Recomendação de Títulos com Base na Similaridade do Cosseno

Thabata Fernandes Zacarias

Maio 2025

Objetivo

O código criado localiza os títulos e descrições mais parecidas, baseando-se na descrição, utilizando a similaridade do cosseno. A similaridade do cosseno avalia o quão parecidos são dois ou mais textos, transformando-os em vetores e calculando o ângulo entre eles.

Quando usamos a similaridade do cosseno, comparamos dois textos para ver o quanto eles se parecem. Essa comparação transforma os textos em números e analisa o ângulo entre eles. Quanto menor esse ângulo (mais próximo de 1), mais parecidos os textos são. Mesmo que não sejam exatamente iguais, essa técnica consegue identificar descrições com temas ou palavras em comum.

Esse tipo de cálculo é muito útil em sistemas de recomendação, pois ajuda a encontrar conteúdos com enredos parecidos. Se os vetores dos textos apontam para direções semelhantes, a similaridade é alta. Se estão em direções muito diferentes, a similaridade é baixa.

Desenvolvimento

O dataset escolhido para este trabalho foi o *Netflix Movies and TV Shows*, retirado da plataforma Kaggle, onde estão reunidos séries e filmes originais da Netflix em inglês. O conjunto de dados totaliza 8.809 linhas e 12 colunas, das quais foram utilizadas apenas duas: `title` (título) e `description` (descrição).

O algoritmo desenvolvido encontra o título e a descrição mais semelhantes ao título inserido pelo usuário, comparando as descrições do dataset. O funcionamento pode ser descrito da seguinte forma:

- A primeira etapa do algoritmo é carregar os dados do arquivo `netflix_dataset.csv`,

que contém informações sobre diversos títulos e descrições. Cada linha do arquivo é armazenada como um dicionário com as chaves "título" e "descrição".

- Para comparar descrições de forma matemática, os textos precisam ser transformados em vetores numéricos. O algoritmo faz isso usando uma contagem de palavras.
- Cada palavra é contada, formando um dicionário onde:
  - chave = palavra
  - valor = número de vezes que a palavra aparece
- Esse processo é feito tanto para a entrada do usuário quanto para as descrições do dataset.
- O algoritmo compara cada descrição usando a fórmula da similaridade do cosseno:

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

  onde $A \cdot B$ é o produto escalar dos vetores (soma do produto das frequências das palavras em comum) e $\|A\|$, $\|B\|$ são as normas dos vetores (raiz quadrada da soma dos quadrados das frequências).
- O resultado final é um valor entre 0 e 1, onde:
  - 1 = textos idênticos (mesma direção vetorial)
  - 0 = textos completamente diferentes (vetores ortogonais)
- O algoritmo percorre todas as descrições do dataset, calcula a similaridade do cosseno entre elas e a descrição inserida, e retorna o título e descrição com maior similaridade.

Resultado

A seguir, apresentamos alguns exemplos práticos, onde é inserido um título e o sistema retorna o título e a descrição mais semelhantes com base na similaridade do cosseno:

Exemplo 1

Digite um título: breaking bad

Resultado mais parecido com base na descrição:
Título: The Heat: A Kitchen (R)evolution
Descrição: Seven female chefs describe what it's like breaking into the restaurant industry's notorious boys' club.

Exemplo 2

Digite um título: vikings

Resultado mais parecido com base na descrição:
Título: The Last Kingdom
Descrição: As Alfred the Great defends his kingdom from Norse invaders, Uhtred – born a Saxon but raised by Vikings – seeks to claim his ancestral birthright.

Exemplo 3

Digite um título: squid game

Resultado mais parecido com base na descrição:
Título: Glitch Techs
Descrição: Two teens work at a game store as a front for their actual job: Hunting video game monsters who've broken out into the real world.

Exemplo 4

Digite um título: stranger thinks

Resultado mais parecido com base na descrição:
Título: The Spiral
Descrição: When an unbearable stranger arrives at a winter mountain resort and vanishes overnight, his dark past comes to light.