CSIC 5011-Project 1
Dream of Red Mansion Analysis

Zhijie Yu

20454611
Department of Civil & Environmental Engineering
HKUST

Abstract

Principal component analysis(PCA) and k-means clustering methods are adopted to analyze the character-event grouping problem in the novel Dream of Red Mansion. Through analysis of the data, Jiabaoyu and Wang Xifeng are the two main characters in this novel. And there are mainly three story lines driving the plot development including love story line, management affair line and outer environment line.

1 Preliminary analysis

Dream of Red Mansion is a very classical traditional Chinese novel which tells a story about a noble family. It covers hundreds of distinctive characters and events and meanwhile reflects the cruel social reality at that time in ancient China. It is always seen as a peak in ancient Chinese literary history and well worth revisited from different perspectives. In this project data analysis methods are taken to analyze characters and events in this classic novel.

Dream of Red Mansion has totally 120 chapters. From these chapters character and event information are collected to form a 374×475 matrix, where 374 represents different characters and 475 represents different events. The element value is 1 if the character is involved in the corresponding event, otherwise 0. From basic analysis it can found that some of the characters involve few or even no event due to some collection errors. We thus erase those characters involved in less than 4 events and finally get a 105×475 matrix with 105 characters and 475 events.

The characters are ranked in the descending order based on the number of events they involve shown in 1. Similarly, the events are also ranked in the descending order based on the number of characters they covered shown in 2. The top 10 characters are 贾宝玉, 王熙凤, 薛宝钗, 林黛玉, 史太君, 王夫人, 袭人, 贾琏, 贾政, 平儿 respectively and the top 10 events are 92 回是否出现, 秦氏后事,91 回是否出现, 紫娟激宝玉生病, 藕香榭聚餐, 庆生辰雅座行令, 贾政升官, 搬进大观园, 众亲戚相认入住大观园, 惑奸谗抄检大观园 respectively.

2 Principal component analysis

Principal component analysis is a commonly used feature extraction method in data analysis. Its ultimate goal is to find orthogonal principal component directions so that by projecting the samples onto these directions, these samples can be separated. The detains are discussed in class and will not be explained here.

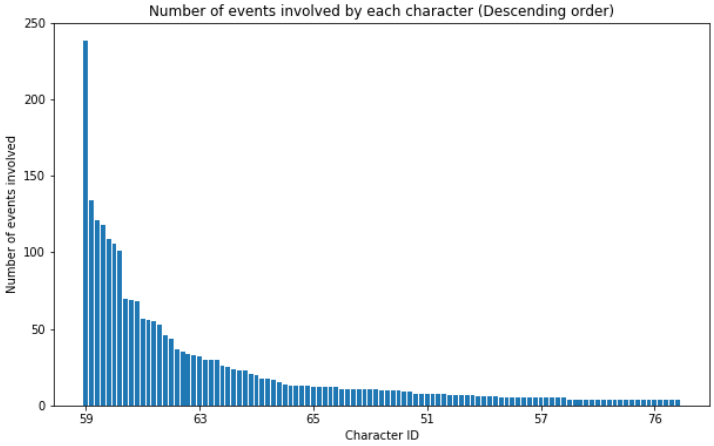

Figure 1: Number of events involved by each character

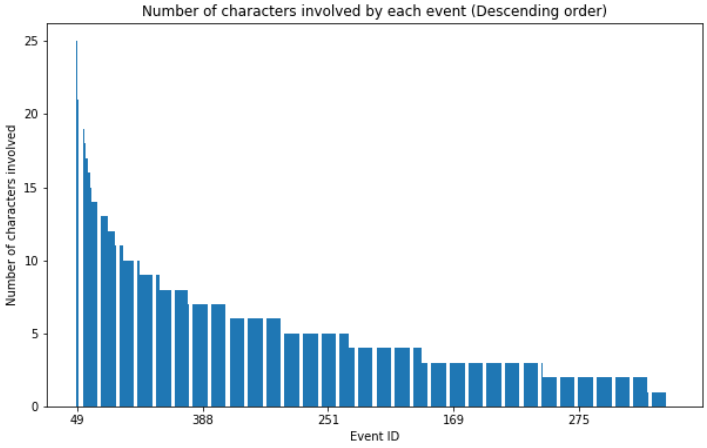

Figure 2: Number of characters involved by each event

2.1 Character analysis

We first apply PCA to the 105×475 matrix in order to separate different characters based on the events they involve.

