Xi'an Jiaotong-Liverpool University
西交利物浦大學

School of Advanced Technology

Project 1 Report

Project Title: Web Scraping

Student Name: Jinhao Pang

Student ID: 1824395

Project field: Data analysis

Supervisor:

Co-supervisor (if applicable):

1- Problem Statement

This project aims to scape data from Mayan top100 movies websites and analyze the data accordingly. In this project, some common properties could be found to assist the movies productors to improve the quality of the movies and some humble ideas are provided for further discussion, which mainly focuses on the correlation between the features and the ranks.

2- Analysis from the data

Primary exploration

	rating	rating number	cumulative income	title	title_en	first week income	type	area	duration	time in CN	director	actors	reviews	awards	number of prize	number of nomination	rank
86	9.1	430	NaN	十二怒汉	12 Angry Men	NaN	['剧情', '悬疑']	美国	159分钟	714517	西德尼·吕美特	['亨利·方达', '李·科布', '马丁·鲍尔萨姆', '杰克·瓦尔登', '西德尼·吕美...	[('9', '这部由亨利方达主演的上古大神黑白电影实在是太过屈利，但它在当年奥斯卡上没能拿...	{'第30届奥斯卡金像奖': {'提名': ['最佳影片', '最佳导演', '最佳改编剧本...	6.0	13.0	87

Figure 1 Sample row

Take a sample of the scraped data as example, besides the required data, some extra features such as *cumulative income*, *actors*, *number of prize*, etc., are scraped as well for this analysis, which are store in the pandas as above.

Basic analysis: Type vs area Example

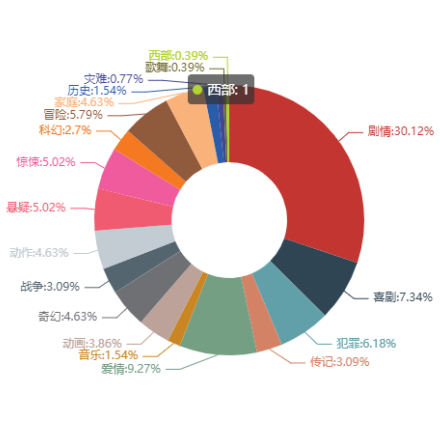

Figure 2 Type distribution

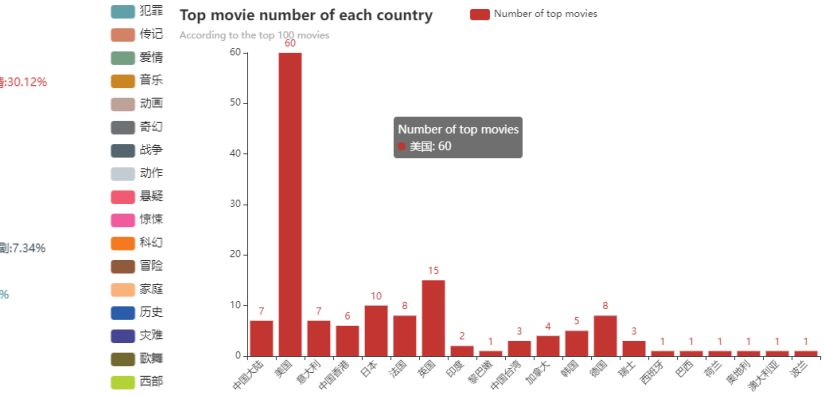

Figure 3 Area distribution

Because of the limitation of the pages, we set on demonstrating by *type* versus *area* as a simple example. Due to the Pie diagram(Fig. 1), the plot and love are the main type audiences pursuit most, while song & dance is of the least attraction. In Figure 2, it indicates that America holds the greatest number of the top 100 movies. Hence, we would be curious if people in different area could have diverse tendency towards the genre of movies(Fig. 3). For example, except the plot, audiences in China tend to watch comedies. Similar to Chinese, Japanese love comedies as

well, with fancies and family types involved in their choices. The other areas are more balanced.

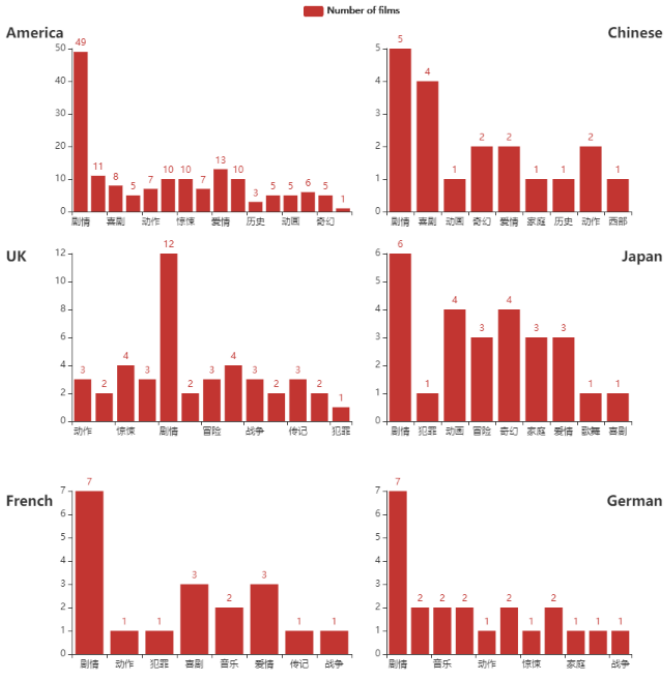

Figure 4 Area tendency

Further analysis: Review as an example

Awards is also another choice for this analysis, as some of which are high gold content for every movie to achieve. Considering the area and tendency of audience, this section continues to explore correlation of audience tendencies in top 100 movies. Figure 4 represents the words with relatively higher frequency in the reviews, whose size of words shows the frequency of the words occurred in the reviews. Clearly, story and life occupy the most obvious position, which means audiences pay attention to most. Surrounded by love, reality, plot, the diagram helps to justify our previous assumptions appropriately.

