SI 618 Project Report – Part B

Movie Scripts Analysis

Huanchen Lu

1. Motivation

As a huge movie fan, movie is a way for directors to convey their feelings and values of their own lives, and a good movie scripts also can be remembered by the people for a long time. For this project, I mainly focus on the relationships among the movie scripts, the genre and etc. For the first part of my project, I calculated and visualized the term frequencies and tf-idf weights of movie scripts for different genre. For the part B of my project, I expand my dataset by collecting more information about the directors and movie length and so on.

My four research questions are listed below:

a. Have movie become longer or shorter? Have movie scripts become longer or shorter?
b. Did actors and actresses speak more words per minute than before? How does these vary across different genres?
c. Have movie scripts become "worse" over past years?
d. How movie topics have changed over past years?

2. Data Sources

a. Movie Quote Database (http://www.moviequotedb.com)

Movie Quote Database is a collection of world movie database with the scripts on it. Since this website do not have the API for fetching their data, therefore, I used BeautifulSoup to fetch the movies with their scripts. From the website's Browse All section, I fetched all the movie showed on the list, which is around 1000 movies, in the meantime, I use paging to fetch all the movie scripts from their database. The dataset contained several important variables, including movie title, movie scripts for each scenario.

b. Simple IMDB API (https://github.com/theapache64/movie_db)

After I fetched all the movies with their movie scripts from Movie Quote Database. I used Simple IMDB API to get all the information for each movie, like genre and rating. Simple IMDB API is a platform for easily access the IMDB database, to get all the data for each movie. It returned a JSON format for each movie that I fetched on Movie Quote Database, and it retrieve 985 records in total.

3. Method

Before manipulating the data, I merged and store the data in a database with one-to many relationships.

a. Have movie and movie scripts become longer or short over past years.

1). How did you manipulate the data prepare it for analysis?

I converted the movie length to numeric data type and factorize the movie genres. I calculate the number of words in each script using R.

2). How did you handle missing, incomplete, or noisy data?

For all the missing or incomplete data, I just get rid of them.

3). Challenge

For this part, the script format is very weird. Each scenario movie scripts always comes with a movie title. In other words, it has the duplicated movie title but with the different script. And it really bothered me a long time to find the way to store them to the data frame. Finally, I stored all the scripts information to the data frame and use sparkSQL group to concatenate the scripts with the same movie title.

b. Did actors and actresses speak more words per minutes than before? How does these vary across different genre?

1). How did you manipulate the data prepare it for analysis?

I converted the movie length to numeric data type and factorize the movie genres. I calculate the number of words in each script using R.

2). How did you handle missing, incomplete, or noisy data?

For all the missing or incomplete data, I just get rid of them.

3). Challenge

The same challenge I mention in last part and solve it in the same way.

c. Have movie script become "worse" over past years?

The variables used in this question were movie_scripts, genre and year.

1). How did you manipulate the data to prepare it for analysis?

For this question, I first obtained a list of "dirty, naughty, obscene, and otherwise bad" words from github, which would be used for calculating the "badness" scores of scripts later. And before the analysis, I did some pre-processing for the bad word list and all the scripts using NLTK, including stop words removal and word stemming. Then, I calculated the badness score for each movie as:

$$Badness = log_{10}\left(\frac{count\ of\ bad\ words}{count\ of\ all\ words}\right) + 3$$

And the average badness score for each year was calculated as:

$$Badness = log_{10}\big(mean(badness)\big) + 2$$

2). How did you handle missing, incomplete, or noisy data?

For all the missing or incomplete data, I just get rid of them.

3). Challenge

Everything went perfectly fine, no more challenge in this part.

d. How movie topics have changed over past decades?

1). How did you manipulate the data to prepare it for analysis?

Since I intended to generate movie topics by decades, I aggregated all the scripts by decade and stored the scripts for each decade in separate txt files. The code for this could be found in scripts_decades.py.

2). How did you handle missing, incomplete, or noisy data?

