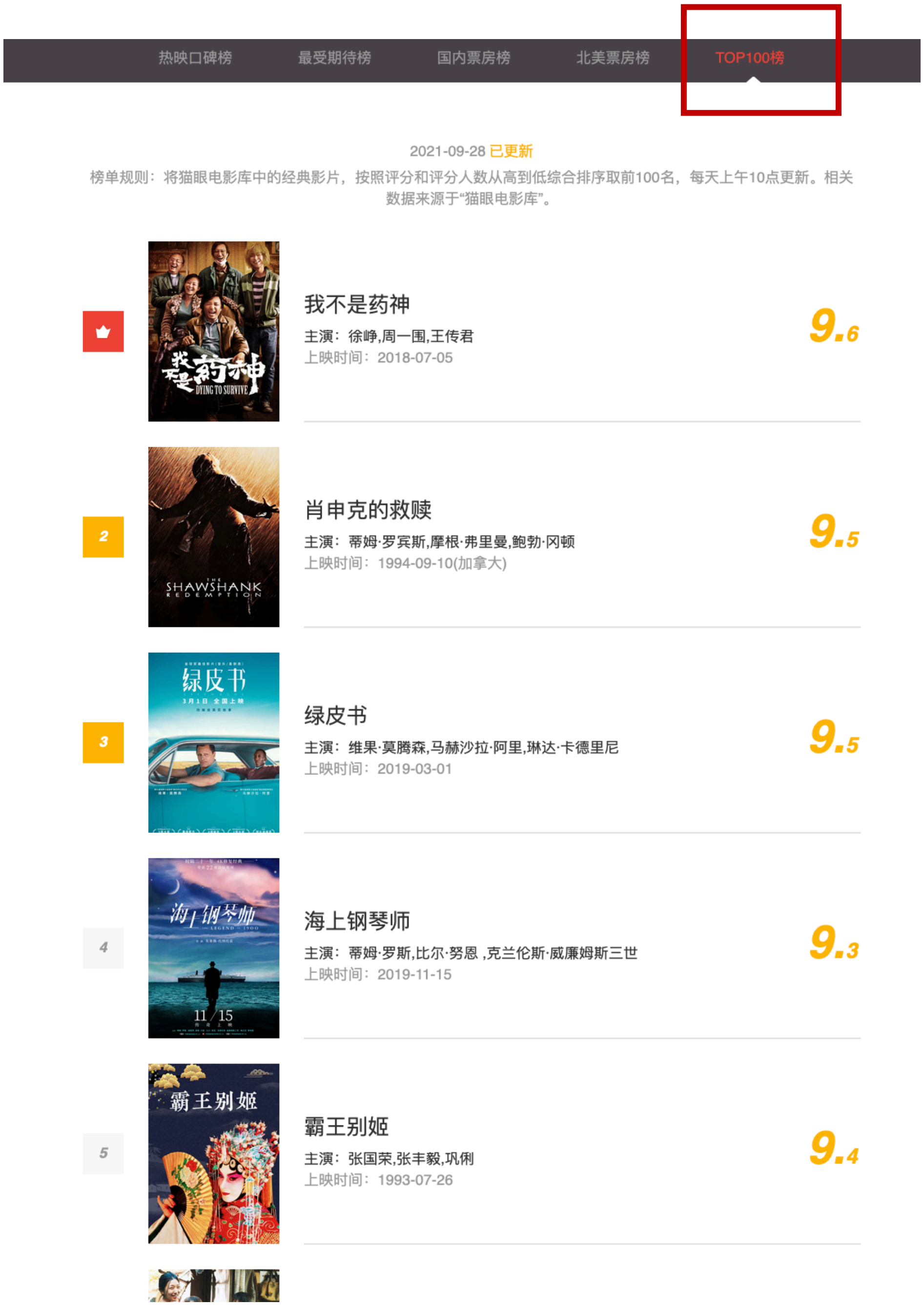
Assignment 1: Web Scraping & Data Analysis

Oct 31, 2021

In this assignment, you should work with data from https://maoyan.com/board/4 (Top 100 Movie)



2021-09-28 已更新

榜单规则：将猫眼电影库中的经典影片，按照评分和评分人数从高到低综合排序取前100名，每天上午10点更新。相关数据来源于“猫眼电影库”。

我不是药神
主演：徐峥,周一围,王传君
上映时间：2018-07-05
9.6

2
肖申克的救赎
主演：蒂姆·罗宾斯,摩根·弗里曼,鲍勃·冈顿
上映时间：1994-09-10(加拿大)
9.5

3
绿皮书
主演：维果·莫腾森,马赫沙拉·阿里,琳达·卡德里尼
上映时间：2019-03-01
9.5

4
海上钢琴师
主演：蒂姆·罗斯,比尔·努恩 ,克兰伦斯·威廉姆斯三世
上映时间：2019-11-15
9.3

5
霸王别姬
主演：张国荣,张丰毅,巩俐
上映时间：1993-07-26
9.4

Maoyan Movies is a favorite viewing platform for domestic audiences and provides you with online ticket purchase services. At the same time, Maoyan Movies also provides you with movie trailers, box office queries, movie rankings, film and television information and other information.

Everyone is interested in bad movies, how can there be so many bad movies? So, it is very important to scrape high-quality movies on various websites. In this project, we will use the requests and regular expressions we have learned before **to scrape the TOP100 movies from Maoyan Movies.**

If you CANNOT read Chinese, please crawl other websites (i.e., https://brickset.com/sets/year-2016). Then report should be written based on your collected data. On the report, please specify which URL you are working with.

Task1. I want you to **scrape the latest 100 movies** from the website and save result into 'your_name+id.csv'. This file should contain the data with the following columns: (40pt)

- Title
- Name of director
- Name of actors
- Rating
- Cumulative income
- Duration
- Type

You are free to explore data **with more properties** if needed.

Task2. I want you to do a data analysis on the data. What do you think is interesting about this data? Tell a story **within 2 pages** (excluding figures) about some interesting thing you have discovered by looking at the data. (60pt)

For example, you might consider whether the director has an impact on movie box office revenue, or whether some directors only focus on making certain type of moves. Another thing you might consider is whether there is a relationship between the box office revenue and the user's ratings on the movie.

The assignment code that runs on its own (Web crawling + Data analysis) should be handed in using a Jupter Notebook file.

Submission Checklist:

Yes/No	**Items**
	Assignment code
	your_name+id.csv
	2 pages Repoort