Text Classification of Movie Plot Summary to Predict Movie Genre

WONUOLA ABIMBOLA

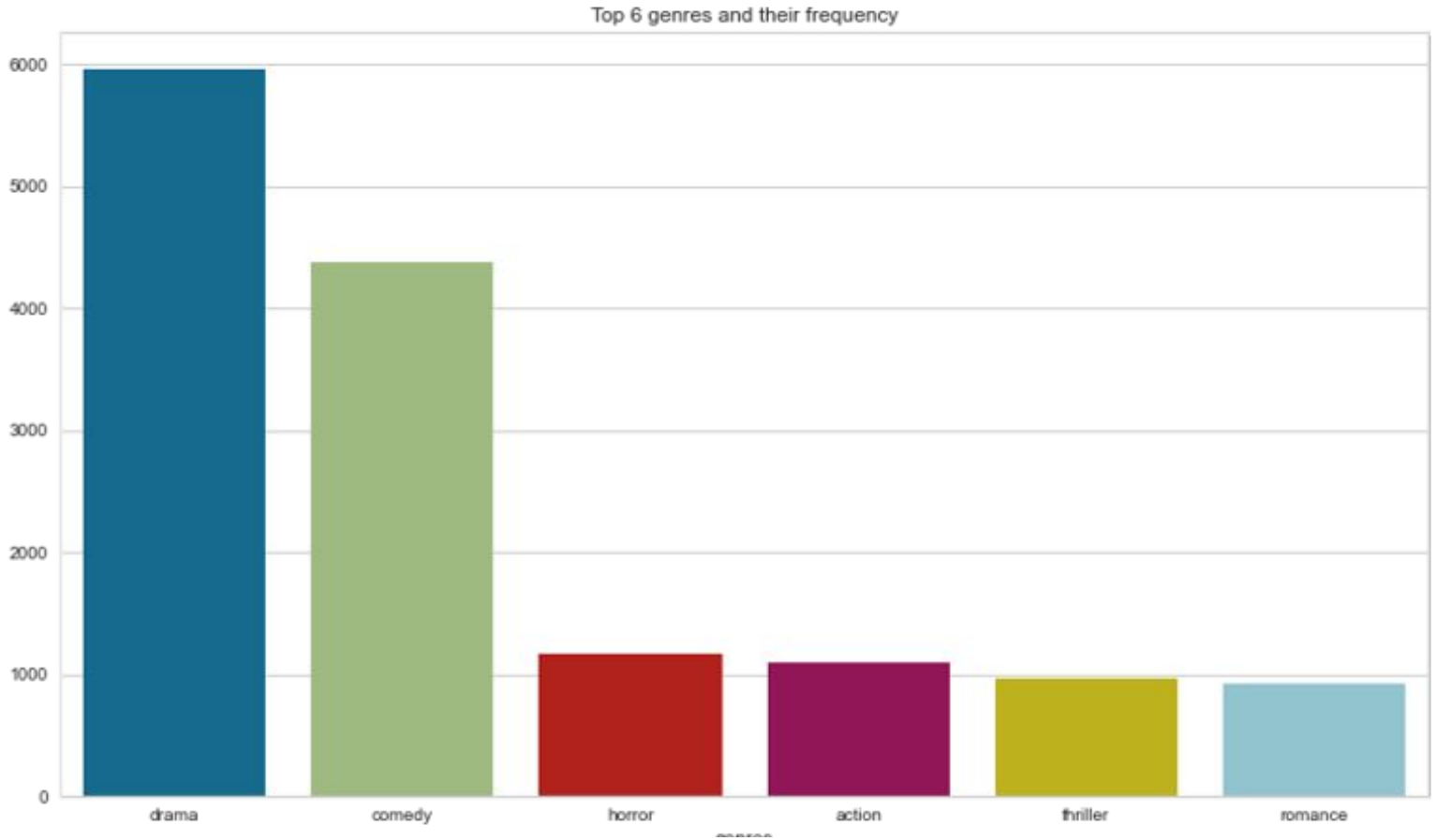
Business Understanding

- 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)

