Web Scraping and Social Media Scraping

Evgeniy Alkhovik 448001
Anastazja Olszewska 356332

To complete the project a web site of movie library was selected (http://www.films101.com/index.htm). According to the description of the web site it comprises more than 11,000 movies with their description and streaming services availability. On the http://www.films101.com/years.htm page movies by years are presented starting from 1891 year.

Each of three python scripts scrap this web page and import movie info (title, year, director, country, media) in .csv format.

First implementation based on using BeautifulSoap library. On the first stage from http://www.films101.com/years.htm page a list with years of movie releases was extracted. Then, this list is passed to function which scraps each web page comprising movie list for one year (f.e. http://www.films101.com/y2022r.htm is a web page with films casted in 2022 year). Last function of the script safe the obtained data to csv file.

Secondly scrapy framework was used to get data in same format. Logically we can split it to three part. Firstly, from http://www.films101.com/years.htm web page csv file with webpages of movies list by year was returned. Reading this file line by line web page of each movie was observed and stored in csv file. Finally, class tableSpider read this file and extract required information.

Finally Selenium framework with ChromeDriveanager was used. Initially, with access to http://www.films101.com/years.htm via ChromeDriveanager list of movie release years was obtained. Iterating over this list ChromeDriveanager object was created to visit movie by year page. And from each page with XPATH language movie information was extracted and saved in csv file.

Output csv file comprises 5 columns:

- Movie title
- Release year
- Director
- Country
- Media
    - T Theaters (US)
    - B Blu-ray Disc
    - D DVD
    - V VHS
    - Y Yes (DVD or VHS, unspecified)
    - S Soundtrack

The main contributor in cinematography is USA over the whole history of it. Rate of European movies is decreasing for the last 30 years. Each decade the amount of produced movies increases more than 15 percent.

Distribution of the work:
Anastazja Olszewska was responsible for finding the topic of the project and writing Beautiful Soup scraper.
Evgeniy Alkhovik was responsible for writing Scrapy scraper.

Selenium scraper was written together: page with years was parsed by Anastazja, films were scraped by Evgeniy.