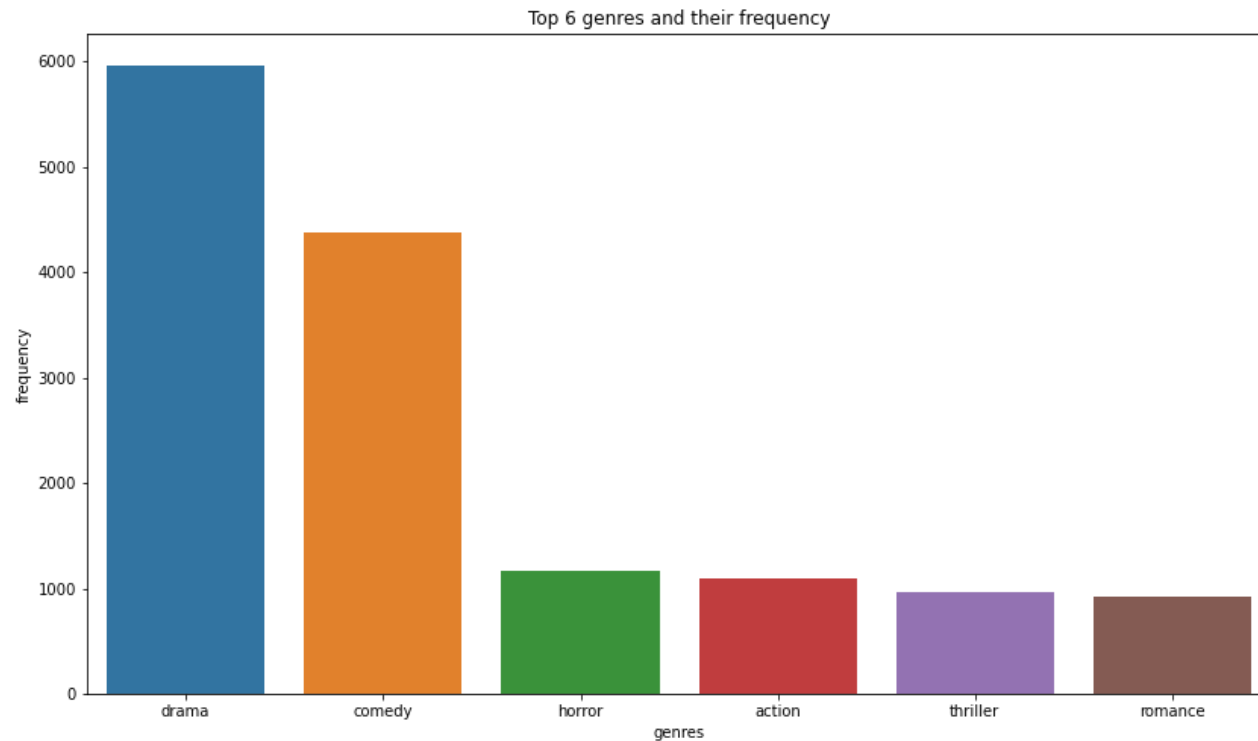# Text Classification of Movie Plot Summary to predict Movie Genre

WONUOLA ABIMBOLA

Business Problem

▶ Movies are one of the most popular means of entertainment.

▶ There are large volumes of movie data being generated and shared online

▶ The genre of a movie can be deciphered from its synopisis most of the time

▶ This project seeks to perform text classification of movie plot summaries in order to predict movie genres

- In this project, I will be performing NLP techniques on the movie plot summaries in order to use them predict the movie genres

- The dataset is from Kaggle which contains plot summary descriptions scraped from Wikipedia.

