1/21/2019

Shravan Ds

E-mail: [shravands69@gmail.com](mailto:shravands69@gmail.com)

Ph.no: 7760028716

IMDB movie analysis

Data analysis on top 250 imdb rated movies

Scrapping Algorithm:

The scrapping program is written on python using libraries BeautifulSoup and urllib, the algorithm has been formulated accordingly.

1. Extraction of contents along with the html elements from the given link (IMDB Top rated movies).
2. Going through the extracted page contents looking for the classes which contain the individual movie page link, getting the link using the class and storing it in empty python list called as movie-list.
3. Iterating over movie-list and extraction of html page for each link stored in the movie-list.
4. Start of data extraction from the obtained individual movie page.
5. Extraction of movie name which was found in first h1 tag of the page.
6. Similarly extraction of release year, rating, movie length, no of reviews and movie summary using respective class and html element.
7. Since we have more than one genre for some movies the repetition of the step will be done for the class which contain the genre values and extraction of top 4 genres (the value will be stored as null if genre is not present).
8. Repetition of step 7 to get the keywords for a movie.
9. It was found that budget, Production Company, Gross USA, cumulative gross movie length, sound quality and release date were present under instances of class “txt-block” which also has a h4 tag containing the label.
10. The iteration was done on the each instances of class “txt-block”, the data label is validated for the required variable and then the data is stored in the variable.
11. Repetition of steps 7 and 10 to get stars list and writers list which was found under class “credit\_summary\_item”.
12. Writing the obtained data into the CSV file.
13. Steps 3 to 12 will be performed by a function which will be passed with each movie link from movie-list.
14. End of algorithm, once the program is run we will get a csv file containing the data for all the top 250 movies.