Ivy Homes SDE Intern Assessment
Problem A
Readme

Authors

Vidish Chandekar cvvijay20@iitk.ac.in

Jan 4, 2024

1 Problem A: Workflow

1. My general idea was to scrape through the website's (**IMDb**) front-end **HTML** codes for relevant information. This included finding urls for different genres, getting information of 20 movies for each genre and getting 10 reviews for each movie. Other than these there are also some extra information like Cast of Actors, Directors, Writers, Ratings, etc.

2. I have used modules like **requests** for making url requests and **BeautifulSoup** for scanning for information through the HTML I got.

3. The starting url was https://www.imdb.com/feature/genre#movie as it has a list of genres readily available for scrapping. Within I scanned through all genres and stored them by their **Genre Name: URL** in a dictionary named *refs*

4. Then I scanned through refs and for every genre created a thread which would independently run. Each thread would request information from the respective genre URL and then scrape for relevant information. So essentially every thread would run the function *scrapping_genre*.

5. After all threads are finished, they will write their output in their respective **CSV** file. So every thread will independently create a *genre.csv* file for their respective genre and dump all the relevant information there. So in total there will be around 23 csv files dumped. Each one having headers as *Title,Ratings,Cast,Directors,Writers,Plot Summary,User Review 1,User Review 2,User Review 3,User Review 4,User Review 5,User Review 6,User Review 7,User Review 8,User Review 9,User Review 10*

2 Setup

1. Download the file *final.ipynb* from my github repository and extract it wherever needed.

2. Make a new python environment with python version as 3.11 and add the following packages **requests** (*pip install requests*) and **BeautifulSoup** (*pip install beautifulsoup4*). Other required packages will already be installed by default.

3. Using a terminal or a text editor (VS Code or Jupyter Notebook) run the script using this particular environment.

4. It will take around 5 minutes to run before dumping the csv files for each genre. I have added print statements so that every thread outputs the number of entries done. It will go from 0 to 19.

3 Challenges Faced

1. The usual challenge was finding the data from the HTML and storing and writing it in csv file properly and efficiently. There were also some edge cases I faced including some genres not having any movies, some movies not having writers/directors, or not having enough reviews.

2. Initially I wrote a standard loop which scans all genres (instead of making multiple threads) but this takes quite a long time (around 45-50 mins). Multi-threading helped reduce time significantly to less than *5 mins*.

3. I also tried converting *.ipynb* to a normal *.py* file but there were some logistical issues as a result it wasn't writing in the csv files. Hence I ended up submitting the *.ipynb* instead.

4. There might be some minor errors encountered while running (such as connection time-out/index error, etc). If so happens you can rerun the script and it will work fine (I've tried so and it worked)

5. I haven't used databases at all. This was partially because I never used them in general as well as the fact that after scrapping and cleaning the data I didn't have enough time left to study Databases and implement it in the program. I realized since the Assignment was about information extraction and output, I could do it without the help of databases.