Text Classification of Movie Plot Summary to Predict Movie Genre

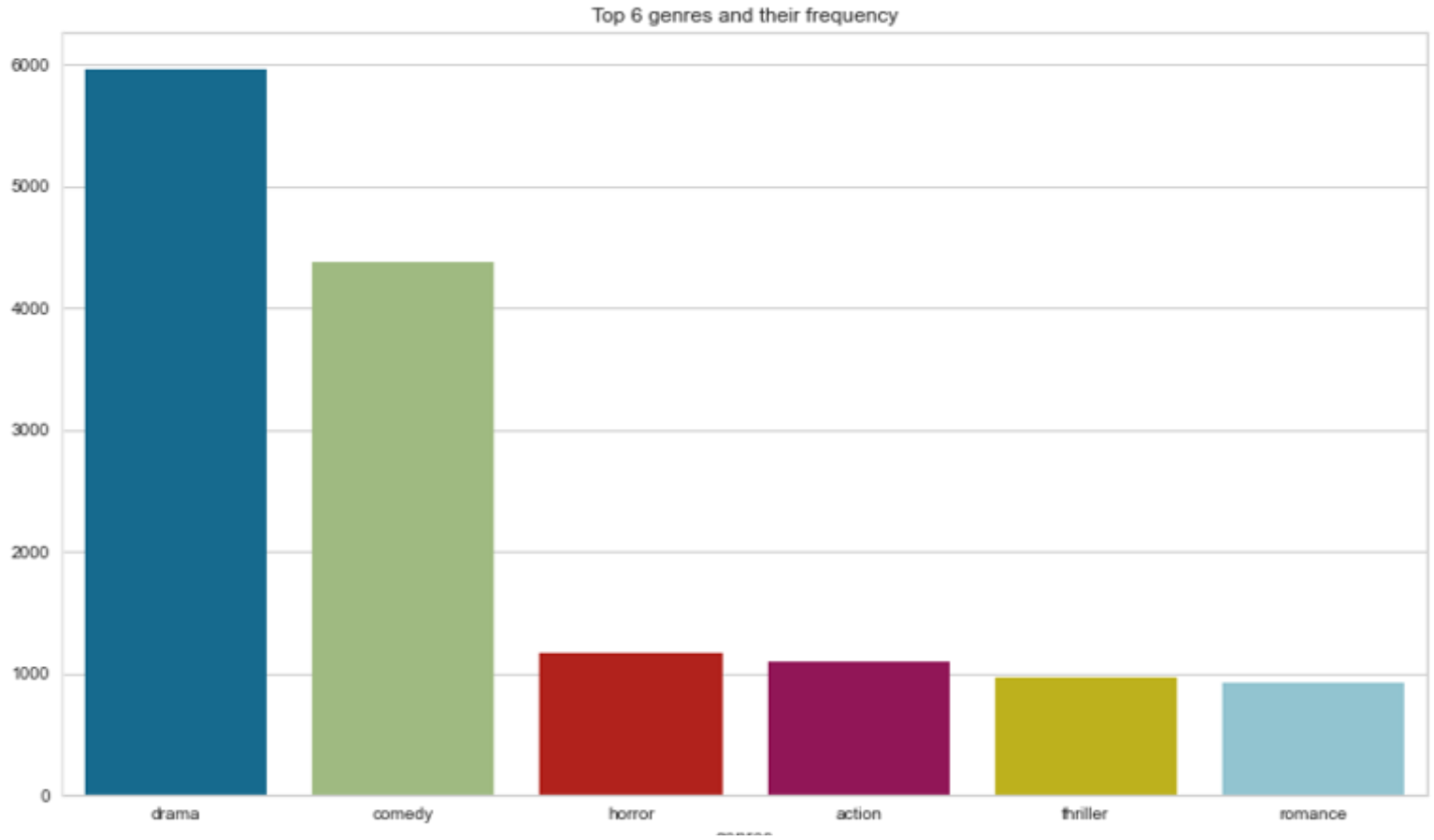
WONUOLA ABIMBOLA

Business Understanding

- Movies are a popular means of entertainment and so there are large volumes of movie data online
- The goal is to build a predictive model that predicts movie genre using the plot summary
- This would ideally be used for automated genre generation on movie streaming platforms

