Movie Analysis Report

Che Jung Lee | 301249826

Introduction

The objective of this project is to determine the success of a movie in terms of popularity by its plot summary. The workflow of the data analysis consists of two main parts: predict movie genres from movie descriptions and analyze the trends of movie genres over time. This combination allows one to classify the genres of a movie and correlate them to the current trend to see if the identified genres are popular. Granted, the trend of the genres does not directly represent the popularity of a movie, but it is correlated and is certainly one of the factors that affect the acceptance of the movie. The result of this project is therefore significant as it analyzes one of the factors that contribute to the overall success of a movie. This report is going to address the details of the workflow with the focus on the aspects of data processing and analysis methodology.

Data Processing

This section is mainly about data transformation, as the extraction part is done by the professor already. There are two data sets used in this project: omdb and wikidata. For omdb, I filtered out movies with no plots, then removed all characters and punctuations other than letters and digits to improve the accuracy of upcoming text classification. The genres are extracted from the same data set by converting the genre Series into a list to get the unique elements of it. This step is a bit tricky as the column contains lists of genres, which can only be obtained through nested loops or inefficient groupby and apply methods in the pandas way. The solution for this is to use np.concatenate to unstack each genre of the lists into multiple rows, then use np.unique to get the unique elements. For wikidata, I filtered out movies with publication years before 1980 and after 2018 to remove outliers and misleading results. Movies produced before 1980 are irrelevant to the analysis, and this data set clearly lacks information about future films judging from its modified date and modern movie business. After cleaning both data sets, I merged them together to match the publication year with movies in omdb, and applied the same technique used in the genre extraction to unstack columns consisting of lists into rows. This allows me to count the number of movies for each genre annually, which can then be used to create a Dataframe shown in Figure 1 by applying groupby and unstack methods on the merged data.

genre	Action	Adult	Adventure	Animation	Biography	Comedy	Crime	\
year								
1980	11.0	1.0	10.0	1.0	5.0	17.0	6.0	
1981	17.0	NaN	15.0	4.0	6.0	20.0	10.0	
1982	19.0	NaN	15.0	5.0	1.0	27.0	6.0	
1983	21.0	NaN	11.0	1.0	2.0	24.0	14.0	
1984	26.0	NaN	19.0	3.0	2.0	29.0	11.0	
1985	28.0	NaN	25.0	7.0	4.0	37.0	13.0	
1986	18.0	NaN	21.0	6.0	5.0	34.0	8.0	
1987	28.0	NaN	22.0	10.0	3.0	51.0	21.0	
1988	26.0	NaN	16.0	11.0	4.0	42.0	25.0	
1989	31.0	NaN	23.0	5.0	6.0	44.0	20.0	
1990	39.0	NaN	17.0	2.0	5.0	40.0	29.0	
1991	33.0	NaN	17.0	4.0	4.0	40.0	27.0	
1992	23.0	NaN	13.0	3.0	10.0	32.0	25.0	
1993	24.0	NaN	20.0	5.0	12.0	40.0	17.0	
1994	27.0	NaN	20.0	6.0	7.0	53.0	27.0	
1995	47.0	NaN	29.0	11.0	10.0	59.0	34.0	
1996	33.0	NaN	23.0	5.0	8.0	56.0	25.0	
1997	35.0	NaN	28.0	9.0	11.0	40.0	21.0	
1998	35.0	NaN	30.0	16.0	9.0	72.0	36.0	
1999	37.0	NaN	23.0	8.0	18.0	77.0	34.0	
2000	41.0	NaN	36.0	15.0	10.0	77.0	38.0	

