Sistema de Recomendação Utilizando Similaridade do Cosseno

1 Introdução

Os sistemas de recomendação têm se tornado uma técnica importante e recorrente no campo da informática, tanto pela sua versatilidade quanto pelo seu poder de otimizar pesquisas, retornando resultados relevantes que satisfaçam as necessidades do usuário.

Esses sistemas são amplamente utilizados, como por exemplo pelo Google para recomendar sites, ou por serviços de streaming como Netflix, Amazon Prime Video e HBO Max, para sugerir filmes e séries com base no que é digitado.

Considerando a importância desses sistemas, este projeto propõe uma aplicação baseada na similaridade do cosseno, um algoritmo que analisa a similaridade entre vetores por meio do cosseno do ângulo entre eles: quanto menor o ângulo, maior a similaridade.

Para utilizar tal algoritmo é necessária uma base de dados legível pelo código. O funcionamento do sistema será descrito a seguir, incluindo o dataset utilizado e os detalhes técnicos da implementação.

2 Explicações do Dataset

O dataset utilizado foi obtido no site `Kaggle.com`, sob o nome "Manga, Manhwa and Manhua Dataset", disponível no formato CSV com o nome `data.csv`. Ele contém seis colunas:

- **title**: título da obra;
- **description**: sinopse ou descrição da obra;
- **rating**: nota de popularidade;
- **year**: ano de lançamento;
- **tags**: gêneros associados à obra;
- **cover**: link para a imagem da capa.

A base de dados já estava organizada, sem necessidade de tratamento adicional.

3 Objetivo

O objetivo do projeto é identificar obras similares àquela fornecida pelo usuário, considerando as descrições e os gêneros. A similaridade é calculada usando o cosseno entre os vetores TF-IDF dessas informações.

A fórmula da similaridade do cosseno é dada por:

$$\text{Sim}_{\cos}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

Onde $\|A\|$ representa o tamanho do vetor A, calculado por:

$$\|A\| = \sqrt{A \cdot A} \tag{2}$$

O cálculo considera pesos diferentes para as descrições (70%) e para os gêneros (30%), pois as descrições tendem a conter mais informações relevantes.

4 Funcionamento do Projeto

O sistema realiza os seguintes passos:

1. Leitura do arquivo `.csv` com as obras;
2. Pré-processamento textual: conversão para minúsculas, remoção de pontuação, tokenização e remoção de stopwords;
3. Vetorização dos textos usando TF-IDF:
   - até 10.000 termos para descrições;
   - até 5.000 termos para tags.
4. Cálculo da similaridade do cosseno entre as obras;
5. Conversão da similaridade em ângulo para facilitar a interpretação;
6. Cálculo da média ponderada da similaridade (descrição com peso 0.7, tags com 0.3);
7. Entrada do usuário com o nome da obra e retorno das 3 mais similares.

5 Resultados

A execução do projeto solicita ao usuário que digite o nome de uma obra. Como exemplo:

Digite o nome de uma obra: `Moriarty the Patriot`

Resultado retornado:

- **1. Young Miss Holmes**
  Ângulo: 71.29° — Similaridade: 0.3207
- **2. Sherlock Holmes**
  Ângulo: 71.56° — Similaridade: 0.3164

- **3. Christie London Massive**
  Ângulo: 72.55° — Similaridade: 0.2999

Esses resultados mostram as obras mais similares à escolhida, considerando a abordagem de pesos desbalanceados entre descrição e gênero.