Exploratory Data Analysis

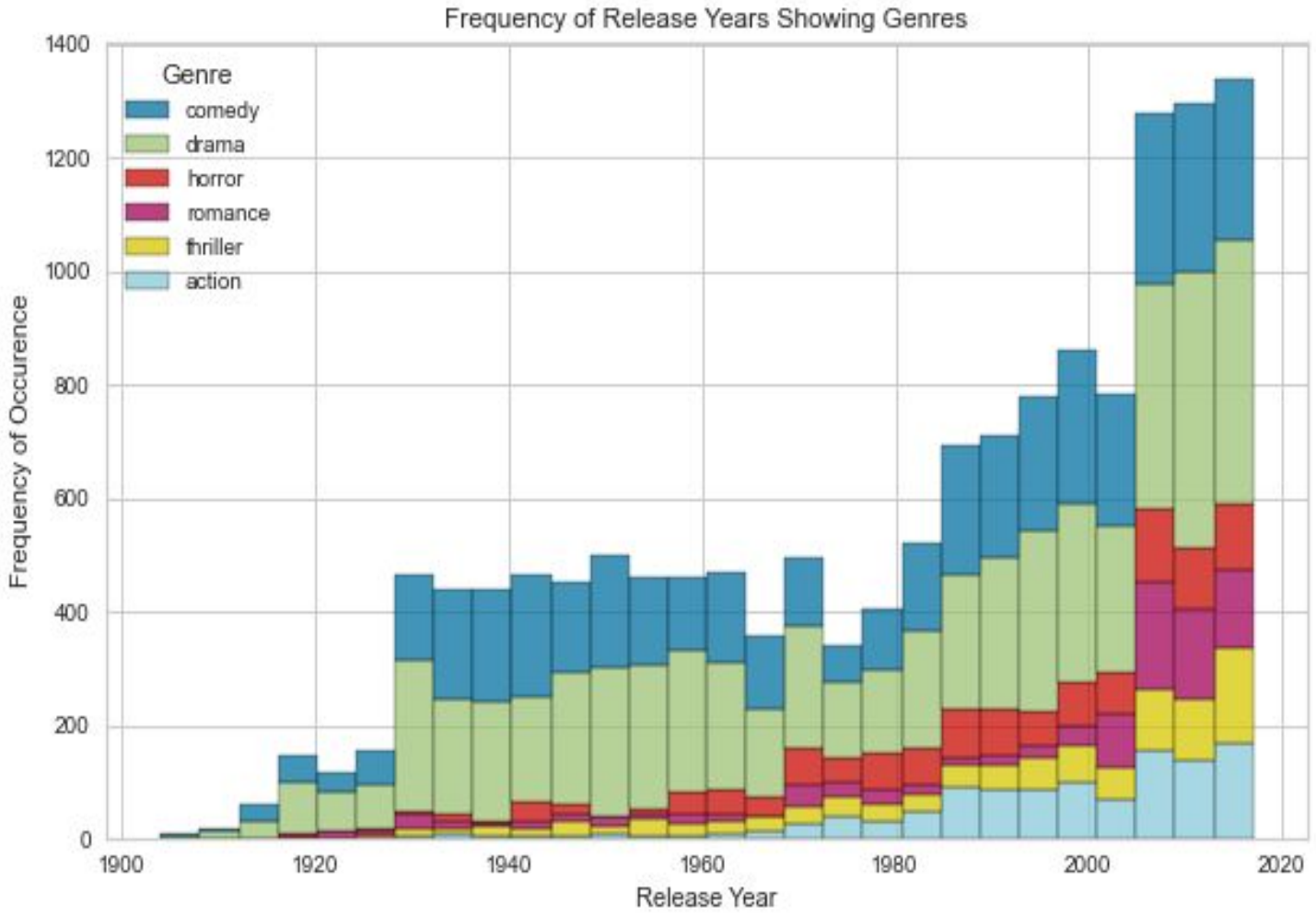


- Only movies that had one genre assigned were used (~15k)