Data Understanding and Cleaning

- The dataset used in this project was obtained from Kaggle.
- It contains 34,886 movie descriptions scraped from Wikipedia
- Dropped movies with more than one genre
- Preprocessing Steps
  - Changed to lowercase
  - Removed stopwords
  - Word tokenization (nltk)
  - Lemmatization of words (spaCy)

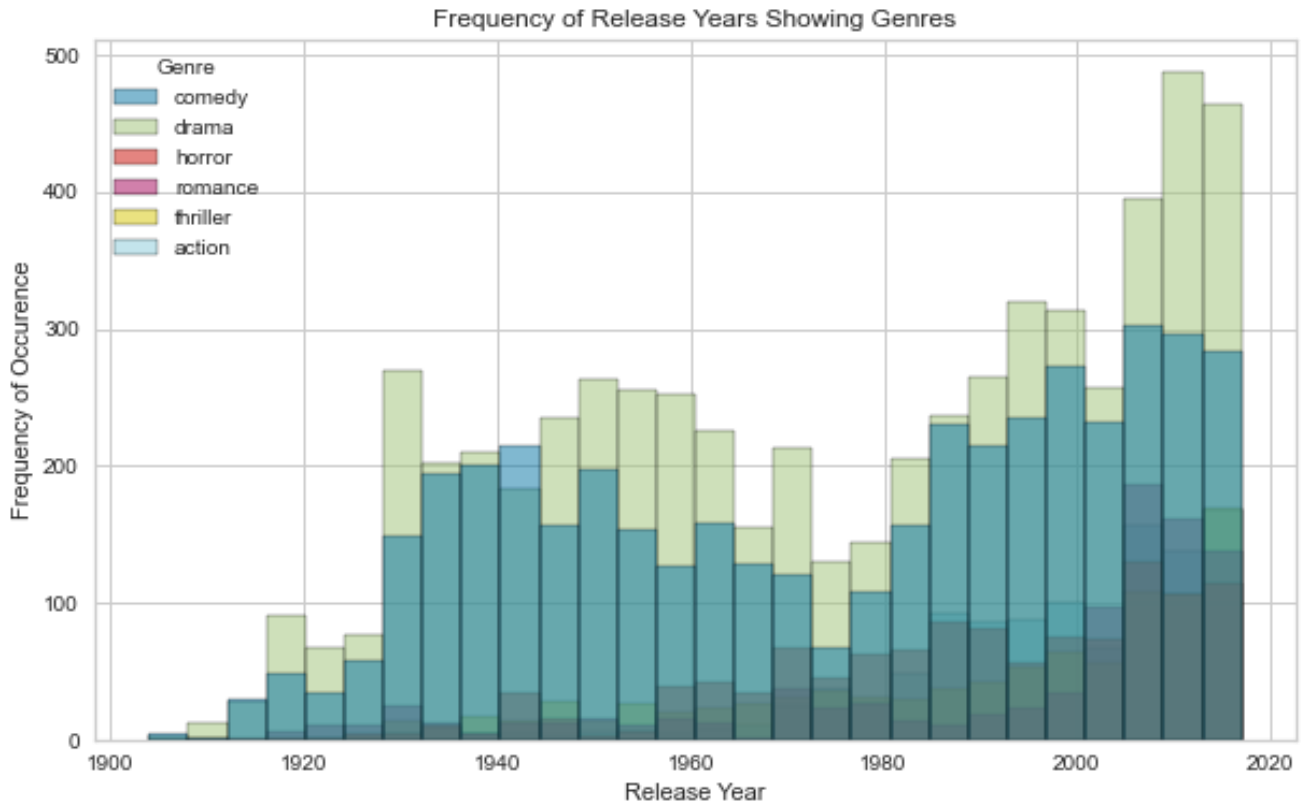


- Only movies that had one genre assigned were used (~15k)

