****

**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.

1. **Analysis from the data**

**Primary exploration**

图表

中度可信度描述已自动生成

Figure 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**

图表

描述已自动生成图表, 条形图

描述已自动生成

Figure Type distribution Figure 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.

图表, 条形图

描述已自动生成

Figure 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.

地图

描述已自动生成

Figure 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.

直方图

中度可信度描述已自动生成

Figure 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.

图片包含 图表

描述已自动生成

Figure Choosing the right estimator[1]

The coefficients are shown as below:

图形用户界面

低可信度描述已自动生成

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/