For all the missing or incomplete data, I just ignored them.

3). Challenge

The challenge of this part of analysis was that the topics extracted from scripts were obscure. I tried to tune the hyper-parameters of the LatentDirichletAllocation(), and the result went a little bit better to some extent.

4. Analysis and Results

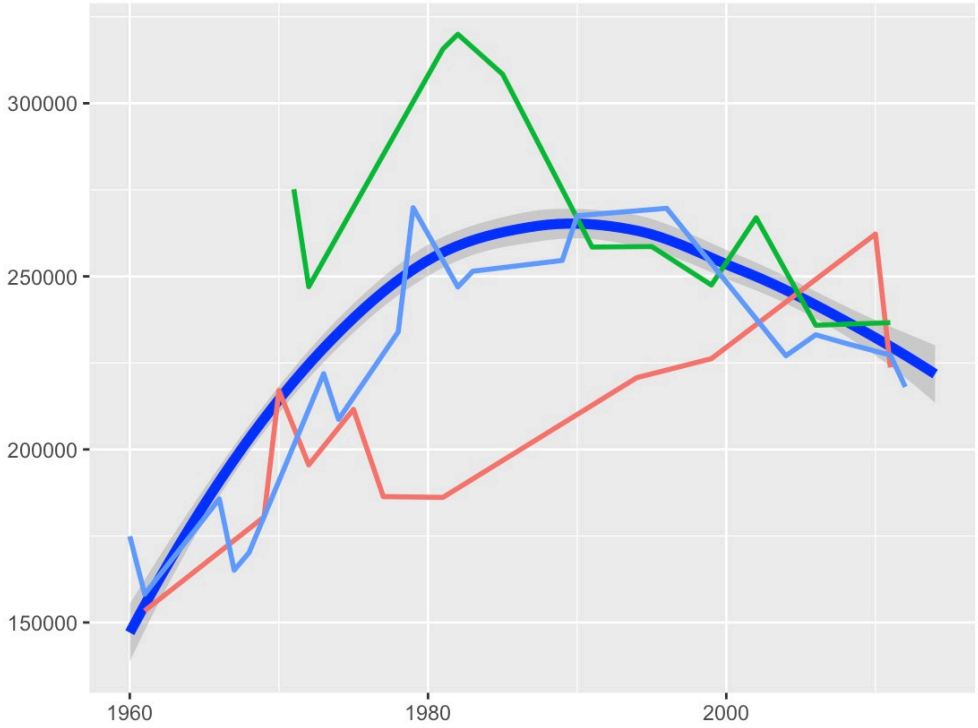
a. Have movie and movie scripts become shorter or longer over past years?

1). How did you perform data analysis in code?

For this question, I mainly used R to complete the analysis and visualization. By connecting to the database, I extracted all the variables needed in this analysis and stored them in data table using the data.table() function provided by R. Data table is a strong data structure that allowed me to perform calculations and aggregations within one step.