The explained variance ratio over the principal components are shown in 3. From this figure we can see that the first 2 principal components explain most of the variance and we will further discuss about them in the following part.

Figure 4 also shows the projection of characters on the first 2 principal components. And the orange points are the top 10 characters involving the most events from section 1.

2.1.1 First principal component analysis

From Figure 4 we can find that in the 1st principal component direction, those characters with larger values are well separated from the others. And these characters are mainly those involving the largest number of events. Thus we can conclude that the first principal component mainly separate the main characters from the other supporting roles.

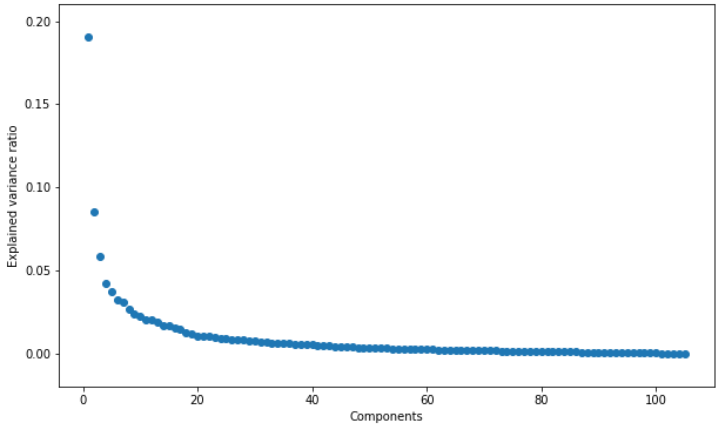

Figure 3: Explained variance ratio of character principal components

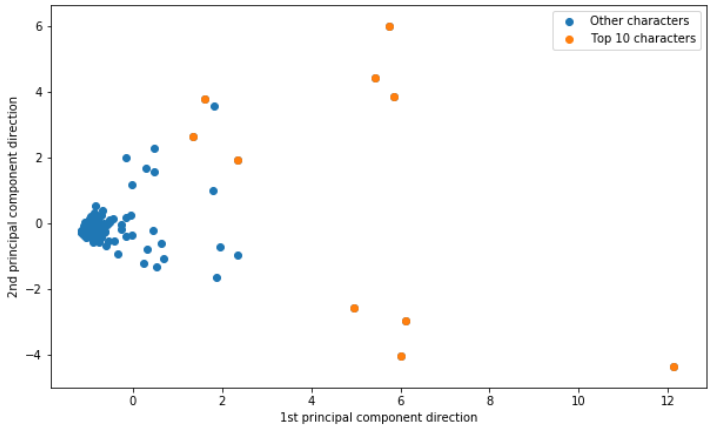

Figure 4: Character projection on 1st and 2nd principal component

And the events with the highest coefficients are also mainly those involving the largest number of characters. This means most main characters participate in main events, which is reasonable for a novel.

2.1.2 Second principal component analysis

In the second principal component direction, the top 10 main characters are separated into two opposite directions shown in Table1, which implies that they may belong to different groups and the core characters of these two groups are 王熙凤 and 贾宝玉 respectively.

2.2 Event analysis

We then apply PCA to the 475×105 matrix in order to separate different events based on the characters they involve.

The explained variance ratio over the principal components are shown in 5. From this figure we can see that the first 2 principal components can also explain most of the variance and we will further discuss about them in the following part.

Table 1: Top 10 characters on second principal component direction

Largest projection value	Smallest projection value
王熙凤	贾宝玉
王夫人	林黛玉
史太君	薛宝钗
贾琏	袭人
平儿	
贾政	

Figure 6 shows the projection of characters on the first 2 principal components. And the orange points are the top 10 events involving the most characters from section 1.

