Webscraping books data from Goodreads

Project for Webscrapping and Social Media scraping course

Authors:
Monika Kaczan, student ID: 410998
Katarzyna Jałbrzykowska, student ID: 220763

Topic and the web page

The main aim of this project is to scrap data on the most upvoted books on the goodreads.com website using Beautiful Soup, Scrapy and Selenium.

Goodreads is a social cataloging website which stores data on books, annotations, quotes and reviews. Its users can sign up to generate their own library catalogs, reading lists, book suggestions and more. In this project we used the Best Books Ever list which contains, as the title suggests, best books ever published in terms of number of votes and average rating.

For each book we scrap its title, author, publication date, publisher, number of pages, main genre, number of ratings and average rating. Based on the data we obtained we conduct a simple analysis and compare performance of all three scrappers.

Scrapers' mechanics

Mechanics of all three scrappers are pretty similar.

Firstly, we access the main site where we can find links to individual book pages. The complete list 'Best Books Ever' is divided into 100 pages with each containing 100 books. Therefore, we decided to limit our scraping to the first 3 pages. However, the variable 'pages' can be changed to a larger number if desired. After accessing the pages we scrap the links for individual book pages in a list (in Scrapy, we additionally export it to csv so it can be used by the next spider).

We access them one by one. Here, the Selenium script has one additional step: each time the browser is navigated to the first book page from the list, we need to dismiss the login pop-up. Next, in each book page, all scrappers search for a title, author, main genre, year, publisher, number of pages, number of ratings and average rating using commands and regular expressions. We store this information in a dictionary, then add

it to the dataframe as a new row. In the end we export the dataframe to the external *.csv file for further analysis.

Output and comparison of scrapers' performance

Firstly, the time it took to scrap the data varied between the scrappers.

Scrapper	**The time it took to scrap 300 pages**
Beautiful Soup	19 min 32 s
Scrapy	27 s [sprawdź może czy to dobrze]
Selenium	22 min 21 s

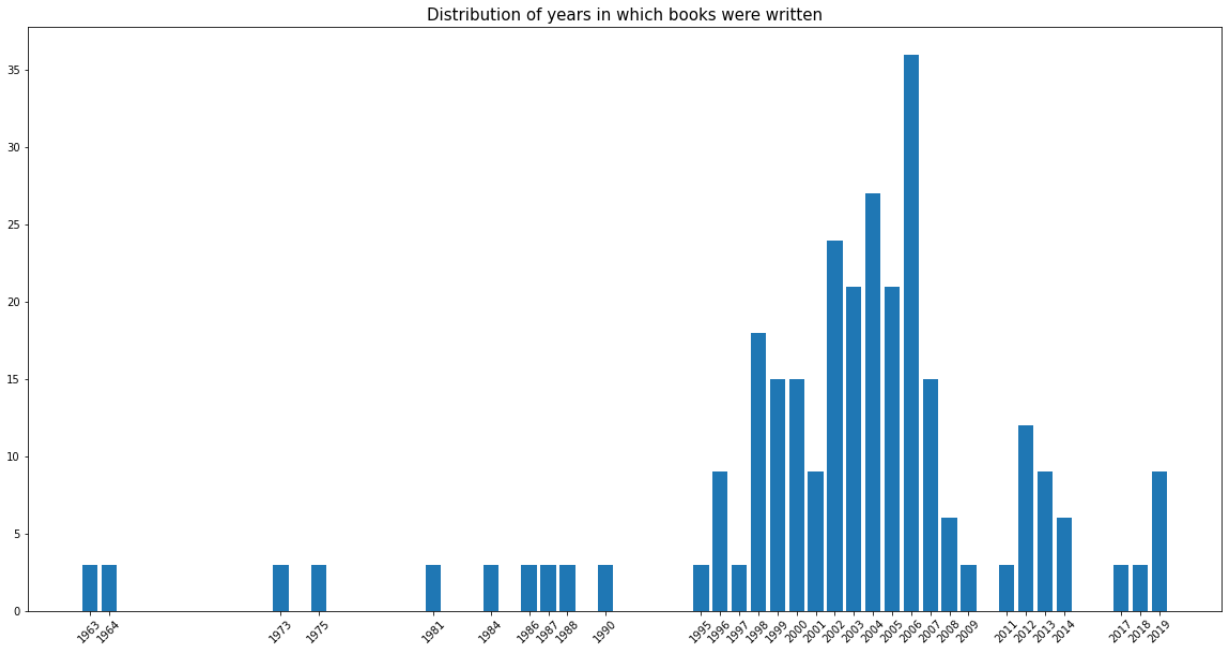
Beautiful Soup and Selenium took similar amount of time (taking into account differences in one's computer and processing power at the moment) while Scrapy turned out to be much faster than the other two.

