Desafio de Cientista de Dados

Sugestão de desenvolvimento de MVP para analisar sentimentos em comentários de leitores livros por Gilmar Santos.

Apresentação do Desafio

O desafio se resume a construção de um MVP para análise de scores de avaliação de livros de uma livraria por leitores e a análise dos comentários de leitores, bem como, a extração de análise de sentimento dos textos registrados nesses comentários.

Objetivo: avaliar as publicações pelo viés dos leitores.

Planejamento de entregáveis (inclusive futuros) - Roadmap

1. jupyter notebook: contendo o passo a passo da construção da análise de dados;
2. mvp: construção de aplicação utilizando linguagem python e llm gpt-3.5-turbo para analisar individualmente comentários de autores da publicação;

Explicação do Processo Utilizado

1. Análise dos dados: estudo dos dados para entender a organização desses dados para extrair insights, os arquivos são: books_data.csv e books_rating.csv;
2. Preparação dos dados: adequação dos dados a fim de atender as necessidades dos requisitos levantados;
3. Construção de código em python para interagir com a api da openAI do modelo de LLM GPT 3.0;

Hipóteses Levantadas

1. Quais os livros mais vendidos;
2. Quais autores mais vendidos;
3. Quais os gêneros mais consumidos;
4. Como estão a avaliação dos livros do ponto de vista dos leitores;
5. ...

Análise Exploratória

1. Compreensão e análise dos arquivos books_data e books_rating.
2. Construção dos DataSet's:
    a. **Descrição dos dados do Data Frame df_books_data**
        i. Quantidade de Linhas: 212.404 -> desta feita, conclui-se que estão registrados 212.404 livros
        ii. Quantidade de Colunas: 10 -> Que são: Title, description, authors, image, previewLink, publisher, publishedDate, infoLink, categories, ratingsCount

    b. **Descrição dos dados do Data Frame df_books_rating**
        i. Quantidade de Linhas: 3.000.000 -> desta feita, conclui-se que estão registrados 3.000.000 de comentários/avaliações
        ii. Quantidade de Colunas: 9 -> Que são: Id, Title, price, User_Id, profileName, score, time, summary, text

