Eduardo Darrazão - 1906399

Marcelo Guimarães da Costa - 1937570

Leandro Batista de Almeida - Professor

Big Data e Aplicações

12 de dezembro de 2023

Projeto 1: Aquisição e Processamento de um Dataset

Este projeto é a parte 1 de um projeto final de duas partes, sendo o objetivo final uma apresentação para a classe que envolveria todos os processos desde a seleção de um *dataset* até a utilização de modelos de *AutoML* sobre o *dataset* escolhido.

Nesta parte do projeto, os processos incluem:

- Selecionar *dataset* (fontes públicas ou privadas)
- Carregamento de dados (CSV, importação de bancos de dados, etc)
- Tratamento de informações
- Validação de erros
- Preenchimento de valores faltantes
- Limpeza de dados

1 – SELEÇÃO DO DATASET

O *dataset* escolhido foi o *IMDB (Internet Movie Database)*, que abrange séries e filmes de vários gêneros, bem como a equipe envolvida e a média das avaliações e números de votos. Os arquivos disponíveis são os seguintes:

- *title.basics*: Informações principais sobre cada título, bem como gênero, e tipo (filme, série, vídeo, etc)
- *title.akas*: Traduções e localizações do nome do título para diferentes culturas. Referencia title.basics
- *title.principals*: Dados sobre cada membro da equipe de um título, bem como sua função naquele título (atores, escritores, diretores, etc). Referencia title.basics
- *title.crew*: Informações sobre a equipe de direção e escritores em cada título. Referencia title.basics e title.principals
- *title.episode*: Dados sobre episódios no caso de séries, conectando a série com o episódio. Referencia title.basics duas vezes
- *title.ratings*: Média de avaliações (1 a 10) e número de votos para certos títulos. Referencia title.basics
- *name.basics*: Mais informações sobre alguns registros de title.principals. Data de nascimento, nome, pelo que é conhecido, etc. Referencia title.principals

O *dataset* foi baixado do site https://developer.imdb.com/non-commercial-datasets/ na data 2023-12-02.

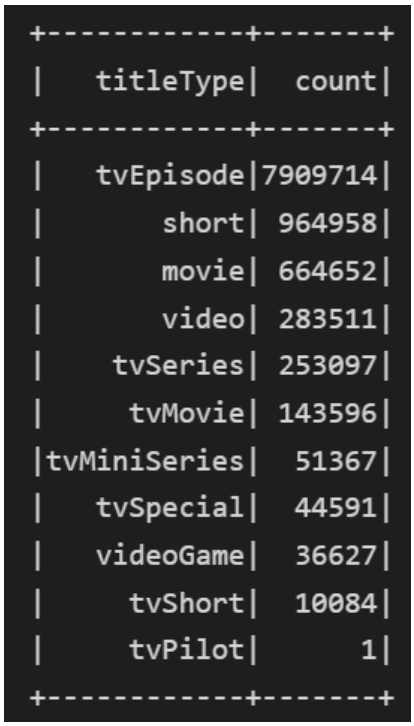
2 – CARREGAMENTO DOS DADOS

Para tratamento dos dados foi utilizada a biblioteca *pyspark*, mais especificamente utilizando a extensão Spark Dataframes. Segue um exemplo de carregamento:

```
title_basics = spark.read.load('../Data/title.basics.tsv', format='csv', sep='\t', inferSchema=True, header=True)
```

3 – TRATAMENTO, LIMPEZA E TRANSFORMAÇÃO DOS DADOS

A primeira etapa foi selecionar apenas os filmes, ou *movies*, que representariam nosso corte de interesse. Dessa forma, começamos a trabalhar com um *dataset* de 664 mil linhas.



titleType	count
tvEpisode	7909714
short	964958
movie	664652
video	283511
tvSeries	253097
tvMovie	143596
tvMiniSeries	51367
tvSpecial	44591
videoGame	36627
tvShort	10084
tvPilot	1

Depois, filtramos pela presença de gênero, no campo *genres*, que seria crucial para o modelo que planejávamos fazer. Isso nos deixou com um *dataset* de 590 mil registros.

Por fim, filtramos nosso *dataset* novamente pela junção interna (*inner join*) com a tabela *title.ratings*. Como *averageRating* será nosso campo *target* para os modelos desempenhados na parte 2, a existência do registro em *title.ratings* é essencial. Isso nos deixou com uma massa de 290 mil linhas.

O *dataset* apresenta algumas colunas com dados relevantes faltantes, como no *title.basics*, as colunas *runtimeMinutes* e *startYear*. Estas contêm o valor \N por vezes, que para este *dataset* significa a ausência de valor, ou nulo. Estes foram substituídos por 0, para que não tenhamos mais perda de dado.