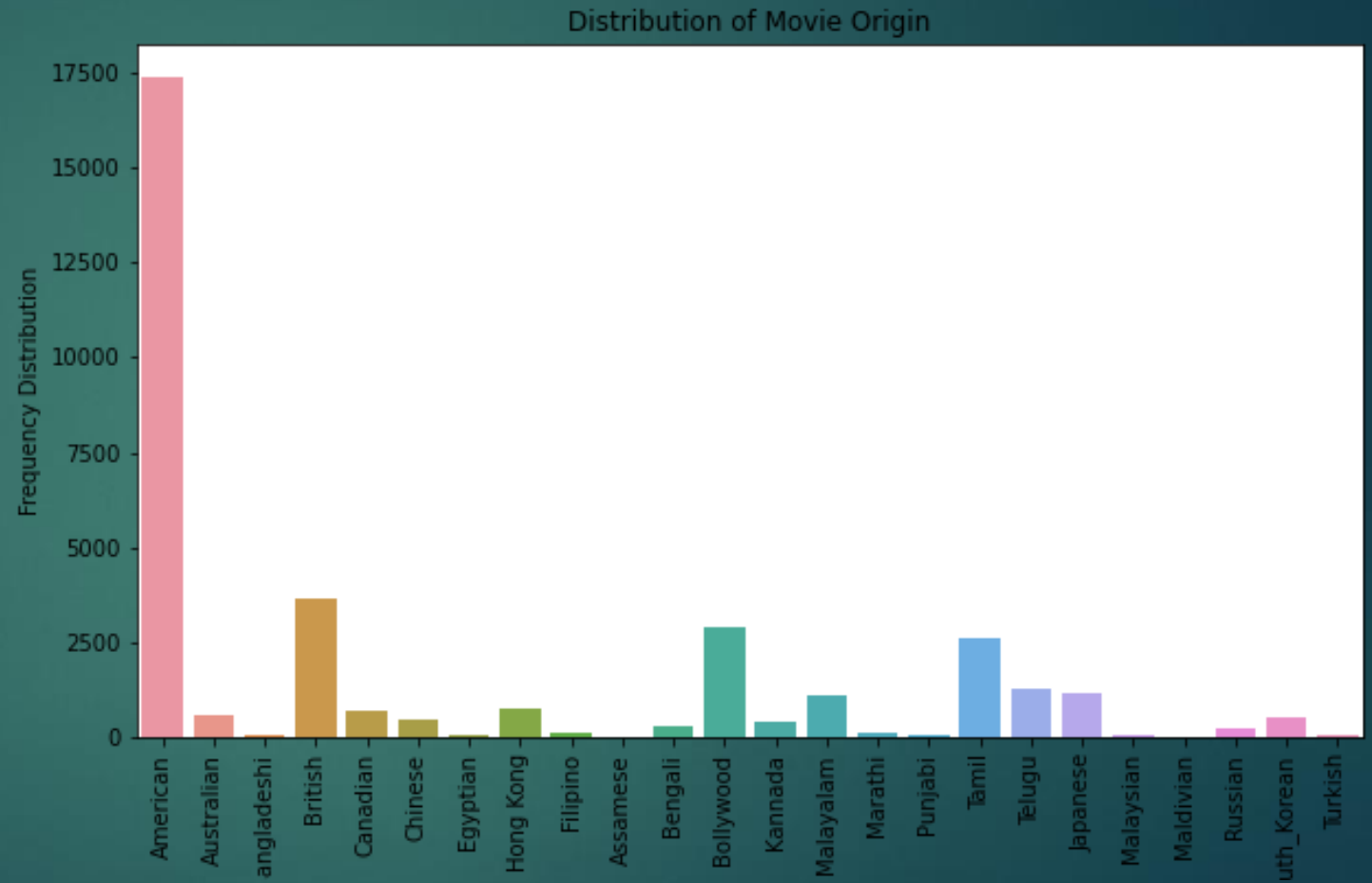- Automated movie genre predictions can have applications in recommendation systems as well