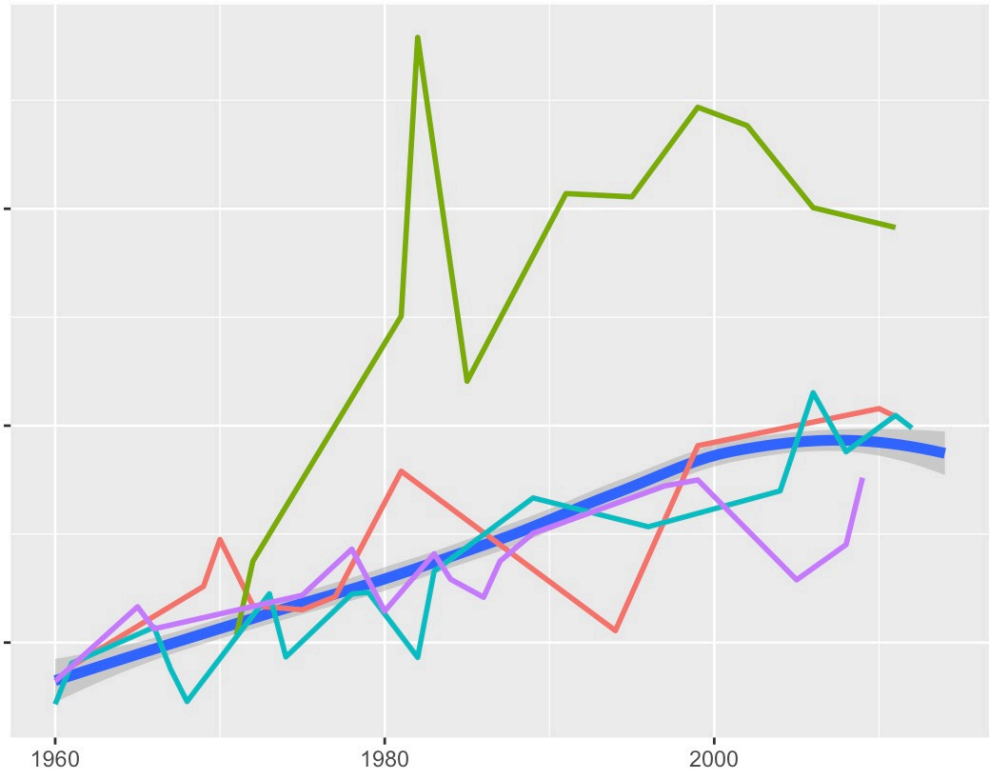
In order to find out how movie length has changed in past years and how does it vary across different genres, I first calculated the mean value of movie length for each year using data table. I also calculated the mean value of movie length for each year and each genre. After the calculations and aggregations were done, I plotted the data using ggplot() with geom_smooth() and geom_line().

2). Summary:



The blue smooth line is the average movie length for all movie of each year for all genres, include Action, Crime, Comedy and Drama. As we can see from the plot, average movie length became longer before 1990 and became shorter afterwards, which

means 1990 is the peak of the movie length.



The blue smooth line represents the average movie scripts length for all movies of each year for all genres, include Action, Crime, Comedy and Drama. As we can see from the plot, Crime has longer scripts length than other genres in general and the average movie scripts length has become longer over past years.

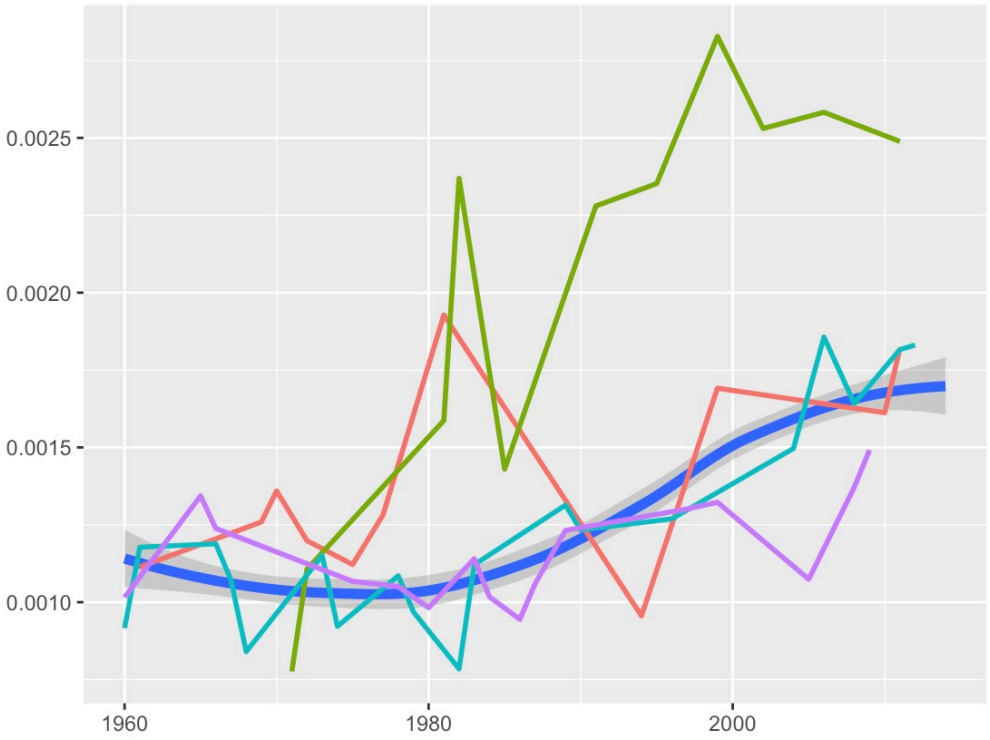
b. Did actors and actresses speak more words per minutes than before? How does these vary across different genre?

