Recomendação de Filmes/Séries

Desenvolver um sistema de recomendação de filmes e séries com base no catálogo de um serviço de streaming, com o objetivo de melhorar a experiência do usuário e aumentar sua satisfação e engajamento. O campo "description" (descrição), que contém informações sobre os filmes, deve ser utilizado, observando que o conteúdo está em inglês e não em português.

O trabalho pode ser desenvolvido em dupla, utilizando a linguagem Python. A documentação detalhada deve ser enviada para o e-mail (alexandreasouza@ufgd.edu.br). Envie o código em Python e prints de tela da execução e dos resultados obtidos.

Análise do problema:

1. Importar os dados via base de dados do catálogo e usar o campo description (descrição) para identificar similaridade;
2. Usar análise de texto (TF-IDF) para transformar descrições de filmes em dados numéricos comparáveis;
3. Realizar agrupamento (clustering) com MiniBatchKMeans para formar grupos de filmes semelhantes;
4. Com base na construção do grafo, modelar relações complexas entre filmes, atores, diretores, etc., e identificar similaridades através das conexões no grafo, usando métricas de vizinhança e pesos de conexão.

Sugestão de desenvolvimento:

Passo 1: Importação de Bibliotecas Necessárias

Será necessário importar as bibliotecas para manipulação de dados, criação de grafos, visualização e aprendizado de máquina. Recomenda-se a exportação das bibliotecas: pandas, networkx, matplotlib.pyplot, numpy e scikit-learn.

Passo 2: Leitura e Pré-Processamento do Dataset

Carregar os dados do arquivo CSV contendo informações sobre títulos disponíveis no serviço de streaming.

Passo 3: Transformação de Texto em Vetores TF-IDF

A transformação de texto em vetores TF-IDF (Term Frequency-Inverse Document Frequency) tem o objetivo de converter **as descrições dos filmes** em vetores numéricos que representem a importância relativa das palavras em cada descrição. Isso é necessário porque os algoritmos de aprendizado de máquina, incluindo métodos de agrupamento (como K-Means) ou cálculos de similaridade (como similaridade por cosseno), trabalham com valores numéricos e não com texto bruto.

Observação: Usar **TfidfVectorizer** para transformar os textos em uma matriz TF-IDF.

Passo 4: Agrupamento com K-Means

Nessa etapa iremos agrupar filmes em clusters com base na semelhança de suas descrições, recomendo que usem o MiniBatchKMeans que é uma variante do algoritmo KMeans. O MiniBatchkMeans é usado para agrupar os vetores TF-IDF das descrições dos filmes, um número razoável de ser adotado é de **100**

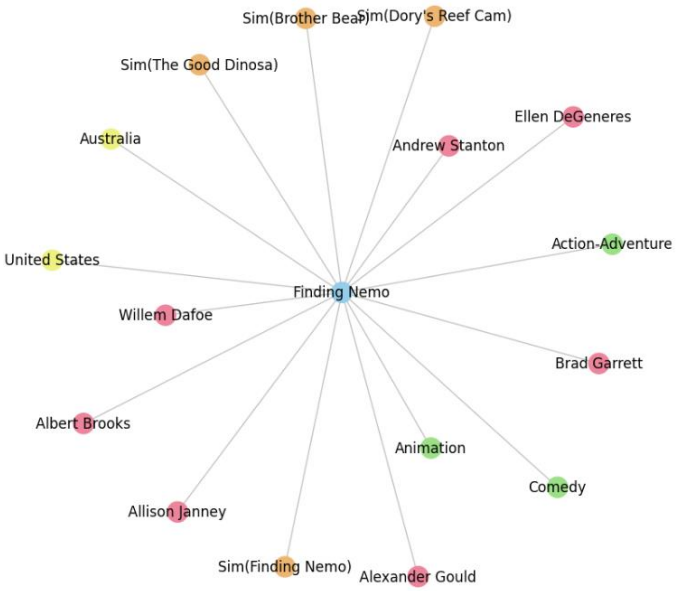
clusters. Assim, podemos agrupar filmes com descrições semelhantes em grupos distintos, facilitando a análise, exploração e recomendações baseadas em características semânticas das descrições dos filmes.

Passo 5: Construção do Grafo com NetworkX

Criar um grafo onde os filmes são conectados a atores, diretores, países e categorias. Deve-se usar networkx para construir e adicionar nós e arestas no grafo.

Passo 6: Desenho e Visualização do Subgrafo (opcional)

Visualizar partes do grafo para explorar conexões e validar os resultados.



Passo 7: Recomendação com Métrica de Adamic-Adar

Com base na elaboração do grafo, fornecer recomendações de filmes com base na similaridade de conexões no grafo. Sugestão: usar a métrica de Adamic-Adar para calcular recomendações.

Passo 8: Recomendação

Ao passar o nome de um filme ou série, indicar 5 filmes ou séries com os maiores índices de similaridade.