Top 6 genres and their frequency

Exploratory Data Analysis

DISTRIBUTION OF GENRES AMONG MOVIES THAT WERE ASSIGNED ONE GENRE

Distribution of Movie Origins

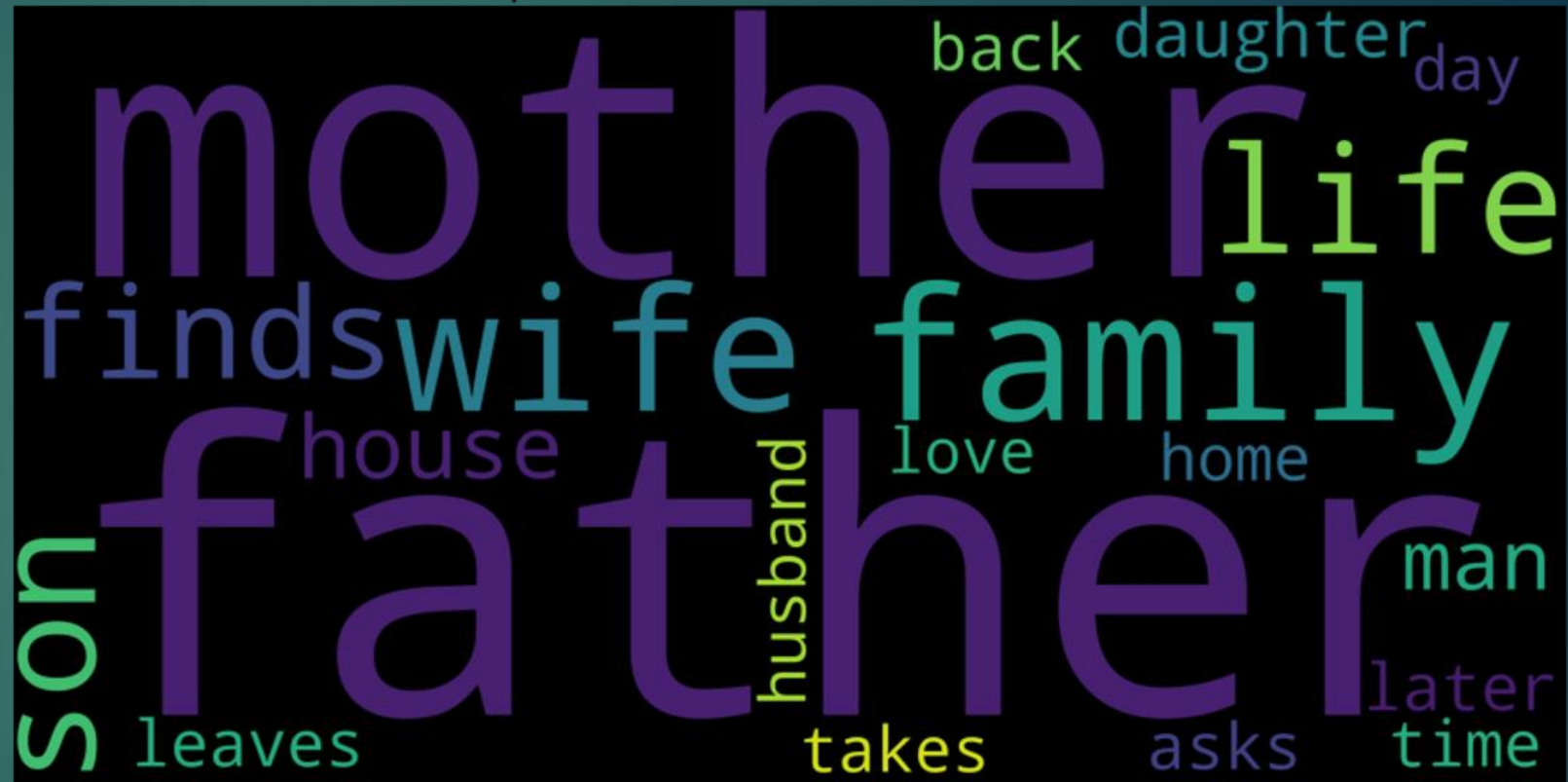Most of the movies in the dataset were American movies