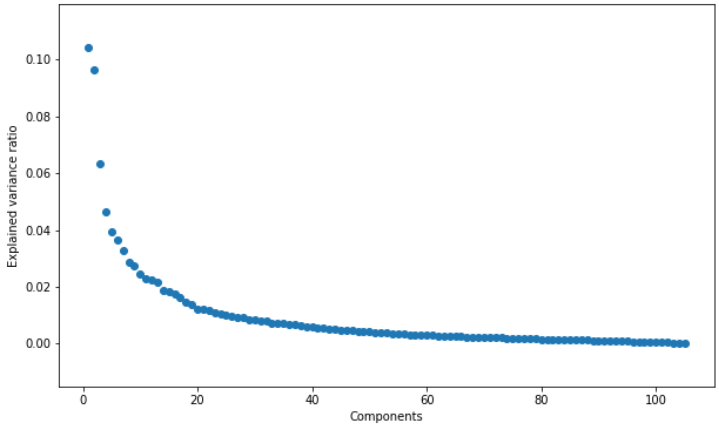

Figure 5: Explained variance ratio of event principal components

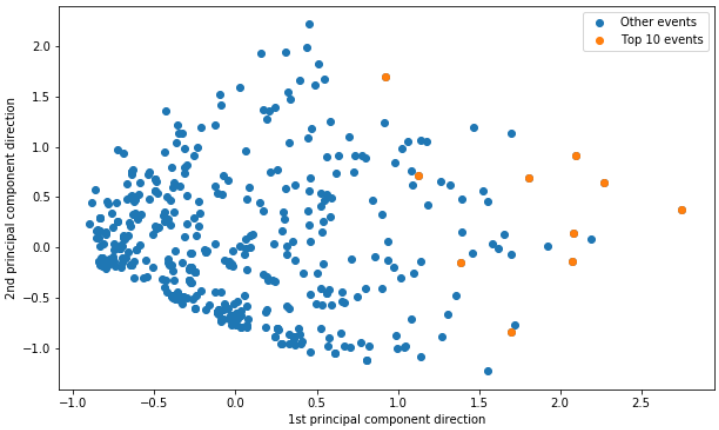

Figure 6: Event projection on 1st and 2nd principal component

2.2.1 First principal component analysis

From Figure 6 we can also find that in the 1st principal component direction, those events with larger values are separated from the others. And these events are mainly those involving

the largest number of characters. And the characters with the highest coefficients are also mainly those involving the largest number of events. This result is consistent with those in section 2.1.1.

2.2.2 Second principal component analysis

In the second principal component direction, the top 5 events with the largest projection values and top 5 events with the smallest projection values are listed in Table 2.

In order to have a better understanding of this, we also list 10 characters with the largest absolute coefficients in Table 3. From this table we can clearly find that these characters forms two groups. 贾宝玉, 林黛玉, 袭人 have negative coefficients while others have positive ones. This result is also consistent with the conclusion in section 2.1.2. Furthermore, these coefficients indicate that in the second principal component direction, those events with larger projection values center on Wang Xifeng's group while those with smaller projection values more likely on Jia Baoyu's group.

Table 2: Top events on second principal component direction

Largest projection value	Smallest projection value
贾琏纳秋桐为妾	诗社作诗
春祭恩赏	黛玉问话宝玉
众人斗牌哄贾母高兴	寻宝玉，宝钗接针线
挑唆张华	放风筝放晦气
凤姐生日凑钱	宝玉作谒

Table 3: 10 Characters with largest absolute coefficients

Character	Coefficient
王熙凤	0.44359011257560904
王夫人	0.35246456220661915
史太君	0.31659719726833724
邢夫人	0.2812301116015962
贾琏	0.26901365037093106
平儿	0.19548053053756675
贾珍	0.1704363301196441
贾宝玉	-0.321553779750935
林黛玉	-0.2416205945946526
袭人	-0.16123923490455275

3 K-means clustering

K-means clustering is also a data analysis method. It is an unsupervised method which aims at dividing the whole samples into different clusters. Here we combine this method with PCA to have a better understanding of the character-event grouping problem in Dream of Red Mansion.

