processo de ETL (Extração, Transformação e Carga) e filtragem por qualidade, idioma, duração e data, a base final utilizada para análise contém **30.612 títulos**. O teste aplicado demonstrou a eficácia do modelo em correlacionar perfis complexos ("Filmes de Super-Heróis") com *blockbusters* relevantes do universo cinematográfico.

2 Introdução e Fundamentação Teórica

2.1 Objetivo do Projeto

O principal objetivo deste trabalho é desenvolver e validar um sistema de recomendação de filmes. O sistema deve ser capaz de transformar um perfil de preferência textual fornecido pelo usuário (a *Query*) em um vetor numérico e, subsequentemente, ranquear os filmes mais semanticamente semelhantes em uma grande base de dados, utilizando a distância angular entre os vetores.

2.2 Similaridade de Cosseno (*Cosine Similarity*)

A Similaridade de Cosseno é uma métrica fundamental da Álgebra Linear utilizada para medir o quão parecidas são a **direção** de dois vetores em um espaço multidimensional. No contexto do Processamento de Linguagem Natural (NLP), ela mede a similaridade de conteúdo semântico entre dois documentos, ignorando o tamanho do documento.

A métrica é definida pelo cosseno do ângulo (θ) entre o Vetor da Query ($\mathbf{Q}$) e o Vetor do Documento ($\mathbf{D}$):

$$\text{Similaridade}(\mathbf{Q}, \mathbf{D}) = \cos(\theta) = \frac{\mathbf{Q} \cdot \mathbf{D}}{\|\mathbf{Q}\| \cdot \|\mathbf{D}\|} \tag{1}$$

Uma Similaridade ($\mathbf{S}$) próxima de **1** indica um ângulo θ próximo de $0°$, significando alta similaridade. Um $\mathbf{S}$ próximo de **0** indica baixa similaridade.

2.3 TF-IDF (*Term Frequency-Inverse Document Frequency*)

TF-IDF é a técnica de vetorização que transforma o texto em um formato numérico. Ela pondera a importância de uma palavra em um documento em relação a todo o *Corpus* (a base de dados de filmes). O *Term Frequency* (TF) mede a frequência da palavra no documento, enquanto o *Inverse Document Frequency* (IDF) penaliza palavras que aparecem em muitos documentos, valorizando termos raros e específicos que ajudam a diferenciar o conteúdo. A combinação resulta em vetores que representam a **importância semântica** das palavras.

3 Metodologia

3.1 Origem e Adaptação da Base de Dados

O *Corpus* foi construído a partir do **TMDB Movies Dataset** do Kaggle.

- **Fonte:** TMDB (The Movie Database).
- **Dataset Carregado:** `asaniczka/tmdb-movies-dataset-2023-930k-movies`.
- **Arquivo Utilizado:** `TMDB_movie_dataset_v11.csv`.

As colunas carregadas foram: `title`, `overview`, `keywords`, `genres`, `tagline`, `vote_average`, `runtime`, `adult`, `release_date`, e `original_language`.

3.2 Transformação (Filtros e Limpeza)

A Transformação (T do ETL) foi crucial para criar um *Corpus* de alta qualidade.

Tabela 1: Filtros Aplicados e Justificativas

Filtro	**Condição**	**Justificativa**
Qualidade	`vote_average` > 6.0	Garante a recomendação de filmes com avaliação média razoável, elevando a qualidade percebida.
Duração	`runtime` > 60	Remove curtas-metragens e *trailers*, focando apenas em filmes de longa-metragem (acima de 1 hora).
Relevância Temporal	`release_year` ≥ 1995	Foca a base em conteúdo mais recente (pós-1995), com maior completude de metadados e relevância atual.
Idioma do Conteúdo	`original_language` $==' en'$	**CRUCIAL para o NLP:** Alinha o idioma do *Corpus* com o idioma da *Query* (`english`), validando a comparação vetorial do TF-IDF.
Censura	`adult` $==$ False	Exclui filmes classificados como conteúdo adulto, padronizando a classificação do filme recomendado.