Gráfico de Quantidade de comentários por Score no dataframe books_rating

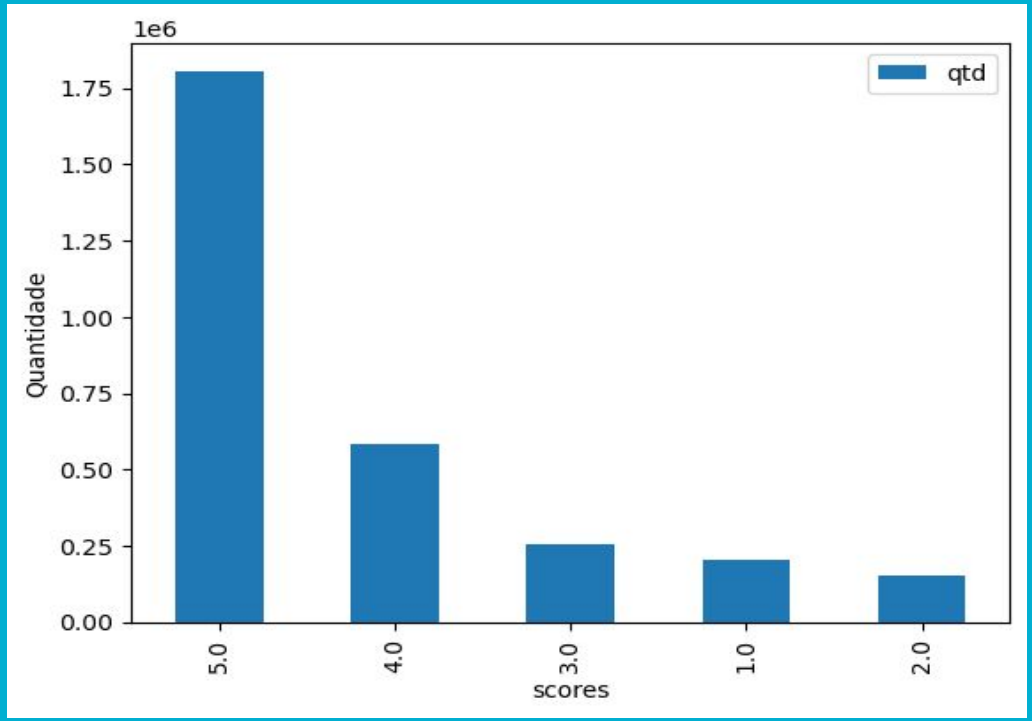
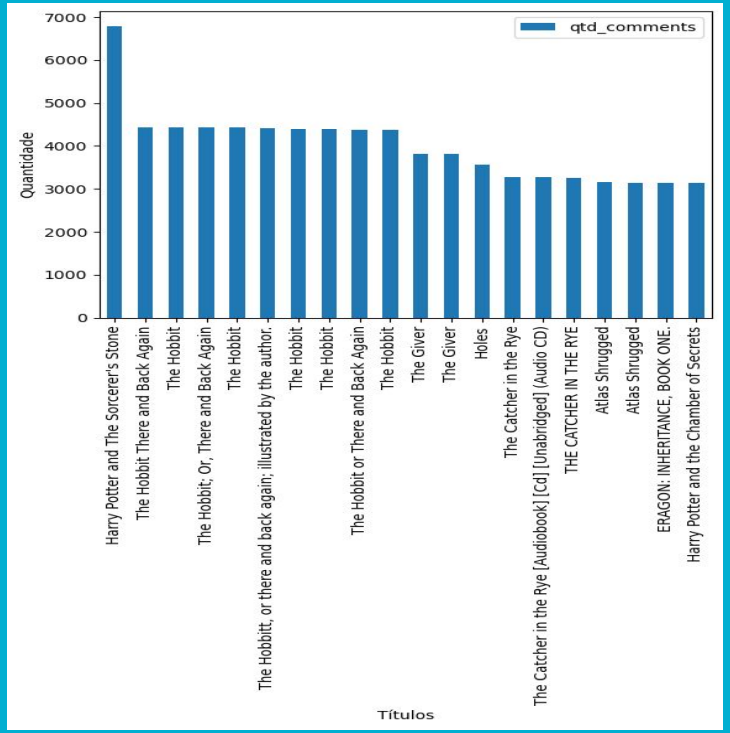


Gráfico de quantidade de Comentários por ISBN



Sumarização das informações textuais (Pelo menos a proposta com utilização de alguma tecnologia).

Para análise de sentimento dos comentários, utilizou-se mecanismo de modelo LLM, gpt-3.5-turbo-instruct, analisando o atributo summary, do dataset df_books_rating, que armazena os comentários de leitores dos livros.

-> Para execução, construiu-se aplicação com linguagem Python para conexão com a API da OPENAI.

Uso de bases de conhecimento (Proposta de utilização).

-> Para o caso específico, análise de sentimentos sobre os comentários dos leitores, fez-se necessário apenas a geração da lista dos comentários dos livros com os respectivos Id's como base de conhecimento.

Neste momento, não foi preciso utillizar-se de RAG (Retrieve Augmented Generate), porém esse recurso poderá ser utilizado quando necessário para incorporar conhecimento específico ao modelo. Bem como, agentes para conexão, via API, com outras funcionalidades da empresa.

Métricas para avaliar a qualidade do resultado.

A métrica de avaliação da qualidade do resultado foi alcançada utilizando BLEU Score, cujo acrônimo em inglês é Bilingual Evaluation Understudy que é uma métrica usada para avaliar a qualidade de resposta gerada por modelo de linguagens. Essa métrica tem um valor entre 0 e 1 e, quanto maior é a correspondência entre a resposta gerada e a referência original. Utilizou-se a fórmula:

$$BLEUScore = BP * \exp(\sum_{i=1}^{N}(w_i * \ln(p_i)))$$

LLM	Teste BleuScore
GPT-3.5-Turbo	0.981112

Fim!!!

Contato: Gilmar Correa dos Santos - gilmarsan@gmail.com