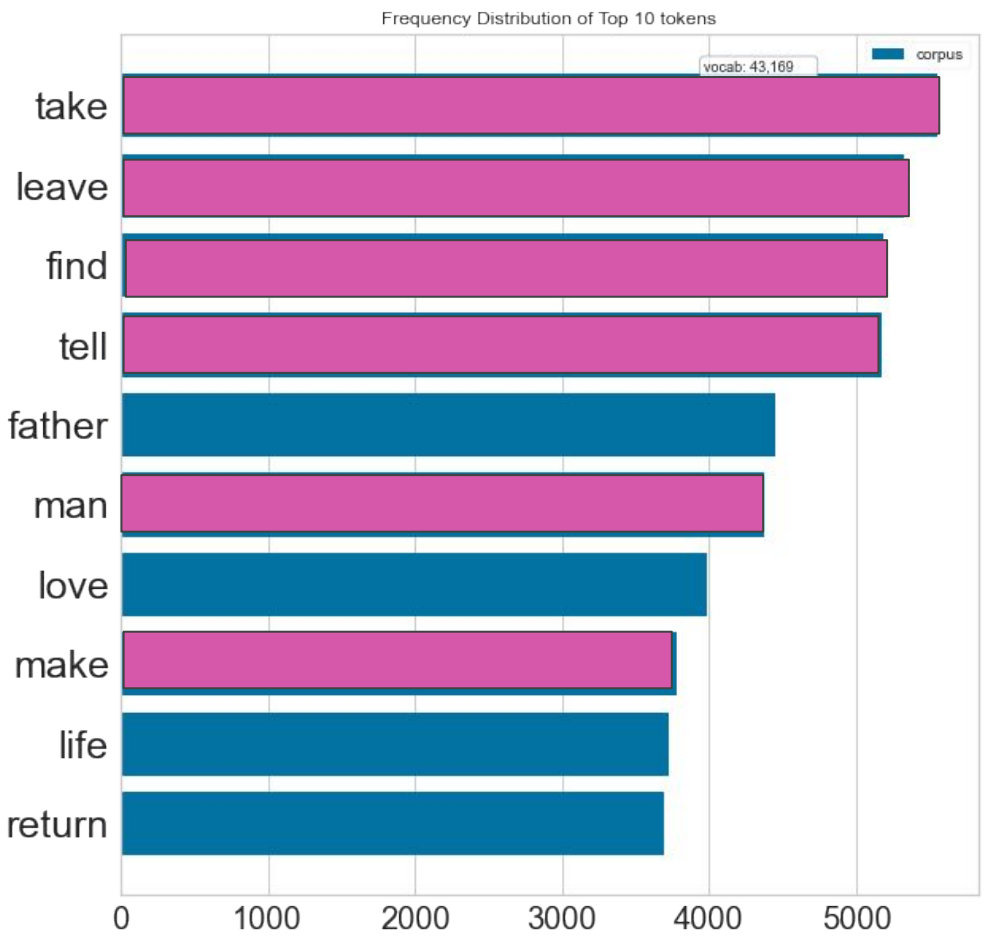
- Drama is the most frequent while romance is the least

