AGEMC do projeto FMF

Matéria: Pensamento Analítico de Dados (PAD)

Professor: Fernando Federson

Alunos: Igor Dias, João Pedro de Castro, Bruno Calura, Vinícius Tormin, Samuel Lopes e Cleiver Batista

- **A (ASK):**

No projeto de sistema de recomendação FMF, a questão central a ser respondida é: "Dado um filme que um usuário gosta, quais outros são suficientemente similares para que ele também goste?".

Para responder a essa pergunta, toda a modelagem de dados e a estrutura do projeto são focadas em calcular a similaridade entre um item de referência e os demais disponíveis (neste caso, os filmes).

- **G (GET):**

Os dados utilizados no projeto são os de filmes disponibilizados pelo **MovieLens**, especificamente as versões com 10 milhões e 20 milhões de avaliações. Este conjunto de dados contém diversos arquivos CSV separados, que descrevem principalmente as relações entre filmes, suas avaliações e tags associadas. Além desses datasets de base, ao final do projeto, Prateek Gaurav utiliza **outros dados** do MovieLens para enriquecer as informações e aprimorar a recomendação. Estes dados novos contêm mais detalhes sobre atores, notas, episódios e datas.

- **E (EXPLORE):**

O trabalho com os dados exige uma etapa de exploração inicial, seguida de limpeza e tratamento, para viabilizar a aplicação de técnicas de machine learning. Como Prateek Gaurav desenvolve múltiplas implementações, os processos de preparação dos dados variam em cada uma delas. Por essa razão, este trabalho não detalhará cada tratamento específico, mas sim a estratégia geral de seleção de atributos.

Inicialmente, nas implementações mais simples, o autor opta por utilizar os gêneros dos filmes como o principal guia para a recomendação. Em um segundo momento, ele enriqueceu essa abordagem ao combinar os dados de gênero com as tags associadas a cada filme. Por fim, nas soluções mais sofisticadas, Gaurav utilizou um conjunto de dados mais completo, resultado da união de diferentes tabelas. Nessa etapa final, ele explorou atributos como título, duração, gênero, diretor, roteirista, nota média, número de votos e elenco para construir o sistema de recomendação.

- **M (MODEL):**

O cientista de dados apresenta diversas possibilidades de modelagem com alguns exemplos cada vez mais completos. Sendo assim, ele aborda basicamente as técnicas e ferramentas abaixo:

1. Similaridade do cosseno;
2. Bag of Words (BoW);
3. TF-IDF;
4. Binary Feature Matrix;
5. LSA;
6. Word2Vec.

- **C (COMMUNICATE):**

No projeto FMF, a etapa de Comunicação é o item menos explorado, devido ao foco do material ser em formato de videoaula. Por essa razão, o autor do conteúdo apenas apresenta os resultados diretamente no ambiente Google Colab, sem desenvolver uma interface de apresentação. Adicionalmente, embora ele disponibilize os códigos e os comente no artigo de apoio, a comunicação dos resultados em si não é o foco do projeto, que prioriza a construção do código. Conclui-se, portanto, que este é um ponto que pode ser aprimorado em nosso projeto.