Web Scraping Project

Łukasz Pieńkowski 443734

Artur Nowak 397246

Description of the topic and the web page:

Our goal was to scrap data about top rated movies from IMDB site. Output for every movie contains 4 factors: title, user ratings, popularity score and genre.

IMDb is an online database of information related to films, television series, etc. – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews

Description of scrapers mechanics:

BS_scraper – at first it scrapes extension of link (https://www.imbd.com) for each of the first 100 (if variable limit_100_pages is true) sites form top 250 list. Then is matches extensions with root link. At the end bs fills titles, ratings, popularity_scores, genres lists with data taken from movie page and compiles list to data frame.

Scrapy: link_list – this scrapy creates list of links and fills it with links for first 100 (if variable limit_100_pages is true) sites from top 250 list.

Scrapy: movie_scraper – role of this scraper is simple, it takes link for each page from link_list.csv and fill a new csv file with data (title, IMDB_rating, popularity, genre) for each movie.

Selenium_scraper – at first it creates a list of links to first 100 (if variable limit_100_pages is true) sites from top 250 links and then i headless form it go through every single page and scrap title, rating, popularity score and genres. An output is presented in form of data frame with data for every movie.

Description of the output:

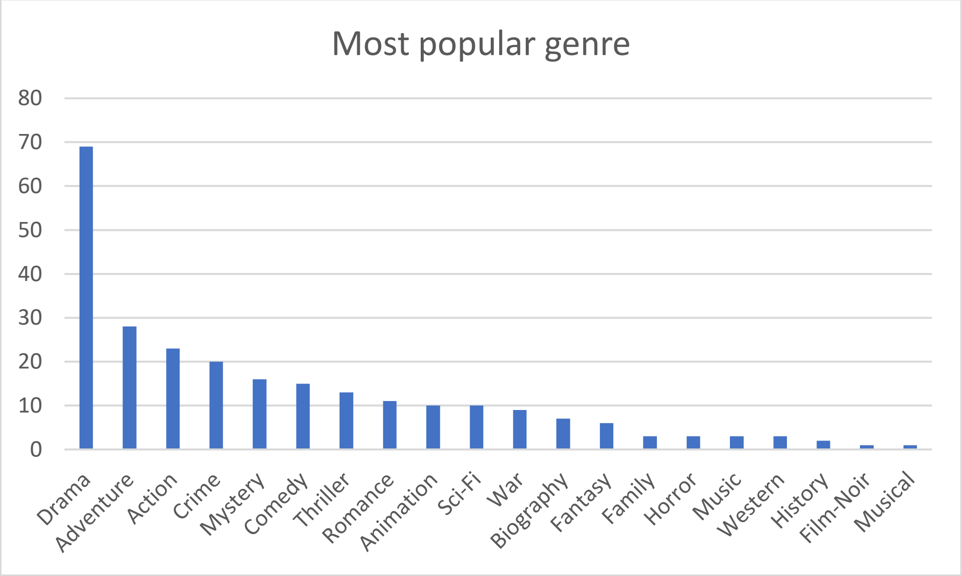
Every scraper give the same output. It is a table (in case of using Scrapy) or dataframe (output of BS or selenium scraper) which contains: Title of the movie, IMDB rating, Popularity score and the Genre of Genres of the film.

Data Analysis:

Based of the collected data we can find out what is the average score of top 100 movies in IMDB.

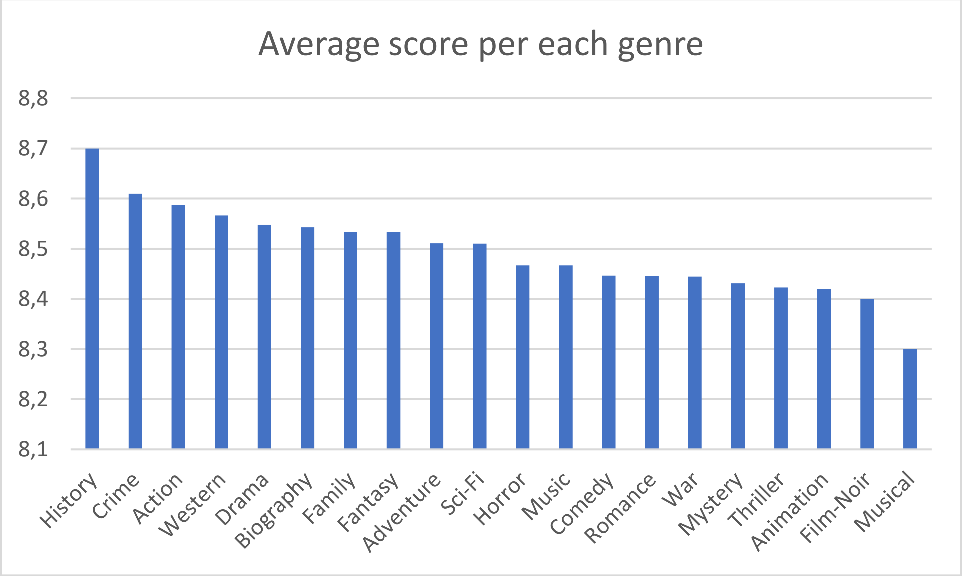
- Average score: 8,5

What is more we can create a chart to see what is the most popular genre of these movies



As the results show the most popular genre is Drama, which is featured in 69 movies. Then other results are not much different. There is Adventure with score 28 on the second place, Action with 23, Crime with 20 etc. The least popular genres are Musicla (1 movie), Film-Noir (1 movie), History (2 movies) and Western/Music/Horror/Family (3 movies).

The average popularity score of top 100 movies in IMDB is equal to 266,4.



This chart presents average score for each genre. As we can see the most rated type of movies are history movies or partly-historical (we need to remember that the movie can have more than one genre). What is more the least graded genre is Musical. So we can say that in 100 top rated movies, the musical ones are the worst (according to users rating).

Role Description:

BS_scraper – Łukasz Pieńkowski

Scrapy scrapers – Artur Nowak

Selenium – Łukasz Pieńkowski, Artur Nowak

Raport – Łukasz Pieńkowski, Artur Nowak