Figure 1: genre trend dataframe

Text Classification

Movie genre prediction is a multi-label text classification, because each movie can be associated with multiple genres. For this type of problem, I chose term frequency - inverse document frequency (tf-idf) bag-of-word model, as suggested by the professor, to extract genres from movie plots. The first step of the analysis is to use MultiLabelBinarizer to encode movie genres in binary label representation, as per the requirement of tf-idf. I then train three classifiers – Naïve Bayes, SVM linear, and Logistic Regression, to find the best performer in terms of overall prediction accuracy. Due to the multi-label scenario, OneVsRestClassifier, also known as Binary Relevance, is needed to wrap the classifiers to train them one for each different genre. The above process is done through Pipeline with Cross-Validation – GridSearchCV, to validate and select the optimal parameter configurations for an algorithm at the same time. After the tests, it seems that tf-idf with SVM linear yields the best results, with an average accuracy of 42% being correct (Figure 2, f1-score). Note that some genres have precision of 0%. This is because the number of movies with those genres are too few (Figure 4). So, they were not in the prediction and were never predicted; hence resulted in 0% in f1-score fields. Also note that certain genres have higher precision. This is due to the skewed distribution of the provided data, where some genres have more occurrences than the others in this data set and thus allow more training to be done on them (Figure 4). This implies that the average accuracy will be significantly higher if there is more data or if the data is more balanced. However, I considered the accuracy of 42% to be acceptable in predicting the genres from movie plots, given that there are only 9508 movies to train on. Therefore, I think the text classifier is successful in this case.

(Figure 3 shows the sample outputs of the new input movie plots and their genre predictions)

	precision	recall	f1-score	support
Action	0.60	0.41	0.49	470
Adult	0.00	0.00	0.00	2
Adventure	0.58	0.31	0.41	411
Animation	0.50	0.11	0.18	134
Biography	0.53	0.19	0.28	166
Comedy	0.57	0.43	0.49	771
Crime	0.64	0.40	0.50	386
Documentary	0.81	0.49	0.61	170
Drama	0.68	0.65	0.67	1156
Family	0.54	0.12	0.19	188
Fantasy	0.56	0.10	0.16	199
Film-Noir	0.00	0.00	0.00	24
History	0.50	0.02	0.04	100
Horror	0.57	0.32	0.41	277
Music	0.55	0.23	0.32	75
Musical	0.00	0.00	0.00	35
Mystery	0.31	0.07	0.12	156
N/A	0.00	0.00	0.00	0
News	0.00	0.00	0.00	1
Romance	0.62	0.28	0.39	388
Sci-Fi	0.64	0.29	0.40	214
Short	0.00	0.00	0.00	17
Sport	0.60	0.08	0.14	38
Talk-Show	0.00	0.00	0.00	0
Thriller	0.42	0.07	0.13	294
War	0.45	0.13	0.20	69
Western	1.00	0.03	0.06	34
avg / total	0.59	0.36	0.42	5775

Figure 3: precision summary report

Figure 2: genre prediction sample

```
detective chris kenner was orphaned as a... => Action, Crime, Drama

following the theft of a highly-secured ... => Action, Comedy, Crime

cole is an aspiring dj who spends his da... => Comedy, Drama, Romance

good and evil scanners combatting when a... => Action

forty-eight year old will keane is a suc... => Comedy, Drama, Romance

top cover girl and fashion model jennife... => Drama

life has its downs for james living with... => Comedy

the cuddly care bears star in this charm... => Adventure

a somewhat daffy book editor on a rail t... => Comedy, Crime
```