Exploratory Data Analysis

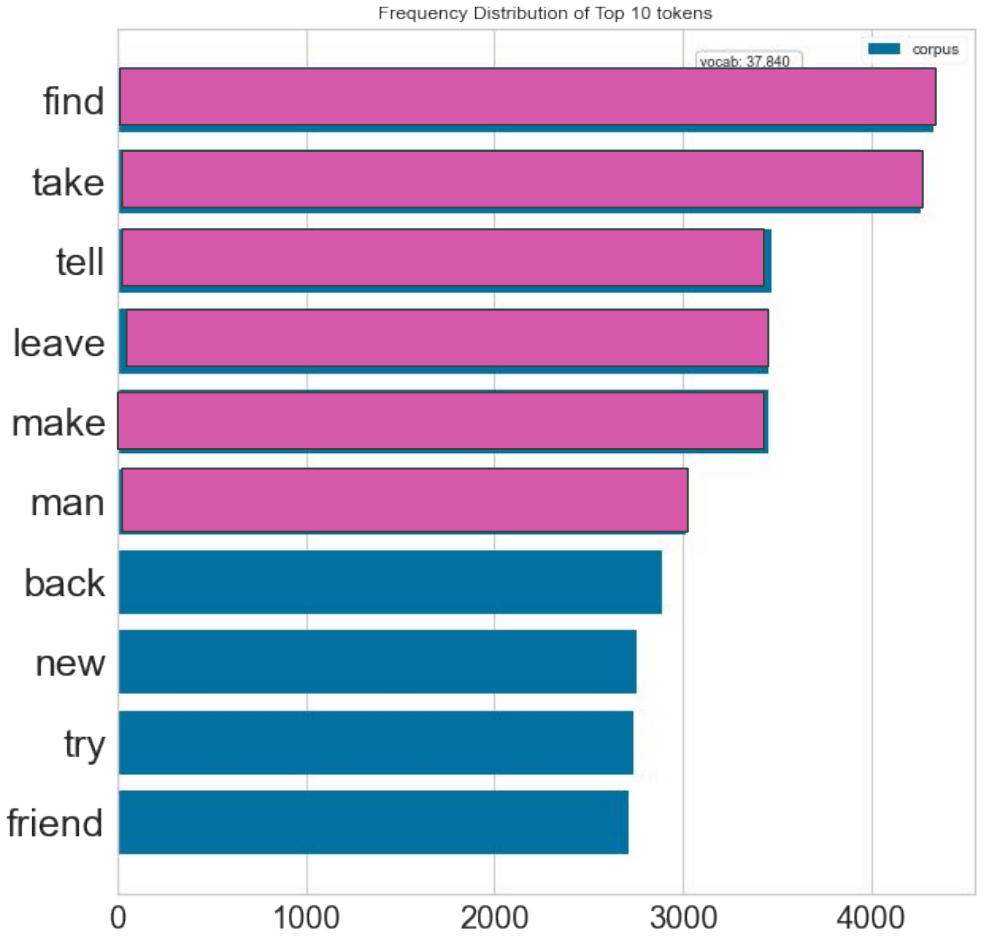
- Release years range 1904 – 2017
- Drama is the most popular in most decades