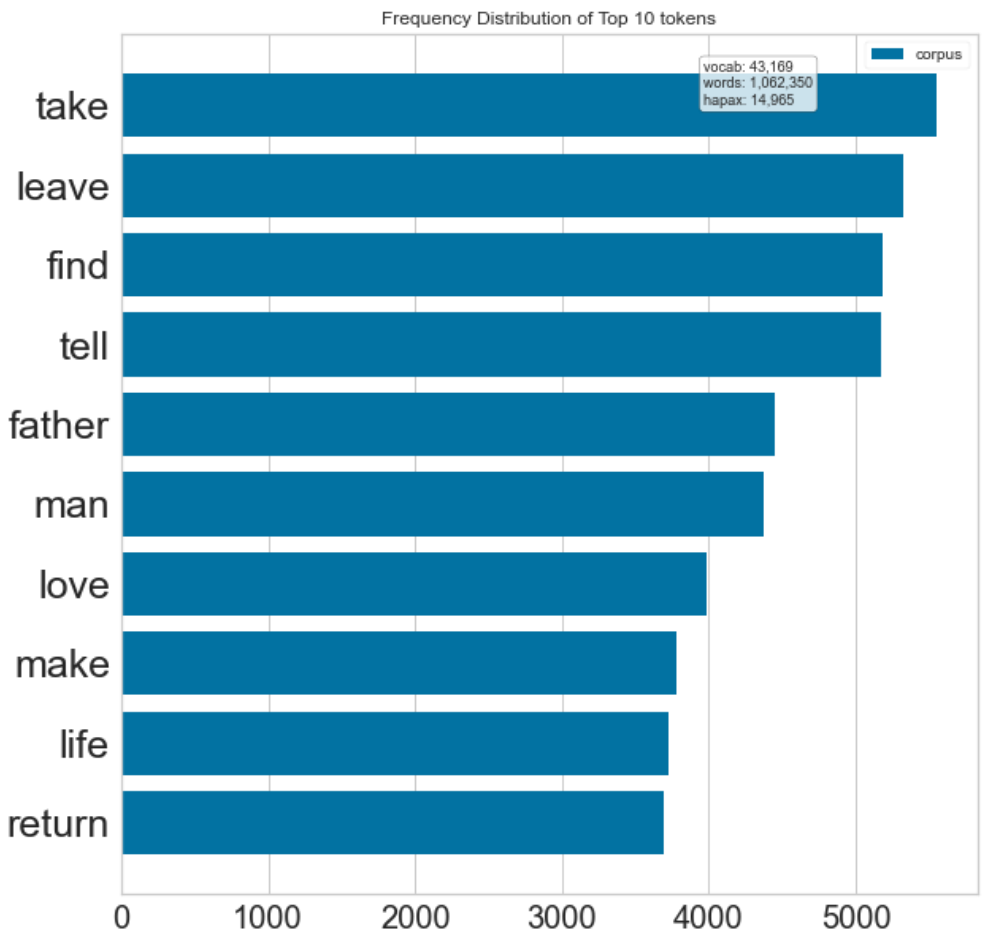
- Top six genres are drama, comedy, horror, action, thriller and romance in descending order.