3.1 Character analysis

We use k-means clustering method to divide all the characters into 3 separate groups and use these groups as labels to label the characters on the PCA projection map Figure 4 and get a labelled map Figure 7.

Each character cluster has clear meanings. Cluster 1 includes 贾宝玉, 薛宝钗, 林黛玉, 袭人. Cluster 2 includes 王熙凤, 王夫人, 史太君, 贾琏, 平儿, 贾政 and cluster 3 includes all the other supporting characters.

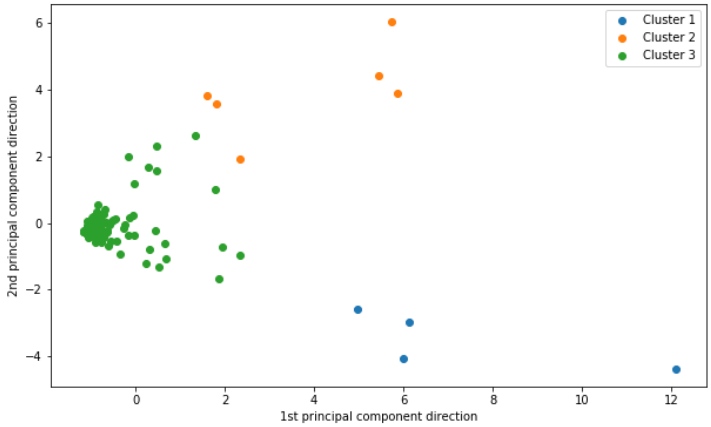

Figure 7: Labelled character projection map

3.2 Event analysis

We also use k-means clustering method to divide all the events into 3 different groups and use these to label the events on the PCA projection map Figure 6 and finally get another labelled map Figure 8.

Based on previous analysis, these clusters also have specific meanings. Events in cluster 1 mainly reflect the effects of outer environment on the big novel family while other events mainly focus on the inner affairs. Events in cluster 2 are mainly the love stories between Jia Baoyu and his lovers. And events in cluster 3 are the daily management affairs centered on Wang Xifeng.

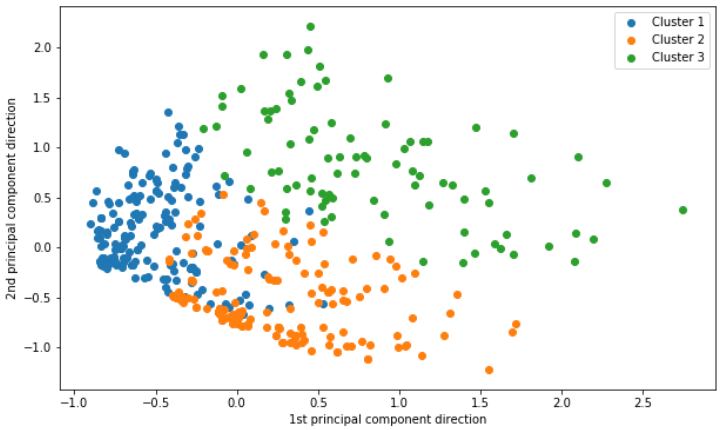

Figure 8: Labelled event projection map

4 Conclusion

By using different data analysis methods including PCA and k-means clustering to analyze Dream of Red Mansion, several conclusions can be drawn. Jia Baoyu and Wang Xifeng are the two main characters in this novel. There are mainly 3 story lines promoting the plot

development. One is love story line centered on Jia Baoyu. Another is management affair line centered on Wang Xifeng. And the final line is outer environment line which depicts the interaction between this big family and the outer environment including the emperor and other novel families.

References

[1] CSIC-5011 textbook. A Mathematical Introduction to Data Science