Drama : Frequency Distribution of Top 10 Words in Plot Summary

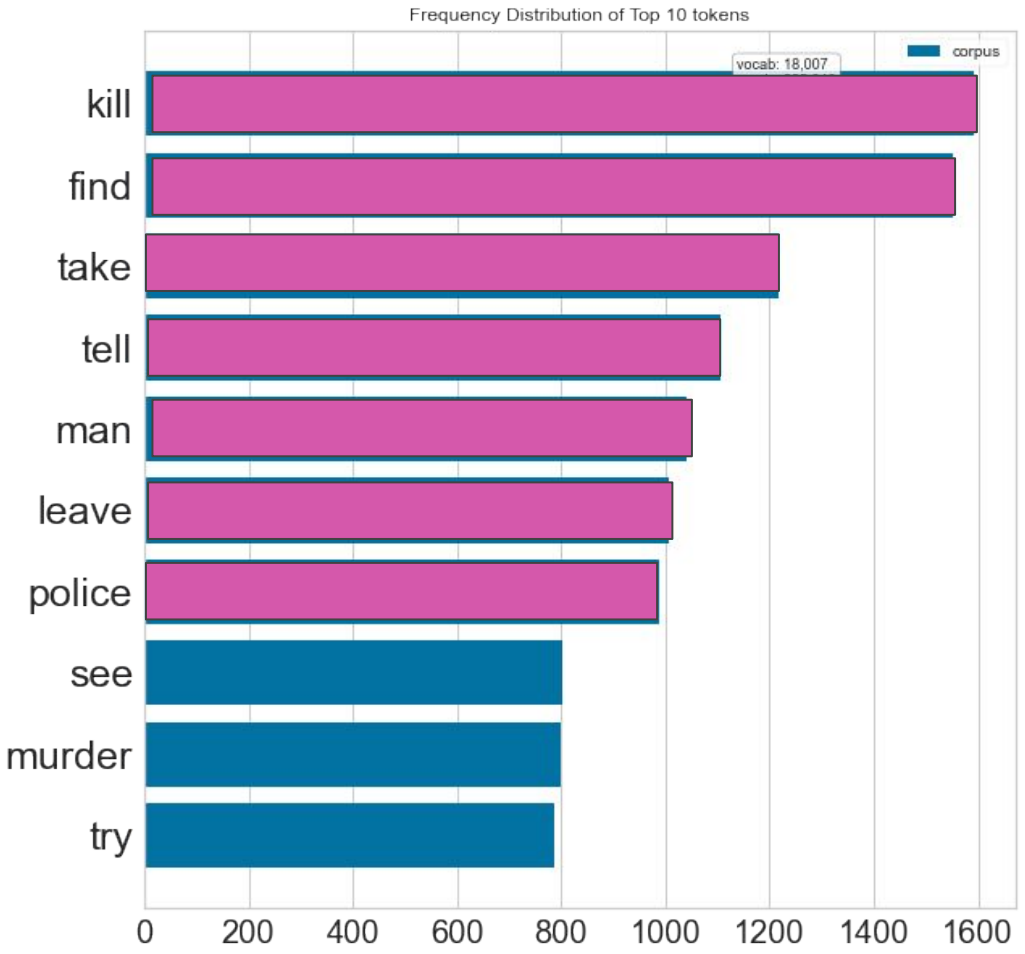


Comedy : Frequency Distribution of Top 10 Words in Plot Summary

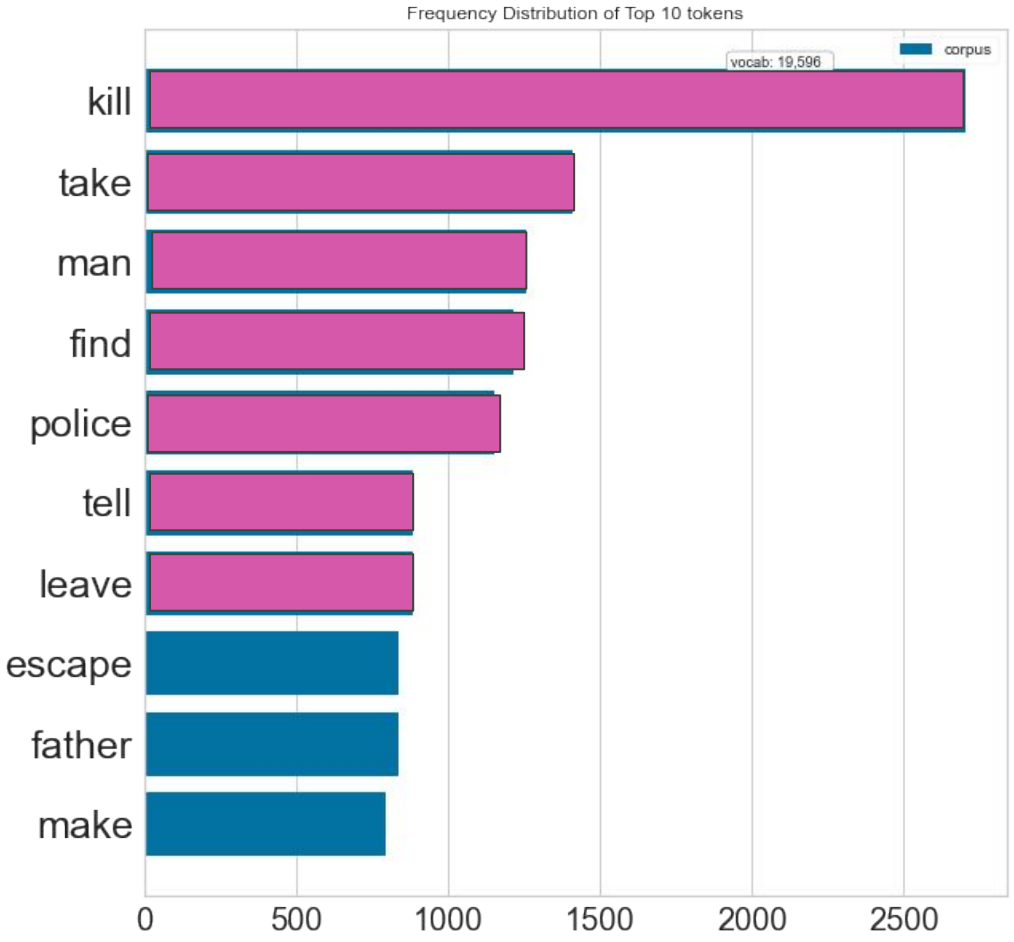


Common words: take, find, tell, leave, man, make

Thriller : Frequency Distribution of Top 10 Words in Plot Summary