1). How did you perform data analysis in code?

For figuring out how scriptss length has changed in past 55 years, the method I used was pretty much the same as above. The only difference was that I had to calculate the number of words in each script first. I completed this step by using three functions: strsplit(), unlist() and length().

And finally, by dividing the scripts length by movie length and plotted the data, I was able to figure out the question of did singers sing more words per minute than before.

2). Summary:



According to the plot, the blue smooth line represent the average words actors and actresses tend to speak per minute. And it tends to increase after 1975. And it decreases slightly before 1975. In addition, actors and actresses speak more words per minute in Action movie according to the graph.

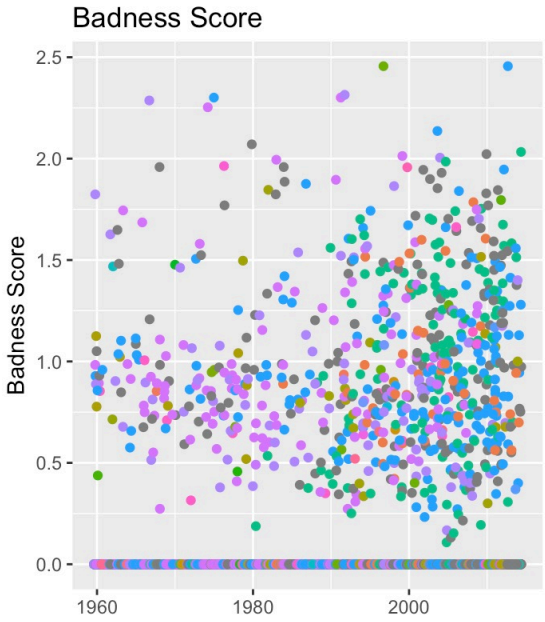
c. Have movie script become "worse" over past years?

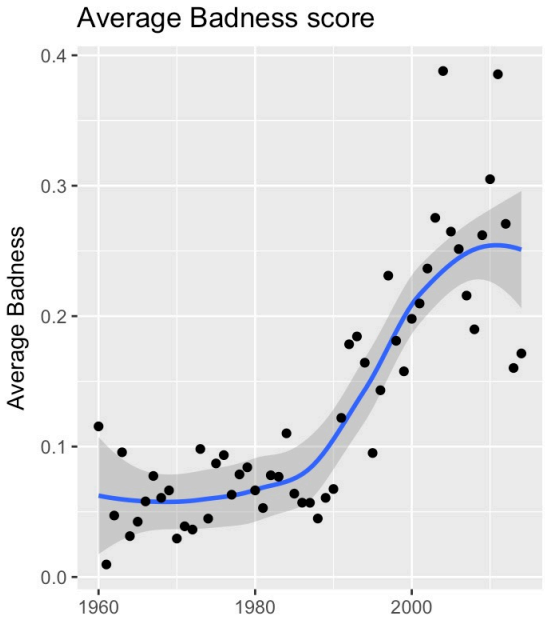
1). How did you perform data analysis in code?