Exploratory Data Analysis

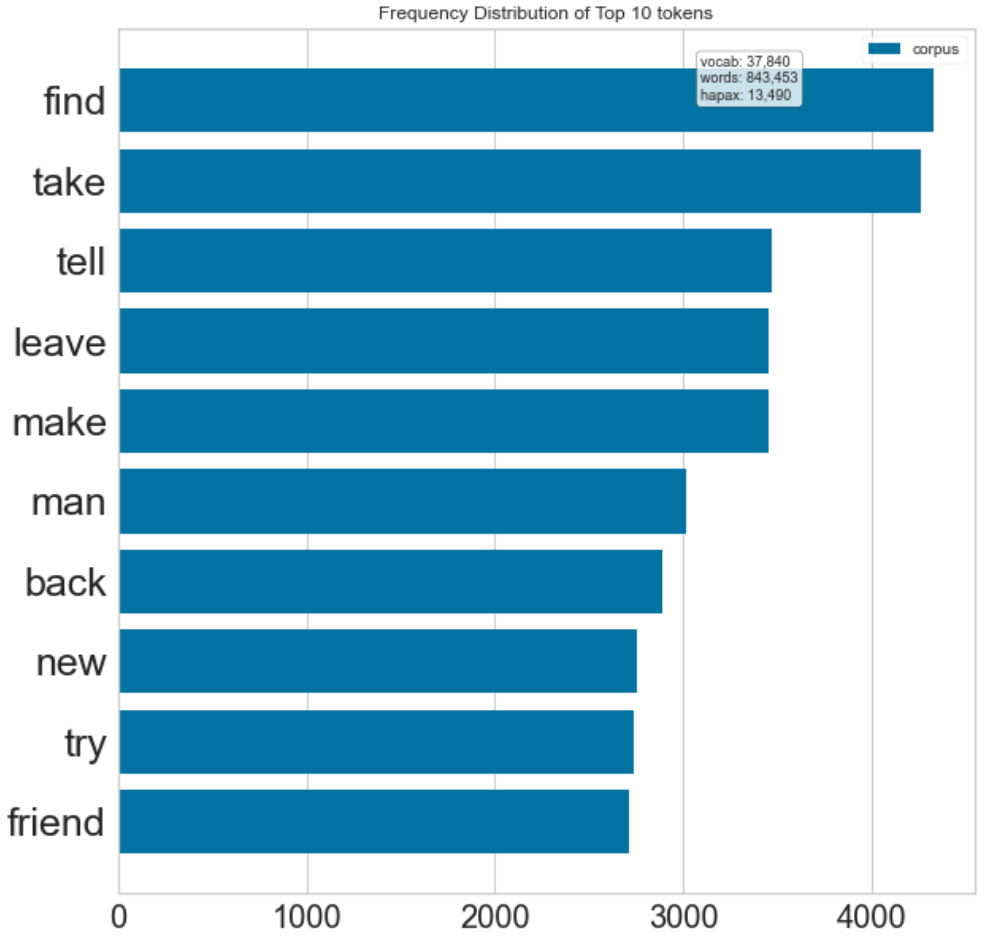
- Release years range 1904 – 2017
- Drama is the most popular in most decades