Figure 5 wordcloud

Explore rank secrets

Data collected from the web are not always useful. In our analysis, for example, *names, release time in CN*, are not as that useful as other features. To reduce the dimensions of the features, we prune some of them. Here, we use *Cumulative income* and *First week income* to explain this.

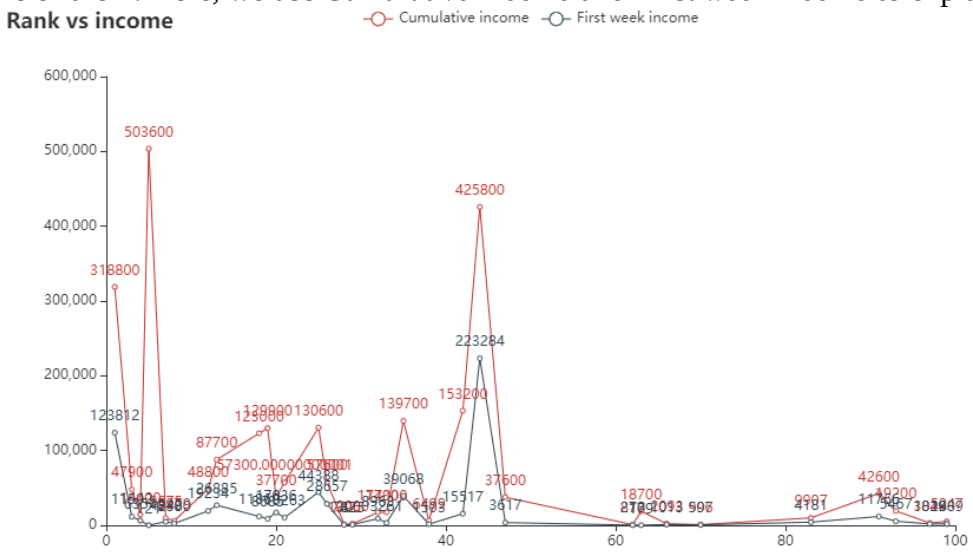

Figure 6 rank vs income

Comparatively, the average income of top 50 movies is higher than the income of the remaining movies. However, it is not absolute. Some old movies' income is much less than the movies nowadays, but the movie still have a high rank(rank 30 versus the movie in the middle who has a little summit. Hence, the income is not a critical factor of the movies. Other features can also be analyzed like this.

Therefore, after preprocessing, we attempt to apply some machine learning algorithms to explore the secrets of the rank. As the limitation of the data size, regression methods are selected to utilized in this analysis.

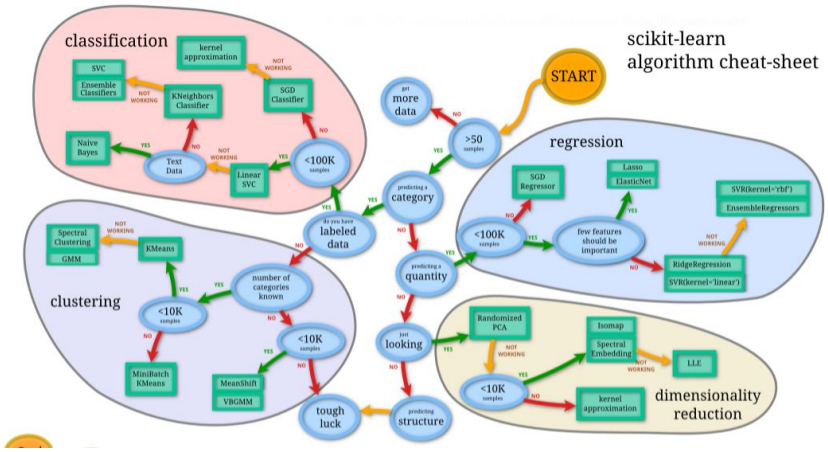

Figure 7 Choosing the right estimator[1]

The coefficients are shown as below:

	rating	rating number	area	duration	time in CN	director	number of prize	number of nomination	喜剧	音乐	...	加拿大	韩国	德国	瑞
lasso	-7.648158	-0.661428	0.122114	1.646673	-1.063192	0.099693	-0.371185	-0.279945	8.68406	-6.113414	...	6.681065	17.414308	-11.824057	16.69435
elastic	-7.601216	-0.663125	0.155219	1.599784	-1.062925	0.101699	-0.368574	-0.27684	8.479876	-5.786519	...	6.221762	17.452829	-12.009427	15.86790
ridge	-4.108698	-1.416888	0.307902	-0.214034	-0.918036	0.079514	-0.499426	-0.051025	0.044468	0.389125	...	1.11474	2.907877	-1.160675	-0.65977

From the machine learning results, although it is quit underfitting, it also enhances our assumptions that area plays an important role in the quality of the movies and *director* counts more than the *price*, etc.

Reference

[1] Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python*[Online].Available: https://scikit-learn.org/stable/tutorial/machine_learning_map/