In terms of data content, all scrapers provided the same results. There were no problems with missing data or incomplete values. However, the order in which Scrapy accessed pages with individual books from the list seemed random. There was no such problem with Beautiful Soup and Selenium which accessed pages in order from the highest on the main list page to the lowest.

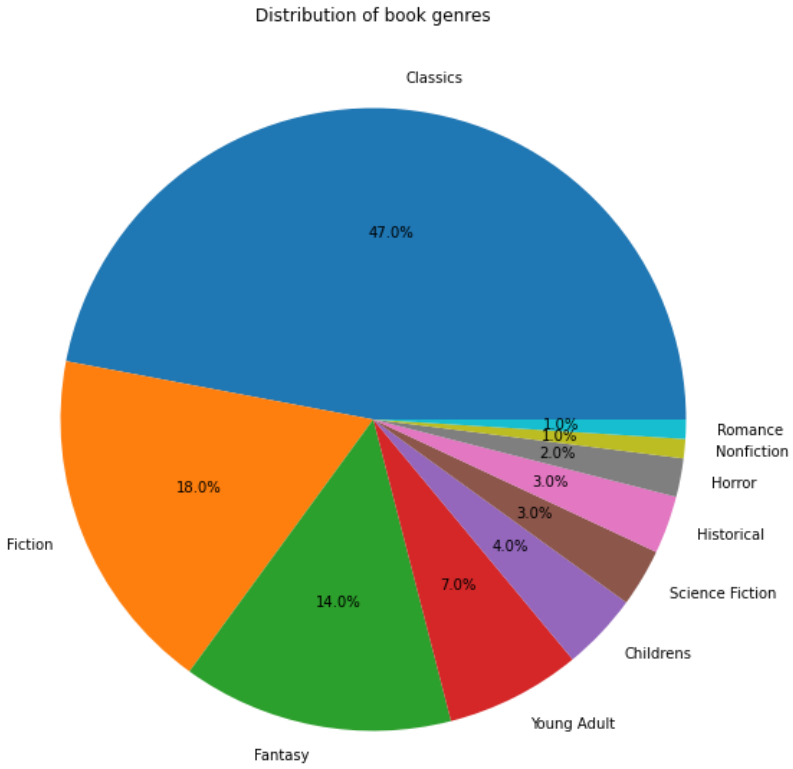
Data analysis

We performed a simple analysis of the data we obtained.

Majority of the books we scrapped were written in the twenty-first century. The year 2006 was exceptionally abundant with 36 books written in that year.



The most popular main genre among the books we analyzed was Classics which is rather predictable as most books on the list were considered canon. What is interesting, a large proportion of the books were Fantasy and Young Adult books.



Additionally, we found that an average book has 445 pages. The longest book on the list is The Hobbit and Lord of the Rings Boxed Set by J.R.R. Tolkien with 1728 pages. The shortest book is Where The Wild Things Are by Maurice Sendak with 38 pages.

Work division

Monika Kaczan was responsible for the Beautiful Soup scrapper, analysis, and this description file. Katarzyna Jałbrzykowska was responsible for Scrapy and Selenium scrappers. We also helped to check and debug each other's codes.