I mainly used python to complete the natural language processing part and calculate the badness score of each movie. The specific code can be found in badness.py. The stopWordsRemoval() function takes a sentence as input and returns a list that contains all the words in the sentence except for the stop words. And the nltkStemmer() function utilizes the Snowball Stemmer from nltk to conduct word stemming. After the pre-processing was done, I calculated the badness score and stored all the scores in the database.

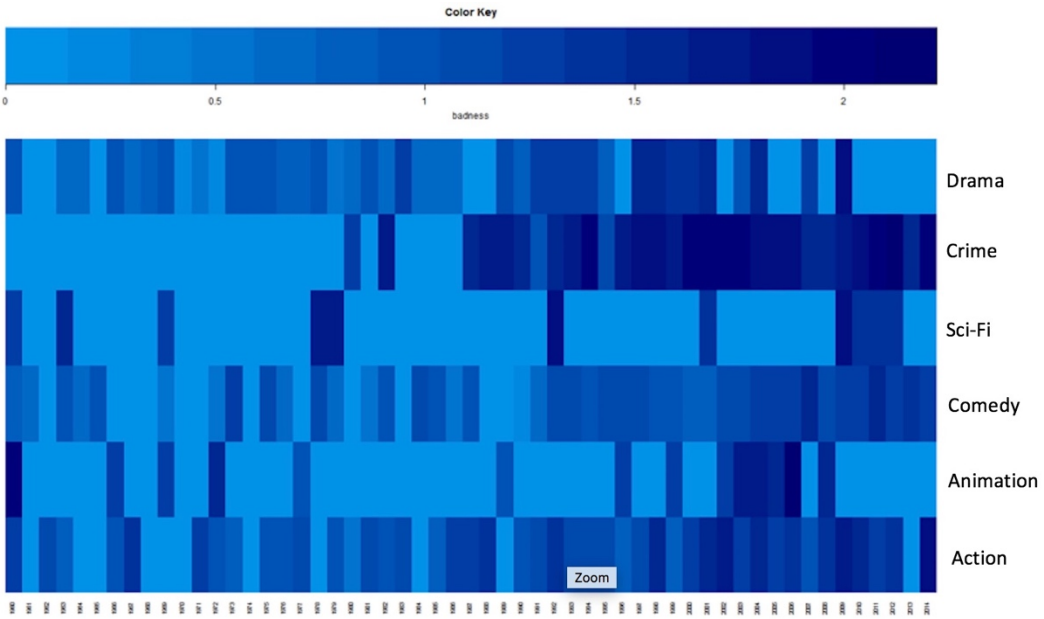
The R script for this question was pretty similar to the first question. After grabbing the data from my database, I stored all the data in data table and calculated the average badness score for each year and each genre. Then I plotted the data using ggplot() and heatmap.2().

2). Summary:





As we can see from the two plots above, scripts have become much "worse" over past 55 years. The badness score and the average badness score both increase a little bit to some extent.



According to this plot, Crime movie is the "worst" genre compared with other genres. And Sci-Fi is the "nicest" genre among all these genres. Based on the analyses above, I think it is necessary to rate the movie and just like we did.

d. How movie topics have changed over past years?