Drama : Frequency Distribution of Top 10 Words in Plot Summary

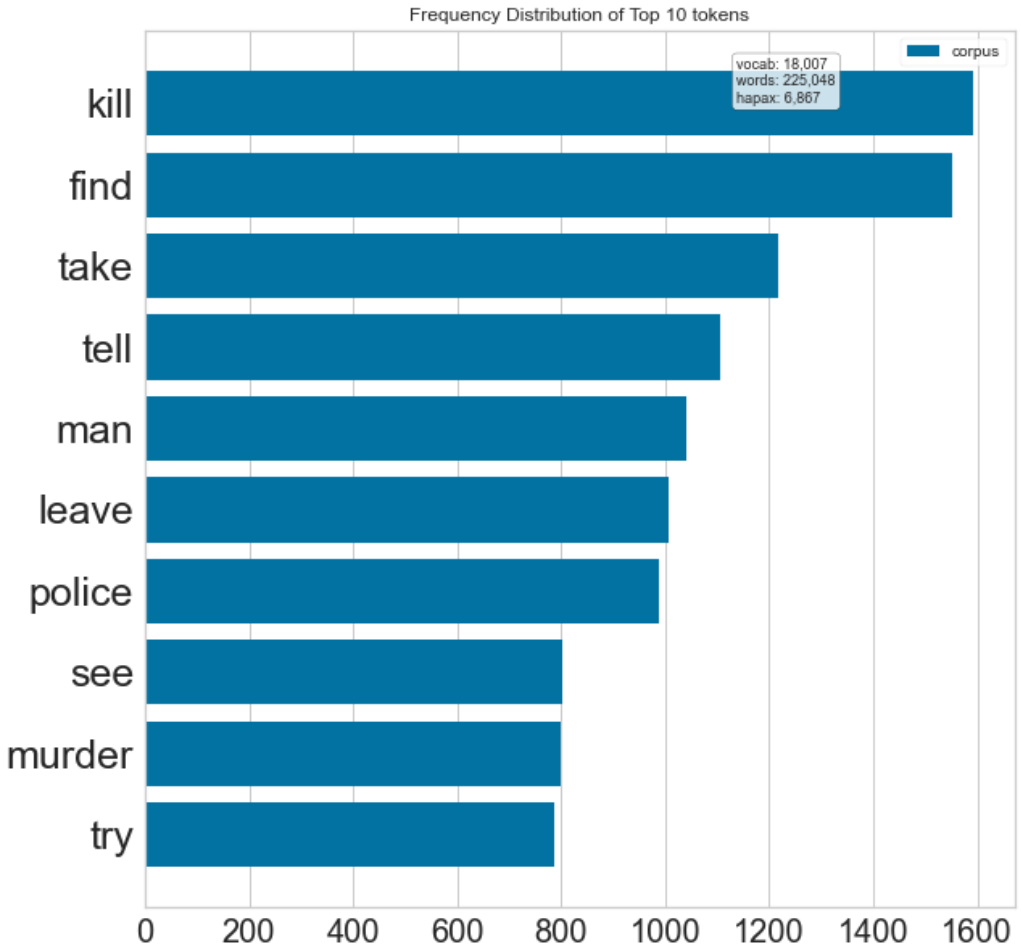


Comedy : Frequency Distribution of Top 10 Words in Plot Summary

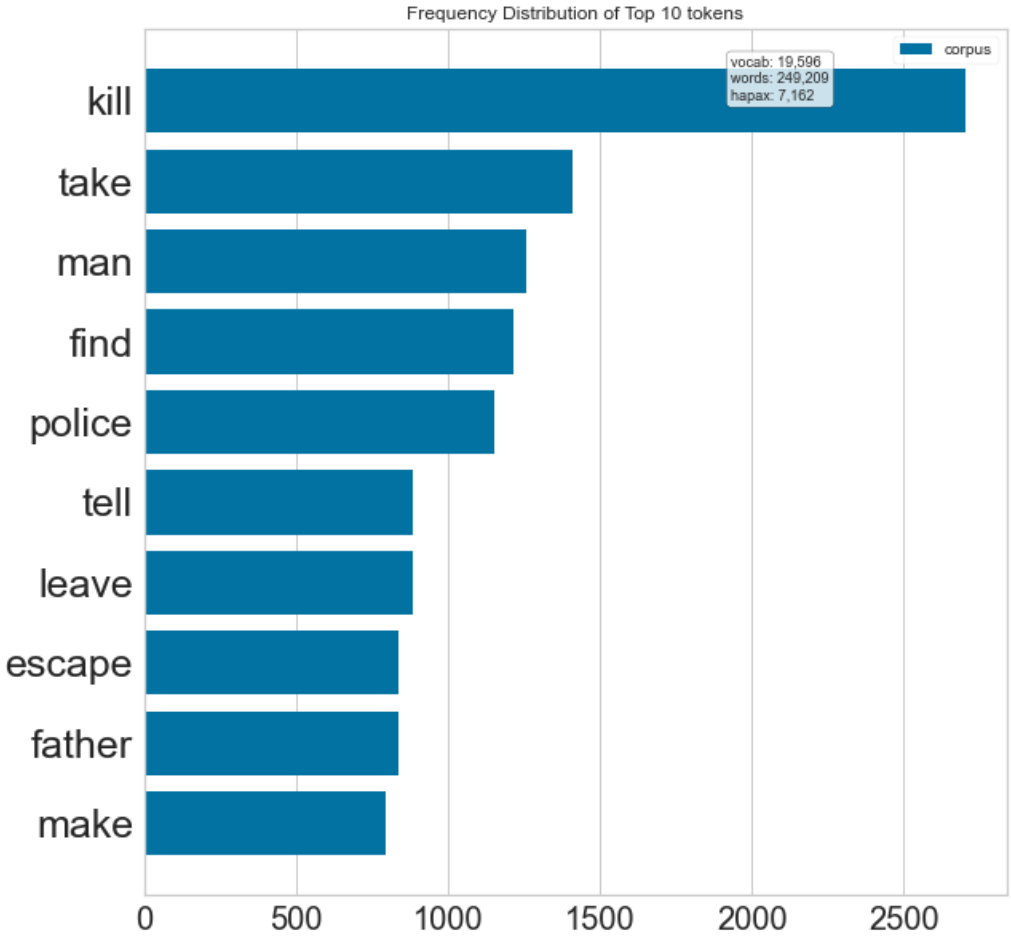


Common words: take, find, tell, leave, man, make

Thriller : Frequency Distribution of Top 10 Words in Plot Summary

