Data Scraping and Analysis of IMDb's Top TV Shows

Name: Ahmed Ayman

ID:221341

Name: Mariam Moataz

ID:223675

Name: Nada Mostafa

ID:221427

Introduction

This project focuses on data scraping and analysis of IMDb's top TV shows. IMDb is a widely used online database for movies, TV shows, and celebrities, and the top TV shows list is a significant resource for understanding popular and critically acclaimed television content. The primary aim is to extract relevant information about these top TV shows and perform a detailed analysis to derive meaningful insights.

Objectives

1. **Data Extraction**:
    - Scrape IMDb's top TV shows page to gather information about each show.
    - Extract specific details such as title, release year, rating, genres, creators, stars, and show image.
2. **Data Processing**:
    - Clean and format the scraped data for consistency and readability.
    - Convert lists of creators, genres, and stars into comma-separated strings.
3. **Data Analysis**:
    - Identify correlations and insights about the popularity and characteristics of top TV shows.
    - Use Power BI to create visual representations of the data for better insights.
4. **Data Storage**:
    - Save the cleaned data into an Excel file for further analysis and reporting.

Planned Approach:

Step 1: Web Scraping

Tools and Libraries:

- **requests** for sending HTTP requests to IMDb.
- **BeautifulSoup** for parsing HTML content and extracting data.
- **pandas** for data manipulation and storage.

Process:

1. Send a GET request to the IMDb top TV shows page.
2. Parse the HTML response to extract links to individual TV show pages.
3. For each TV show page, extract the following details:
    - Title
    - Release year
    - Rating
    - Genres
    - Creators
    - Stars
    - Show image URL

Step 2: Data Cleaning and Processing

Tools and Libraries:

- **pandas** for data manipulation.

Process:

1. Compile extracted data into a pandas DataFrame.
2. Clean and format the data:
    - Extract the release year from the year string.
    - Convert lists (creators, genres, stars) to comma-separated strings.
    - Handle any missing data.

Step 3: Data Storage

Tools and Libraries:

- **pandas** for data storage.

Process:

1. Save the cleaned and processed DataFrame to an Excel file (**ImdbTopTvShows250.xlsx**).
2. Ensure the Excel file is well-formatted and includes all relevant details for further analysis.

Step 4: Data Analysis and Visualization with Power BI

After the data has been cleaned and saved to an Excel file, Power BI is used to perform detailed analysis and create visualizations. The following key relationships and visualizations are generated:

Max of Rating by Release Year:

A donut chart showing the maximum rating of TV shows for each release year.

Helps identify the years with the highest-rated TV shows.

Count of Titles by Release Year:

A clustered column chart showing the number of TV shows released each year.

Provides insight into trends in TV show production over the years.

Count of Titles by Genre:

An area chart displaying the number of TV shows in each genre.

Highlights the most common genres among the top TV shows.

Max Rating by Genre:

A line chart showing the highest rating for each genre.

Identifies the genres with the highest-rated TV shows that people love.

By using Power BI, these visualizations provide a comprehensive overview of the data, making it easier to derive insights and understand trends in popular TV shows.