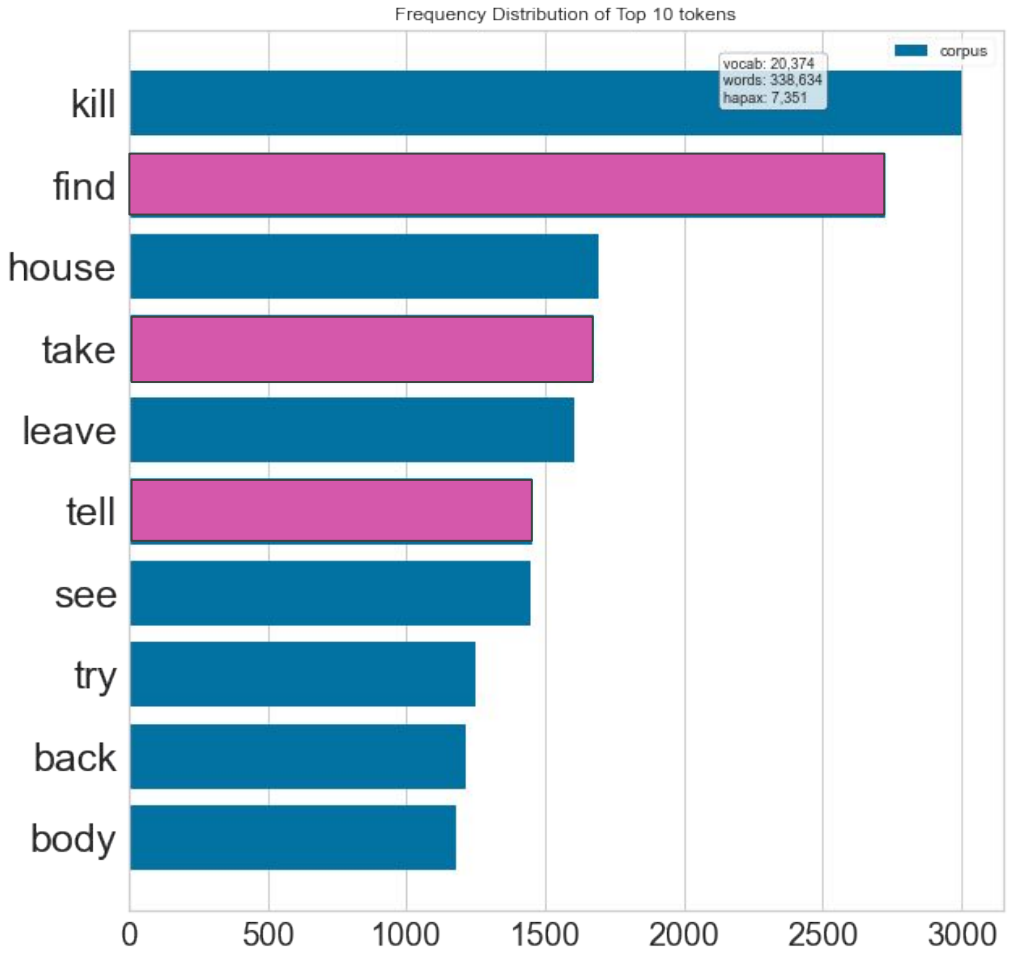


Action : Frequency Distribution of Top 10 Words in Plot Summary

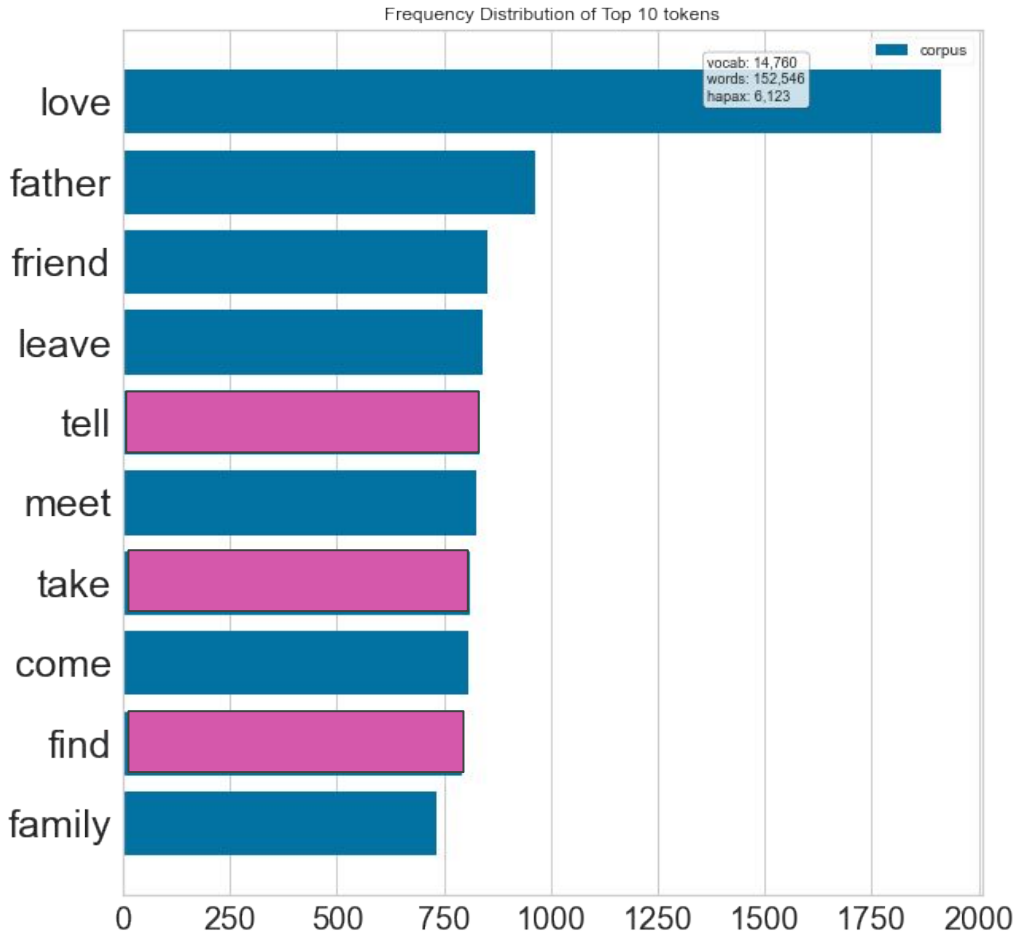


Common words : kill, find, police, tell, leave, take, man

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

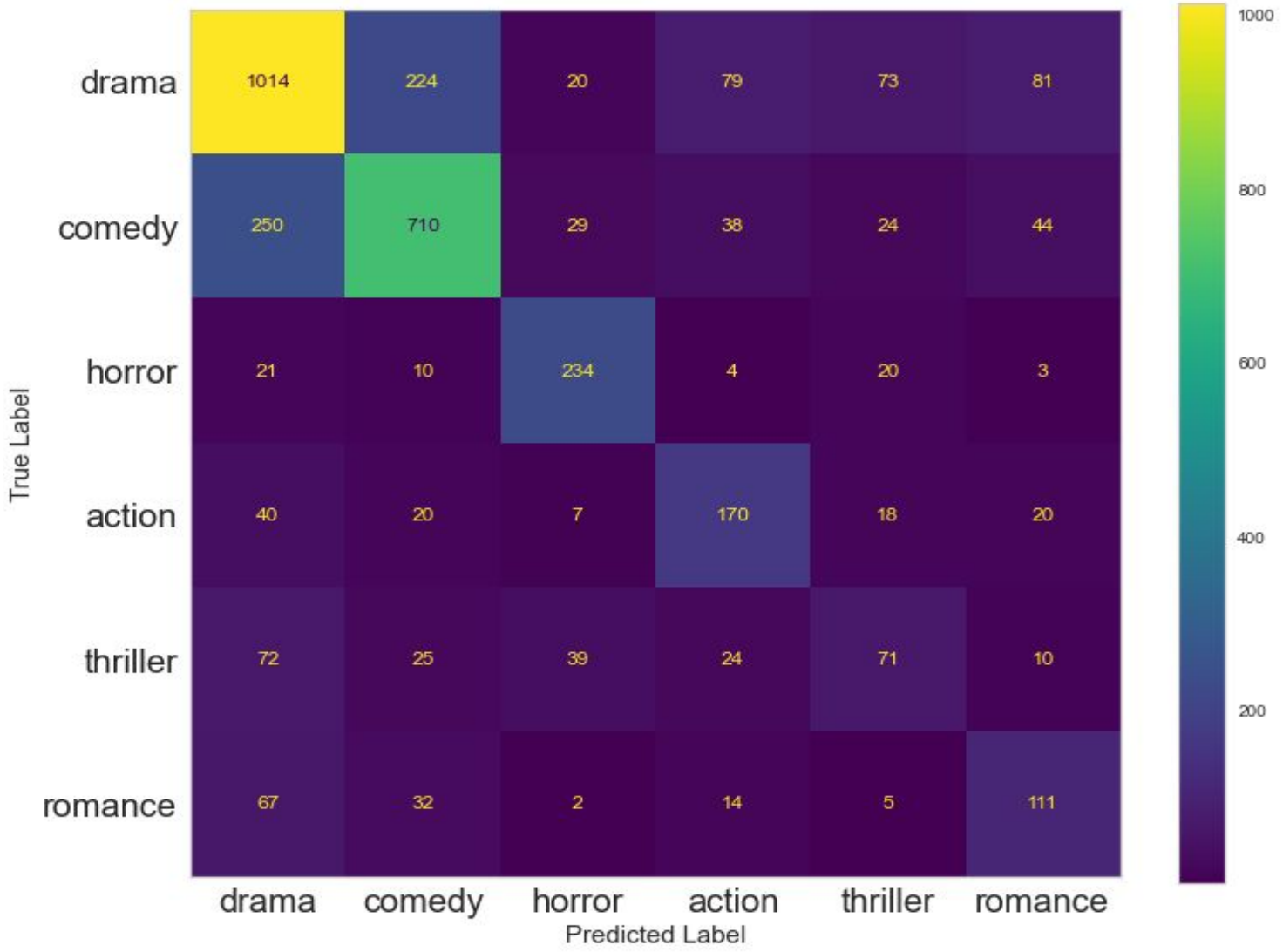
Modeling and Results

- Baseline accuracy of 41%
- Logistic regression and SGD models performed best during cross validation
- However the Logistic regression performed on the test set with an accuracy score was ~64%

	model	tfidf_score	count_vec_score
0	Multinomial Bayes	0.617551	0.603659
1	Logistic Regression	0.636129	0.615801
2	Decision Tree	0.468820	0.470843
3	SGD	0.636220	0.624724

Conclusion

- Classes were most commonly falsely predicted as drama.
- Drama and Comedy had the highest overlap because they have a lot words in common.



Thank you for Listening!

https://github.com/Wonuabimbola
http://www.linkedin.com/in/wonuola-abimbola
https://wonuolaa4.medium.com/