Major Project -2

This web scraping project aims to extract IMDb ratings, movie titles, and release years from the Top 250 movies list. By following the outlined steps, we will use Scrapy to efficiently scrape and organize the data, ultimately saving it into a CSV file for analysis.

Guidelines for Web Scrapping

Pick a website and Set Up Scrapy Project:

1. Create a new Scrapy project using the command: **scrapy startproject imdb_scraper**
2. Navigate to the project directory: **cd imdb_scraper**

Create a Spider:

1. Inside the spiders directory, create a new spider, e.g., **imdb_spider.py**.
2. Define the spider to start scraping from the provided in the Assignment.

Inspect the HTML Structure:

1. Use the browser's developer tools to inspect the HTML structure of the IMDb page.
2. Identify the HTML tags and attributes containing the required information (ratings, titles, release years).

Extract Information with Scrapy:

1. In the spider, use XPath or CSS selectors to extract the relevant information from the HTML.
2. Populate the item with the extracted data.

Save Data to CSV:

1. Modify the spider to save the scraped data into a CSV file.
2. Define the CSV export format (e.g., columns for rating, title, release year).

Execute the Scrapy Spider:

1. Run the Scrapy spider using the command: scrapy crawl imdb_spider -o output.csv.
2. Verify that the CSV file is created with the extracted information.

This comprehensive web scraping project aims to extract IMDb ratings, movie titles, and release years from the Top 250 movies list. By using Scrapy and following the outlined steps, the data is efficiently scraped, organized, and saved into a CSV file for further analysis. Utilizing Visual Studio Code ensures an organized development environment, and sharing the project on GitHub allows for collaboration and version control.

Web Scraping Project " Titles, IMDb ratings & Release Year"

This is the project is to scrape IMDb's Top 250 movies list and extract information such as movie ratings, titles, and release years. The extracted data will be saved into a CSV file for further analysis.

Project Idea:

Scrape the data out of the website and save the data in the CSV format.
WEBSITE.

Recommended Web Scraping Tool:

Scrapy is a Python-based open-source web crawling framework, ideal for scalable data extraction. With modular design, asynchronous operations, and support for XPath/CSS selectors, it efficiently navigates websites. Features like item pipelines, respectful crawling, and an active community make Scrapy a powerful choice for ethical and effective web scraping.

Submission: The Entire assignment should be submitted by (Sunday, 08/12/2024). You can use VS Code or Google Colab. Upload the Project Folder and the CSV file in a repository in your GitHub Account.