Project 5.47. Goodreads Web Scraping

Contractor: LEI Tong 3035329778

Requirement:

1. The program scrapes Goodreads.com for up-to-date information on upcoming movie releases in a certain country/region. First, prompt the user to input a keyword to search in Goodreads.
2. Scrape the search result page, e.g. <https://www.goodreads.com/search?page=1&q=song+of+ice+and+fire&search_type=books&tab=books&utf8=%E2%9C%93>. Retrieve from the html the number of book results returned and the number of pages. If there are no results, ask the user to enter another keyword and search again.
3. If there is more than one page of results, ask the user how many pages of them he or she would like to be presented.
4. Scrape the web page of each book in the search results within the pages of interest e.g. <https://www.goodreads.com/book/show/13496.A_Game_of_Thrones?from_search=true>, using a randomly chosen user agent as disguise to avoid being blocked. Retrieve the following information, where available, on the book by analyzing the html codes:
   1. Title
   2. Authors
   3. Rating Score
   4. Number of Ratings
   5. Description (complete)
   6. Book Format
   7. Publishing Information
   8. Link
5. In addition to the above, scrape the first page of book review section, i.e. retrieve the full contents of up to 30 reviews on the book.
6. Arrange all the information of the upcoming movies properly in a DataFrame. Display the DataFrame.
7. Ask the user if he or she would like to export the DataFrame to a .csv file, and do so if requested.

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: 12 hours. Time consumption depends on the codes’ quality.

Suggestive points: XXX points + bonus.

Deadline: 23/12/2017 23:59:59 pm

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.