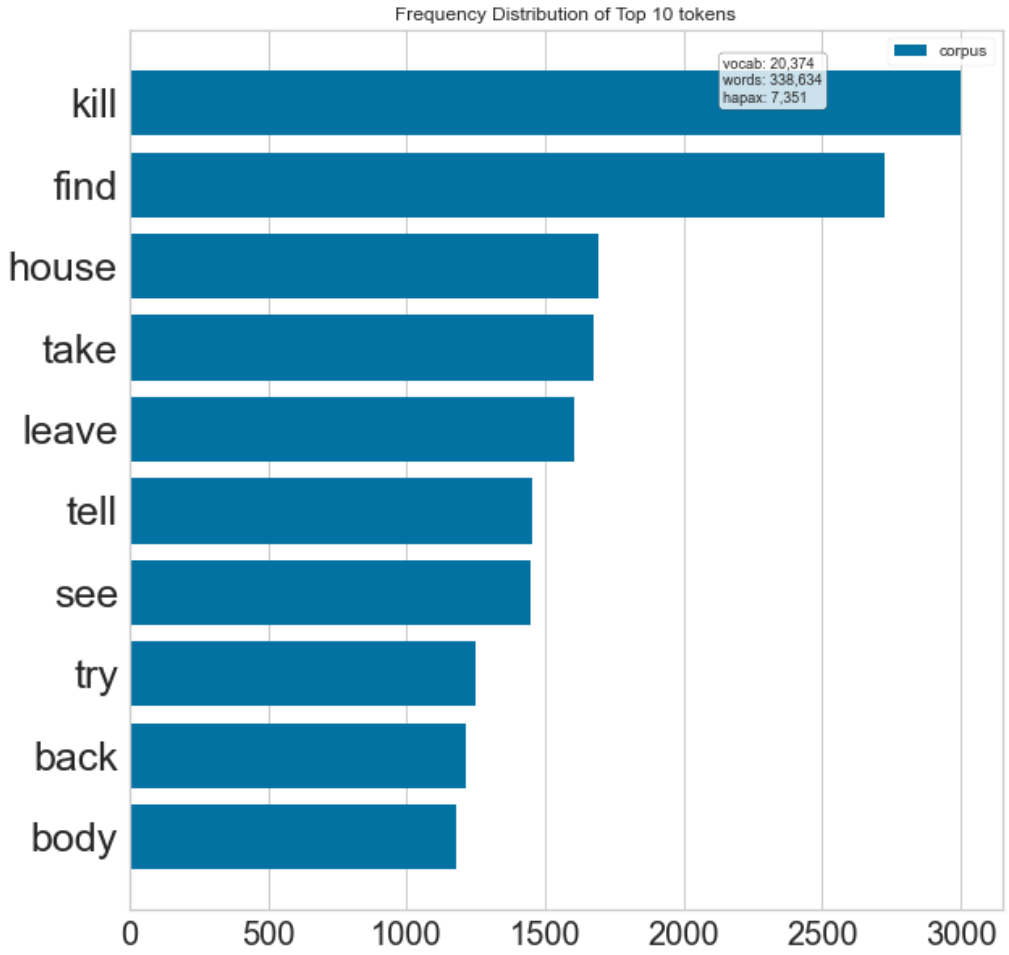


Action : Frequency Distribution of Top 10 Words in Plot Summary

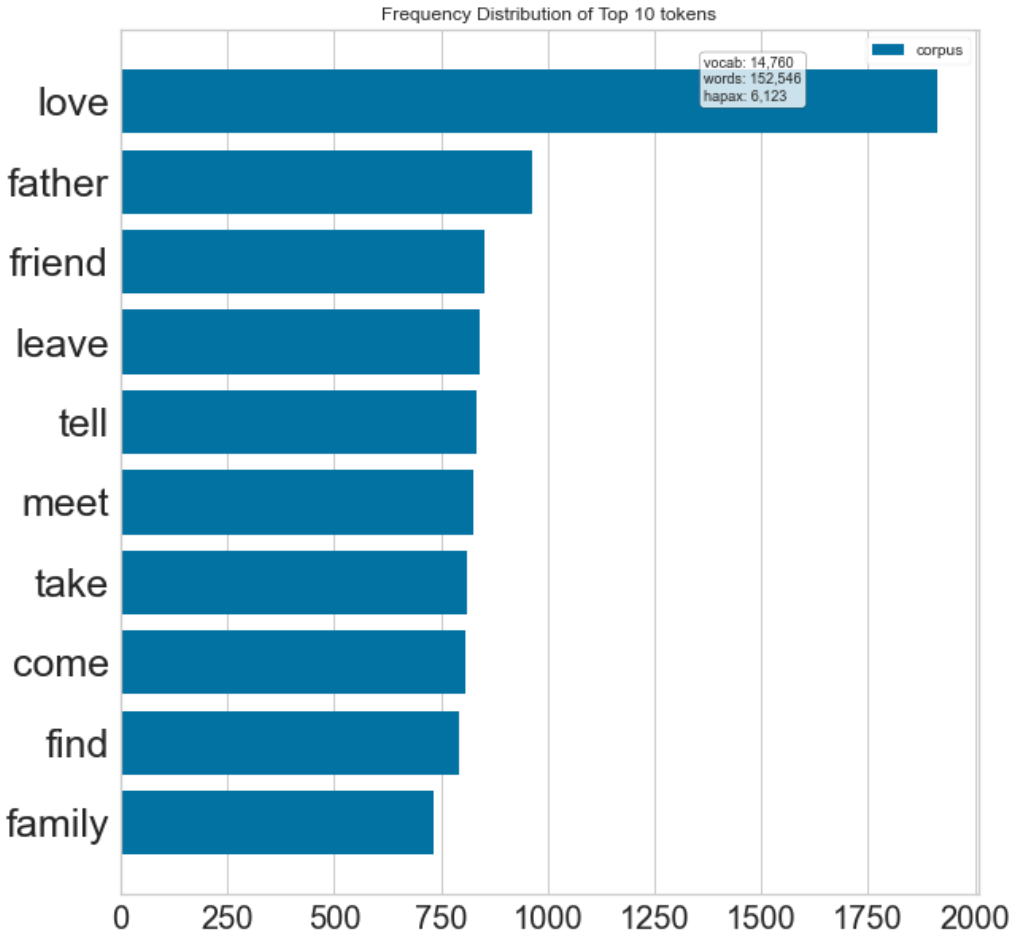


Common words : kill, find, police, tell, leave, take, man

Horror : Frequency Distribution of Top 10 Words in Plot Summary



Romance : Frequency Distribution of Top 10 Words in Plot Summary



Common words : tell, take, find

