Project 5.15. Douban Movie and IMDb web scraping

Basic background: Douban Movie and IMDb are two famous websites providing website viewers with the related information and score rated by audience for numerous released movies. This project aims to scrape the information about movies from these two websites respectively.

Website link:

Douban Movie: <https://movie.douban.com/top250>

IMDb: <http://www.imdb.com/chart/top?ref_=nv_mv_250_6>

Requirement: submit two pieces of code to realize the following function:

1. Enter the movie that you want to search in either code.
2. Each code returns one csv file with the following information from the corresponding website:
   1. The name of movie
   2. The release date
   3. The director(s) of the movie
   4. The main actors/actresses of the movie
   5. The average rating score
   6. The total amount of people who rated the movie
   7. The brief introduction of the movie
3. If entering a list of movies, each code returns multiple csv files.
4. [Bonus1] Write two pieces of codes to return two csv files containing above information of top 250 movies from two websites respectively.
5. [Bonus2] Write another piece of code to merge the data by movie names in two csv files generated from requirement 4 into a combined file. With the help of the combined file, the difference in the rating scores of the same movie listed in top 250 movies in two websites can be easily seen.

Codes will be graded based on:

1. The format of the codes
2. The organization of the csv file
3. The speed of the codes
4. The extendibility of the codes

Estimated completion time: 40 hours + 10 hours + 10 hours. Time consumption depends on the codes’ quality.

Suggestive points:

Deadline:

Bonus points:

1. Work completion before the deadline will generate an extra 20% bonus.

Penalty:

1. Late delivery/missing deadline will incur -10% of finally determined points of the whole project per missing day.
2. Serious mistakes that made the submitted results not useful will incur -120% of the whole project.