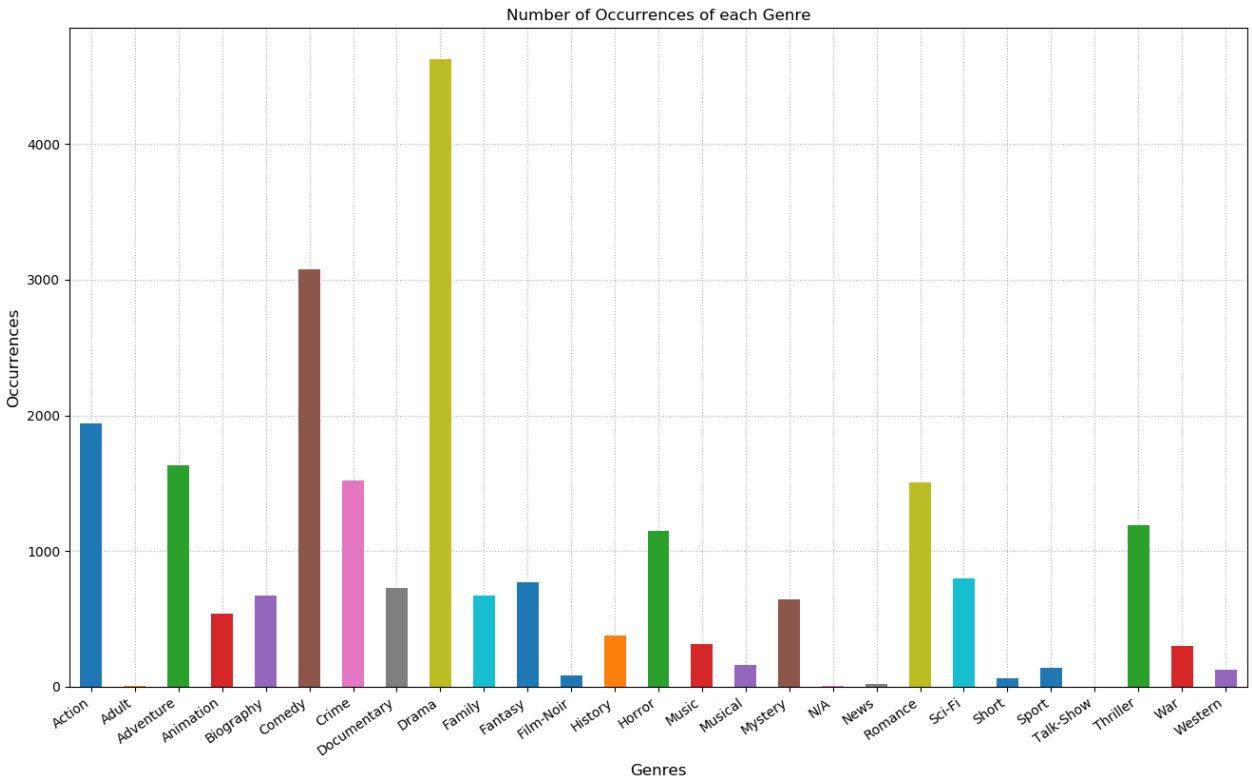


Figure 4: genre vs occurrence in the data set

Trend Analysis

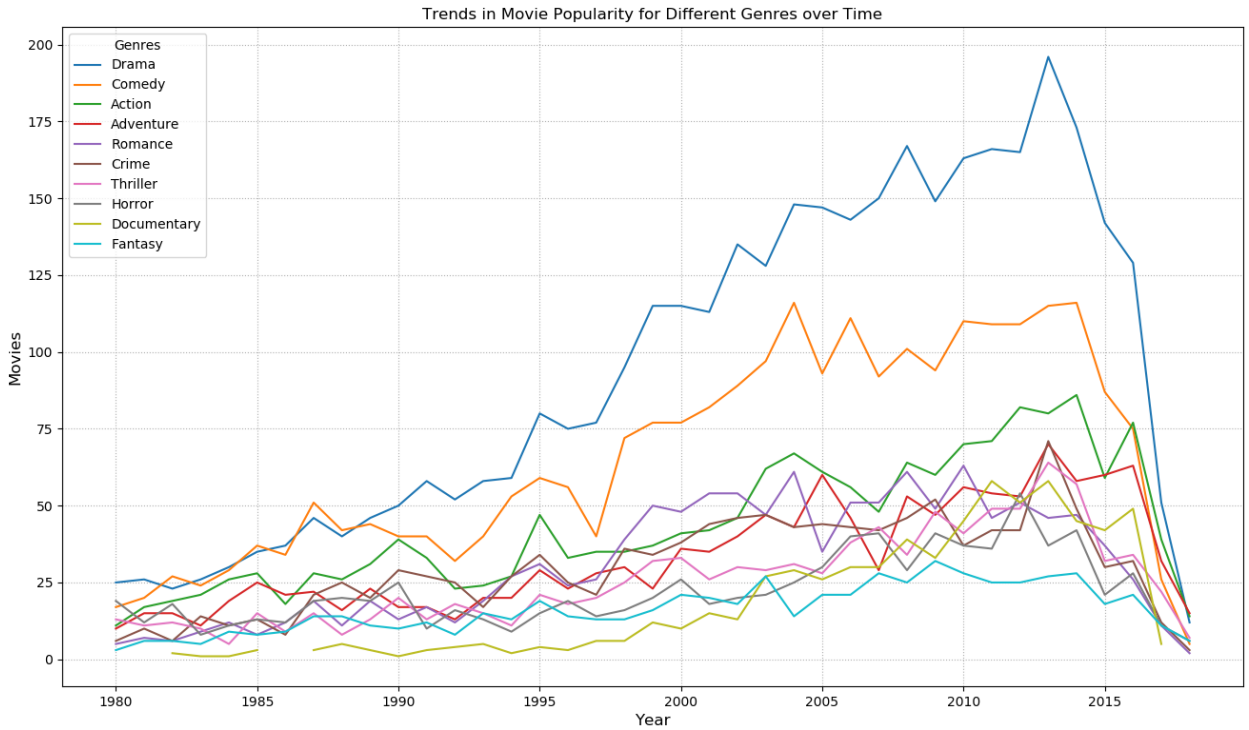
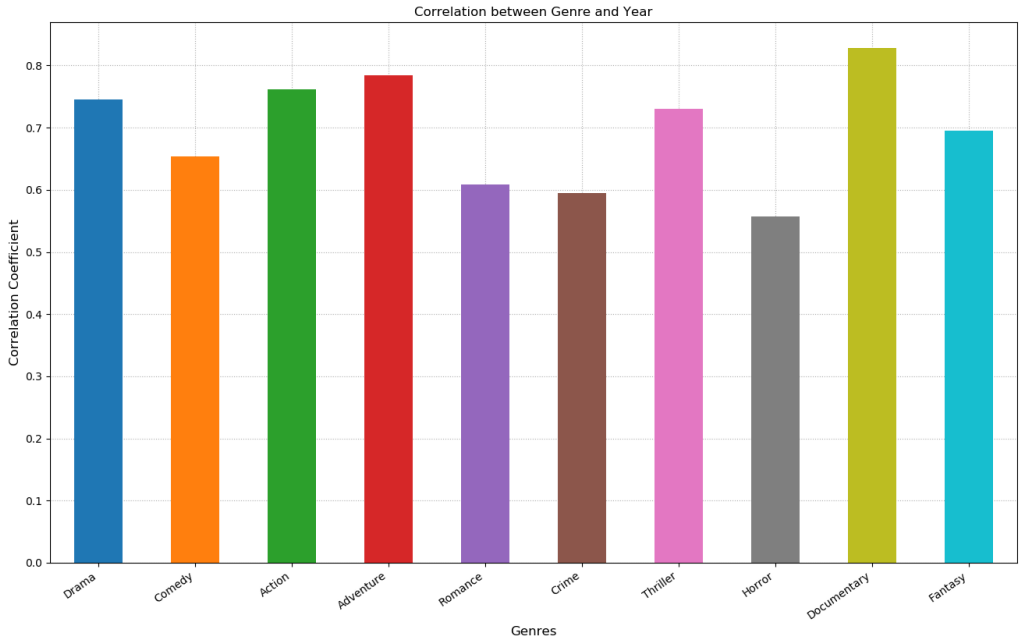


Figure 5: movie genre trend over time

Another analysis in this project is the estimation of movie genre trends over time. This is simpler than the text classification but involved a lot of data transformations. Figure 5 shows the popularity of movie genres from 1980 to 2018. The popularity is simply the count of movie genres per year, which can be calculated from groupby and aggregate. Interestingly, the overall number of movies has a significant drop at around 2013 and seem to be approaching to 0 at around 2016. I am unsure of the cause of such decrease; perhaps the provided data does not include a portion of the movies, or the data is collected before its modified date. Nonetheless, the results excluding the strange part seem promising, and the correlation between the number of movie genres and time proves that there exists a positive linear relationship, such that the two variables are not independent of each other.



Conclusion

This project addresses the multi-label text classification problem to determine the success of a movie in terms of popularity by its plot summary. It estimates the genres from movie plots through the combination of MultiLabelBinarizer, OneVsRestClassifier, tf-idf, and SVM linear, which was tested and compared with other classifiers. It also illustrates the trend of the movie genres over time and calculates the correlation between each other. Based on these results, we can conclude that the text classification can predict success in a useful way by correlating it to the change of movie genre changes over time.