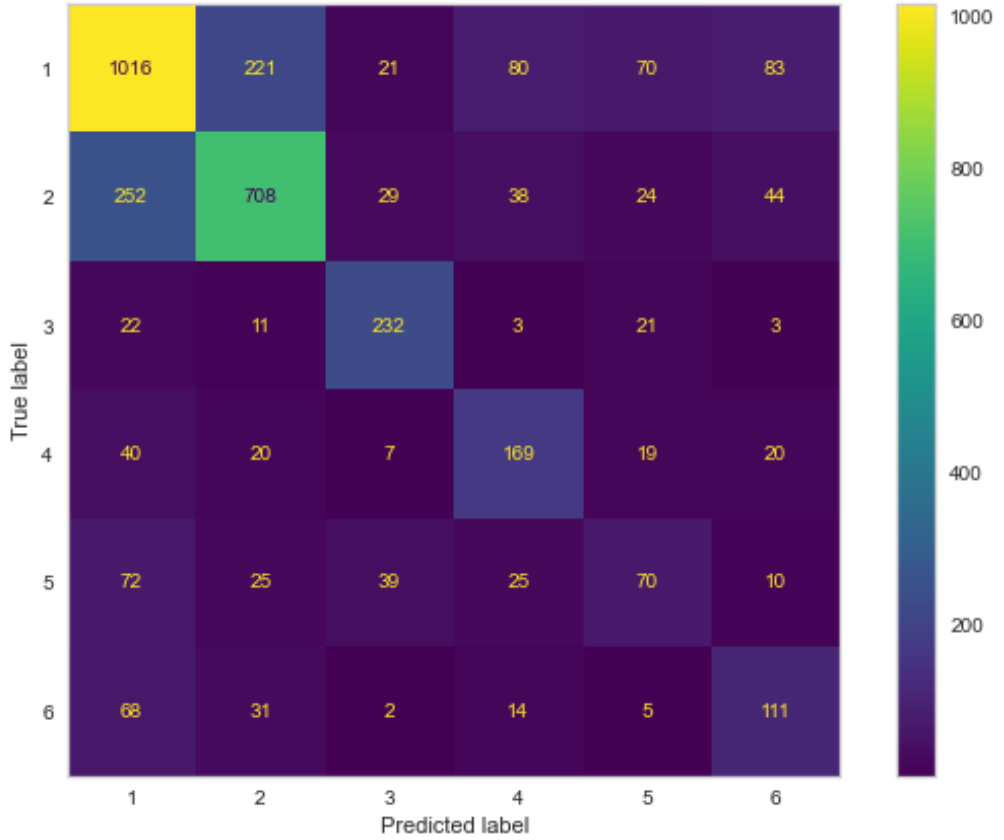
Modeling and Training Results

- Baseline accuracy of 41%
- Best model tfidf logistic regression
- Test set accuracy score was ~64%

	model	tfidf_score	count_vec_score
0	Multinomial Bayes	0.617643	0.604211
1	Logistic Regression	0.635853	0.615341
2	Decision Tree	0.469279	0.467807

Test Results and Conclusion

- Testing accuracy score was ~64%
- The confusion matrix shows the amount how good the model's predictions are.



Next Steps

- Implementing Neural Networks in this problem to possibly achieve better accuracy.
- Generated genres (in addition to other factors like tags and user ratings) could be implemented in movie recommendation systems.