Report on Scraping IMDb for Movie Ratings

Introduction

Web scraping is a valuable technique for extracting data from websites, enabling data analysis and insights. This project focuses on scraping IMDb's Top 100 movie ratings, followed by data cleaning and preprocessing. The goal is to create a structured dataset for further exploration and visualization.

Web Scraping Process

Tools and Technologies Used

- **Selenium**: For navigating and extracting data from IMDb.
- **Pandas**: For handling and processing the dataset.
- **WebDriver Manager**: For automating ChromeDriver installation.

Steps Involved

1. **Setting Up Selenium WebDriver**
    - Chrome WebDriver was initialized in headless mode to efficiently scrape data.
    - A user-agent string was used to bypass bot detection.
2. **Extracting IMDb Movie Data**
    - The script navigated to IMDb's Top 250 movies page.
    - The first 100 movie entries were located using XPath selectors.
    - Relevant data points extracted:
        - **Title**: The movie's name (excluding ranking numbers).
        - **Year**: The release year.
        - **Rating**: IMDb rating (out of 10).
3. **Handling Dynamic Content**
    - WebDriverWait was used to ensure elements were loaded before extraction.
    - Try-except blocks handled missing or changed elements to prevent script failures.

Data Cleaning & Preprocessing

Once the raw data was extracted, the following cleaning steps were applied:

Handling Missing Values

- Some movies had missing values due to incomplete data on the IMDb website.
- Missing **titles**, **years**, or **ratings** were identified and removed to ensure dataset consistency.
- If a value was partially missing but could be inferred, it was filled accordingly.

Data Formatting and Standardization

- **Year values** were converted to integers to facilitate chronological analysis.
- **Ratings** were converted from string format to floating-point numbers to enable numerical computations.
- **Movie titles** were cleaned to remove extra spaces or unnecessary symbols.

Removing Duplicates

- Duplicate entries were checked and removed to prevent bias in the analysis.
- If multiple entries existed with slight variations, only the most recent or complete entry was kept.

Loading and Validating Data

- The cleaned dataset was loaded into a Pandas DataFrame.
- A preliminary check was conducted to ensure:
    - No missing values remained.
    - The dataset contained unique and relevant records.
    - The numerical values (ratings and years) were in valid ranges.

Final Dataset & Summary

Obtained clean Data set that contain Title of the movie, year and rating which has 100 rows *3 columns.

Data Insights

- The average IMDb rating for the top 100 movies is around 8.7.
- Most top-rated movies are from the 1990s and 2000s.
- The oldest highly-rated movie is from 1957.

Challenges & Future Improvements

Challenges Faced

1. **Handling Dynamic Content**
    - IMDb's webpage content loads dynamically, requiring explicit wait times and strategic element selection.
2. **Bot Detection**
    - The website employs bot detection mechanisms, requiring the use of a **user-agent string** to mimic a real browser user.
3. **Website Structure Changes**
    - IMDb occasionally updates its webpage layout, making it necessary to update XPath locators in the script.
4. **Handling Incomplete or Missing Data**
    - Some movie entries lacked full information, leading to manual or programmatic handling of missing values.
5. **Duplicate and Inconsistent Data**
    - Some entries had slight variations in formatting, requiring additional cleaning to maintain data integrity.

Future Enhancements

- Scraping more metadata (e.g., genres, cast, directors) to enhance analysis.
- Sentiment analysis on IMDb user reviews to understand audience perceptions.
- Automating periodic scraping for real-time updates.

Conclusion

This project successfully scraped, cleaned, and analyzed IMDb Top 100 movie ratings. The cleaned dataset is now suitable for further statistical exploration and visualization. Future improvements can expand on this foundation to gain deeper insights into IMDb ratings and movie trends.