Takeo

Project report:

IMDb Movie Ratings Scraping and Cleaning

Submitted by

Aakriti K C

12/10/2/2024

Objective

The main objective of this project is to scrape data from IMDb’s website to prepare a data set of “Most popular music” by extracting ratings and reviews using scraping techniques. The scraping of data is done to later clean and process them to ensure they are accurate, consistent and ready for further analysis and visualization.

Introduction

This project is designed to demonstrate the practical application of web scraping and data cleaning techniques using Python. The primary objective was to extract movie details, ratings, and other related information from IMDb, clean the data to remove inconsistencies, and prepare a well-structured dataset. The structured data generated from this process can provide insights into movie trends, ratings, and genres.

Process Overview

1. Data Scraping

Goal:

Scrape IMDb for the "Most Popular Movies" chart and extract:

- Movie Titles
- IMDb URLs
- Descriptions
- Ratings (Best, Worst, Current)
- Genre
- Duration

Tools and Libraries Used:

requests for sending HTTP GET requests to fetch the webpage.
BeautifulSoup from bs4 for parsing HTML.
json to handle JSON-LD structured data embedded in the webpage.

Steps Taken:

1. A GET request was sent to the IMDb "Most Popular Movies" page using requests with appropriate headers to mimic browser behavior.

2. The HTML response was parsed with BeautifulSoup.

3. Data embedded in JSON-LD format was extracted from a <script> tag containing structured movie data.

4. Relevant fields were processed into Python lists and converted into a pandas DataFrame.

5. The raw data was saved to a CSV file (most_popular_movies.csv).

2. Data cleaning

Goal:

Ensure the dataset is free from missing values, duplicates, and inconsistencies.

Tools and Libraries Used:

pandas for data cleaning and manipulation.

Steps Taken:

1. Inspection:
Inspected the raw dataset for missing values and duplicates.
Checked data types and distribution of values using info() and head() functions.

2. Handling Missing Values:
Replaced missing values in the Description column with "No description available."
Filled missing values in the Genre column with "Unknown."
Dropped rows where critical fields (Title or Rating Value) were missing.

3. Removing Duplicates:
Identified and removed duplicate rows to ensure data integrity.

4. Final Dataset:
Saved the cleaned data into a new CSV file (cleaned_most_popular_movies.csv).

Challenges and solutions

1. Handling Dynamic Web Pages

   Issue:
   The IMDb page contained embedded JSON-LD data, not directly accessible through conventional HTML tags.

   Solution:
   Used `BeautifulSoup` to locate the `<script>` tag with JSON-LD structure and processed the embedded JSON data to extract required fields.

2. Missing and Inconsistent Data

   Issue:
   Missing values and placeholders like "N/A" in certain fields could lead to inaccuracies in analysis.

   Solution:
   Standardized missing values using `pandas.NA`, Filled non-critical missing fields with meaningful defaults (e.g., "Unknown" for genres) and Dropped rows missing critical fields.

3. Managing Data Integrity

   Issue:
   Duplicates and inconsistent data formats could skew analysis.

   Solution:
   Identified and removed duplicates using `pandas.drop_duplicates()`.
   Re-inspected the cleaned data using `info()` and sample outputs.

4. Future Deprecation Warnings

   Issue:
   Deprecation warnings related to `inplace=True` operations in `pandas` were encountered.

   Solution:
   Updated the code to align with `pandas` 3.0 standards by explicitly re-assigning modified DataFrames instead of relying on `inplace` operations.

Results

Raw data

The data consists of over 200 rows extracted from the IMDB's website which includes the movies' titles, urls, description, ratings, genres and durations.

Cleaned dataset

The final cleaned dataset that has been compiled at the end of the project has gone through the following cleansing process:
- Missing and inconsistent values were addressed.

- Duplicates were removed.

- The final dataset was saved as `cleaned_most_popular_movies.csv` and is ready for analysis or visualization.

Conclusion

This project successfully demonstrates the process of web scraping, data cleaning, and preparation for analysis. The final dataset provides a robust foundation for further exploration and insights, such as analyzing rating trends or genre distribution. The challenges encountered were addressed effectively, ensuring the integrity and usability of the dataset.

Recommendations for future work

1. ## Expand Scraping Scope:

   Include additional IMDb pages to gather data for a larger set of movies.

2. ## Enrich Dataset:

   Incorporate metadata such as release year, director, and cast information.

3. Automate Updates:

   Develop a scheduled script to periodically fetch and update the dataset.

4. Data Visualization:

   Utilize tools like Matplotlib or Seaborn to generate visual insights such as rating distributions and genre trends.

References

- IMDb Website: https://www.imdb.com/chart/moviemeter/
- W3schools: https://www.w3schools.com/python/pandas/default.asp