

Relatório do Algoritmo de Similaridade do Cosseno Aplicado em um Dataset de Jogos da Steam

Rodrigo Cavalheiro Dos Santos
FATEC Rubens Lara – Baixada Santista
Disciplina: Álgebra Linear – 2º CD
Docente: Prof. Alexandre Garcia de Oliveira

1 Introdução

A Steam é uma plataforma de distribuição de jogos digitais lançada em 2003 e, atualmente é considerada a maior plataforma de publicação e distribuição de jogos eletrônicos.

Esse projeto tem como objetivo explorar um dataset sobre jogos publicados na Steam desde o seu lançamento até o ano de 2019. Visando aplicar técnicas de álgebra linear, stopwords e TF-IDF (Term Frequency-Inverse Document Frequency) para calcular a similaridade do cosseno entre os jogos.

2 Seleção e Explicação do Dataset

Seleção do dataset pesquisado e estudado na plataforma Kaggle, buscando um dataset simples para facilitar o tratamento e com um número de jogos relevante.

A busca por um dataset com uma quantidade razoável de jogos e com poucas colunas para facilitar a exclusão de colunas insignificantes para o projeto foi uma tarefa demorada. Os principais motivos da escolha do dataset utilizado no projeto foram: quantidade de jogos (cerca de 27 mil), número de colunas (18 no total) e o fato do autor do dataset ter excluído a maioria dos programas que não são considerados jogos. O dataset não foi criado oficialmente pela Steam, foi criado por um estudante de ciência de dados para ele aplicar em trabalhos, projetos etc.

2.1 Explicando o Dataset

O dataset original é composto por 18 colunas, são elas:

1. **appid**

Número único utilizado para identificar cada título na plataforma e no dataset.

2. **name**

Título do jogo.

3. **release_date**

Data de lançamento do jogo.

4. **english**
Se o jogo suporta o idioma inglês (a linha estará com o número 1 caso tenha suporte).
5. **developer**
Nome da desenvolvedora.
6. **publisher**
Nome da publicadora.
7. **platforms**
Em quais sistemas operacionais o jogo está disponível.
8. **required_age**
Idade mínima recomendada.
9. **categories**
Categorias em que o jogo se encaixa, separadas por ponto e vírgula (single-player, co-op etc).
10. **genres**
Gêneros separados por ponto e vírgula (action, RPG etc.).
11. **steampsy_tags**
Gêneros em que o jogo se encaixa por meio de votações da comunidade da Steam, separados por ponto e vírgula (multiplayer, indie etc.).
12. **achievements**
Número máximo de conquistas dentro do jogo.
13. **positive_ratings**
Número de avaliações positivas da comunidade.
14. **negative_ratings**
Número de avaliações negativas da comunidade.
15. **average_playtime**
Tempo médio de jogo por usuário.
16. **median_playtime**
Mediana do tempo de jogo por usuário.
17. **owners**
Números de usuários que possuem o jogo.
18. **price**
Preço total do jogo (ignorando promoções)

As únicas colunas com texto (string) são: name, developer, publisher, platforms, categories, genres e steampsy_tags. Todas elas, menos a coluna platforms e as demais sem string serão descartadas.

3 Tratando o Dataset

A primeiro passo dado após a seleção do dataset é tratá-lo devidamente para tornar o projeto mais simples e fácil tanto para desenvolver o algoritmo quanto para o usuário que irar avaliá-lo.

Nesse processo, foi utilizado a biblioteca Pandas do Python para exclusão de colunas indesejadas e remoção de caracteres fora da formatação ASCII como: ® e ™.

3.1 Stopwords

A aplicação da técnica de stopwords serve para eliminar palavras que se repetem constantemente, o que dificulta a aplicação da técnica TF-IDF que, no algoritmo foi feita antes do processo de stopwords. Como o dataset é uma lista de jogos, contendo gêneros e tags majoritariamente em inglês, foi importada uma biblioteca de stopwords em inglês. Também é feita a exclusão de algumas palavras que são comumente usadas no meio de jogos, que no código é chamada de *'custom_stopwords'*.

4 Matriz TF-IDF

Primeiramente é feita a configuração do código para que seja feita a combinação dos textos (gêneros, categorias, desenvolvedor etc.) para cada título de jogos, ou seja, cada linha do dataset.

Após a aplicação das configurações, é feita a matriz TF-IDF do dataset modificado convertendo texto em números, destacando os termos mais importantes de cada jogo.

5 Insights da Matriz TF-IDF

Essa parte do código serve apenas para mostrar no console os seguintes itens:

1. Uma amostra dos textos originais e os que foram processados na matriz de 3 jogos aleatórios.
2. Mostra os 10 termos que mais aparecem na matriz.
3. Calcula a similaridade do cosseno entre 3 pares de jogos definidos, são eles: Call of Duty 2 e Call of Duty, Portal 2 e Portal, Left 4 Dead e Left 4 Dead 2.
4. Mostra os termos em comum entre cada jogo acima.

6 Análise dos Resultados

Alguns exemplos dos resultados são mostrados no console para o usuário, é visto que existe uma grande similaridade entre os primeiros jogos da franquia Call of Duty, tendo uma similaridade de 0.94, Portal tendo uma similaridade de 0.68 e Left 4 Dead uma similaridade de 0.83. Nota-se que, os jogos da franquia Portal são os que tem a menor similaridade, isso é causado pela falta de alguns termos entre os jogos, por mais que Portal 2 seja a continuação do primeiro jogo, alguns termos são adicionados ao jogo por meio dos usuários da comunidade da Steam, que seria a coluna *'steampsy_tags'*, e a desenvolvedora pode ter adicionado mais ou menos termos nos respectivos jogos.

7 Conclusão

O algoritmo de similaridade do cosseno com TF-IDF funcionou bem para identificar o quanto os jogos da Steam se parecem, com base nas informações de gêneros, categorias, desenvolvedora, publicadora e outras descrições.

Comparações entre jogos da mesma franquia, como Call of Duty, Portal e Left 4 Dead, mostraram resultados condizentes com o esperado, revelando que os jogos realmente compartilham vários elementos em comum, mesmo que tenham algumas diferenças nos termos usados.

Durante o desenvolvimento, foi essencial limpar e preparar os dados, como tirar colunas irrelevantes e remover palavras repetitivas (stopwords), para que a análise fosse mais precisa. A junção dos textos das colunas ajudou bastante a representar melhor cada jogo e facilitou a criação da matriz TF-IDF.

Além disso, o algoritmo possui uma interface extremamente intuitiva e funcional, mostrando alguns jogos aleatórios para pesquisar, permite que o usuário não precise inserir o nome completo do jogo etc. Também conta com um gráfico de barras para melhor visualização do ranking dos 10 jogos mais similares referente ao escolhido.

8 Referências Bibliográficas

Dataset Original:

<https://www.kaggle.com/datasets/nikdavis/steam-store-games/data>

TF-IDF:

<https://letsdatascience.com/tf-idf/>

Stopwords:

<https://www.geeksforgeeks.org/removing-stop-words-nltk-python/>