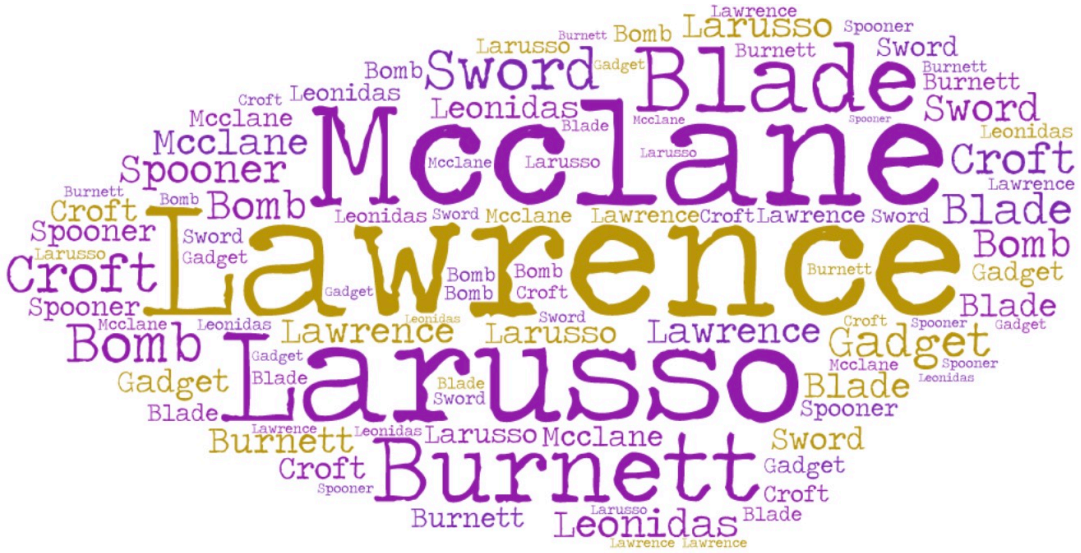
1). How did you perform data analysis in code?

I mainly used a python library, sklearn, for this part of data analysis. First, I used CountVectorizer() to create unigram frequency matrix for scripts of each decade. Then, I used LatentDirichletAllocation() to fit the data and generated the movie topics of each decade.

2). Summary:

Here I will list some word cloud visualization for each decades and their topic.

1960s:



We can see the famous movie Lawrence of Arabia is listed above.

1970s:

We could see that the superman and godfather popped out in 1970s.

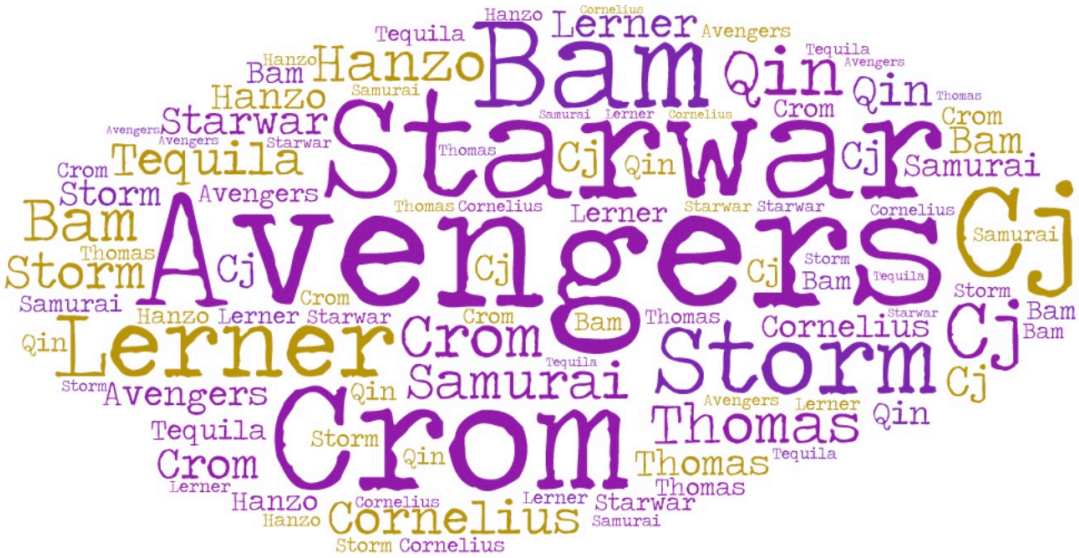
1980s:

We could see that the batman popped out in 1980s.

1990s:

1990s is a really a movie decade, we could see the famous movie The Shawshake Redemption and Forest Gump popped out here.

2000s till now:



We could see some popular movie like Avengers and Starwar here.