Após os filtros e a limpeza inicial dos dados nulos:

- **Feature Engineering:** A coluna `features` foi criada pela concatenação de `overview` + `keywords` + `genres` + `tagline`.
- **Limpeza de Texto:** Foi aplicada uma função de remoção de caracteres não-ASCII (*e.g.*, acentos e caracteres latinos especiais) e transformação para minúsculas, garantindo que o texto da base fosse limpo e homogêneo.

A base final utilizada para análise (carga) resultou em **30.612 títulos** de alta qualidade e relevância.

3.3 Preparação da Query e Simetria

O texto de perfil de preferência do usuário (*Query*) é processado simetricamente à base de dados: é submetido à mesma função de limpeza (`remove_non_ascii`) antes de ser adicionado ao *Corpus*.

3.4 Vetorização e Parâmetros

O *Corpus* (Query + 30.612 Filmes) é transformado pela classe `TfidfVectorizer` com os seguintes parâmetros para otimizar o resultado:

- `stop_words='english'`: Remove palavras comuns do idioma que não possuem valor semântico.
- `lowercase=True`: Remove a distinção entre maiúsculas e minúsculas.
- `token_pattern=r'\b\w+\b'`: Garante a remoção de pontuação e considera apenas sequências alfanuméricas como *tokens*.

3.5 Cálculo da Similaridade

A matriz de *cosine similarity* é calculada entre o **vetor da Query** e todos os **vetores dos filmes**, resultando em um *array* de pontuações de **S** (similaridade). O ranqueamento é feito ordenando os filmes por **S** em ordem decrescente (do mais semelhante para o menos semelhante).

4 Análise de Resultados

4.1 Teste Aplicado: Perfil de Super-Heróis

O teste foi realizado utilizando uma *Query* focada em filmes de ação de Super-Heróis, com alta repetição de termos-chave para garantir um vetor dominante.

```
Query: I am looking for a superhero action film with massive scale and a
focus on team dynamics. The plot must involve powerful heroes uniting
to fight a global threat or a cosmic villain. Key elements include
spectacular visual effects, epic battles, and complex interconnected
storylines within a shared universe. I prefer films about superpowers,
advanced technology, and saving the world, mixed with humor and
emotional stakes. The film should explore teamwork, sacrifice, and the
burden of heroism.
```

4.2 Ranking de Recomendação (TOP 5)

O modelo gerou o seguinte ranqueamento, que valida a correlação semântica entre a *Query* e as descrições dos filmes:

Rank	Filme	Similaridade (S)	Ângulo (θ) (Graus)
1	Spider-Man: No Way Home	0.159903	80.798727
2	Zack Snyder's Justice League	0.155047	81.080503
3	Batman and Harley Quinn	0.133499	82.328149
4	Cyberworld - The future is now	0.131263	82.457394
5	Nirbhay	0.125340	82.799585

4.3 Discussão dos Resultados

O ranqueamento demonstra a eficácia do sistema:

- Os filmes de maior similaridade (Rank 1 e 2) são títulos proeminentes de Super-Heróis (*Spider-Man* e *Justice League*), que possuem alto conteúdo textual em relação aos termos *superhero*, *team dynamics* e *global threat*.
- O ranqueamento é preciso: o filme mais relevante para o gênero obteve o **S** mais alto (≈ 0.16), que corresponde ao **menor ângulo** ($\approx 80.80^\circ$), provando que seu vetor é o mais próximo do vetor de preferência do usuário no espaço vetorial.

5 Conclusão

O projeto validou com sucesso a implementação de um sistema de recomendação baseado em conteúdo utilizando a **Similaridade de Cosseno** e **TF-IDF**. O rigoroso processo de ETL e filtragem resultou em um Corpus otimizado de **30.612 títulos** consistentes e relevantes. A análise comprovou que a distância angular é uma métrica poderosa para mapear a proximidade semântica do perfil do usuário com o conteúdo fílmico, resultando em um sistema de recomendação funcional e coerente.