

Web Scrapping and Social Media Scrapping - Final Project Description

Short description of the topic and the web page.

<https://hdtoday.tv/> is a Free Movies streaming site with zero ads. They let you watch movies online without having to register or paying, with over 10000 movies and TV-Series. You can also Download full movies from HDToday and watch it later if you want.

The goal of the project is to scrape these movies information's:

- Name
- Release date
- Genres
- Cast
- Duration
- Country
- IMDB

Short description of your scraper mechanics - what the program is technically doing.

Selenium-based scraper:

After the Web Site is opened, it automatically click to the "movies" page, clicks on each movie, saves the necessary information from that page in the established database, and finally converts them to Excel file.

Scrapy:

Here, the first spider collects the links of the movies and records them in Excel, while the other spider collects the information about the movies and converts them into Excel.

BeautifulSoup:

Here, information about each film is obtained and stored in an Excel file.

Short technical description of the output you get.

The goal of the project was to scrape these movies information's and the output is a Excel file with the following columns:

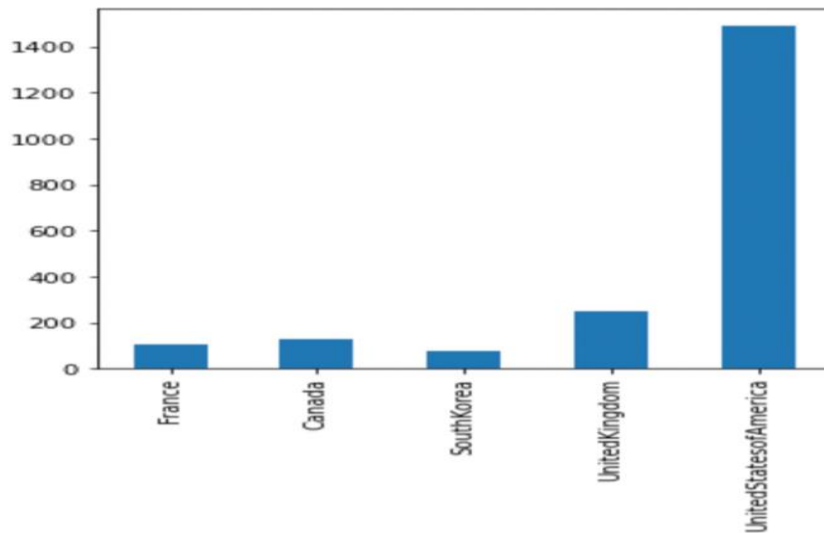
Name	Release date	Genres	Cast	Duration	Country	IMDB
------	--------------	--------	------	----------	---------	------

Extremely elementary data analysis - you need to prove, that collected data can be used for further analysis, but nothing more (hard limit of data analysis: one page).

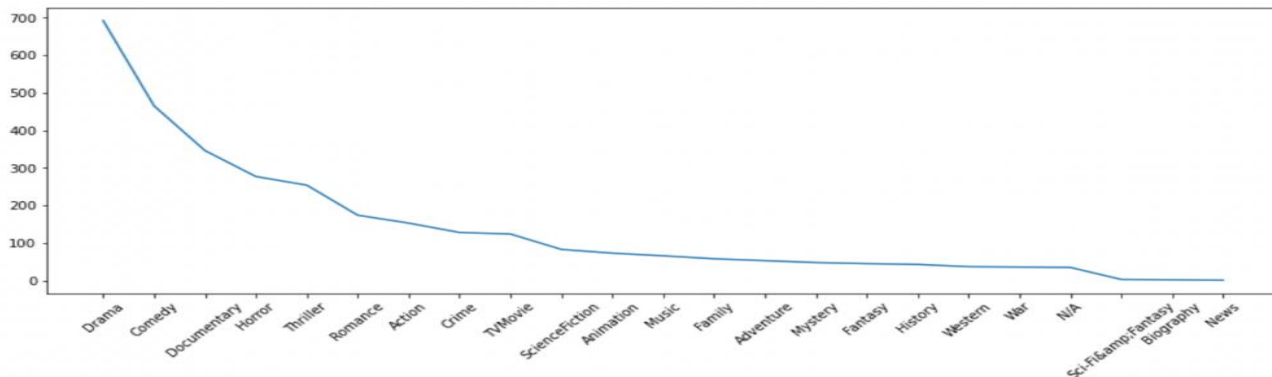
Retro movies have better imdb than the movies that have passed time. Here are the top 5 movies from retro movies:

Name	IMDB
Rashomon	8.2
The Big City	8.2
Don't Cry, Peter	8.2
8½	8.0
Die Nibelungen: Kriemhild's Revenge	7.9

In addition, since it is included in our data in the countries that make the most movies, we can see that according to the data we have, America is recorded as the country that makes the most movies



The most popular genres are drama and comedy.



Detailed description which group participant wrote which part of the project.

The whole project was developed by Elvin Shirinov.