The movie origin with the lowest frequency is Maldavian

Distribution of Movie Origin

Most Frequent words in Plot summary for Drama

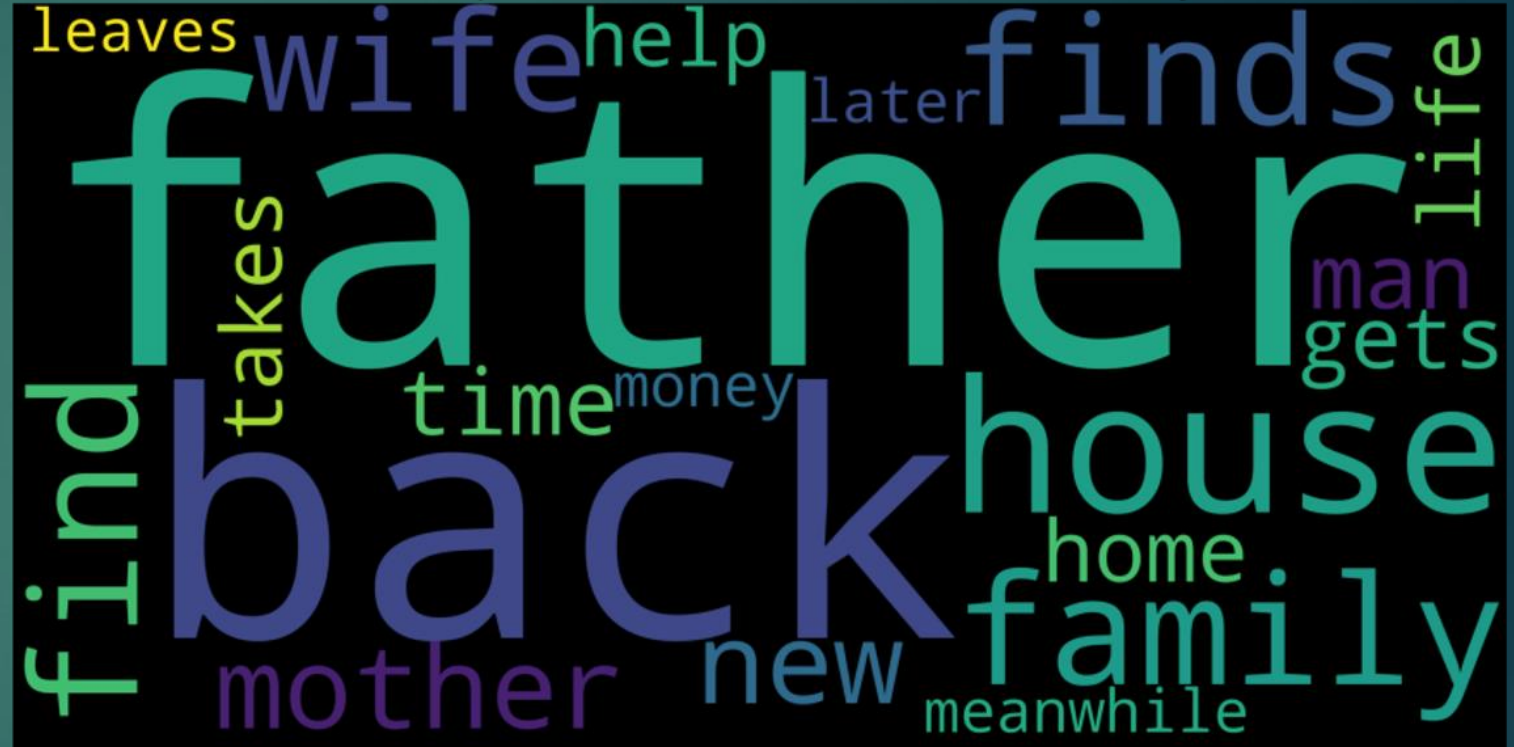- Looking at the wordcloud on the right, we see that drama genres commonly revolve around family, life

Most Frequent Words in the Movie Plot For Drama

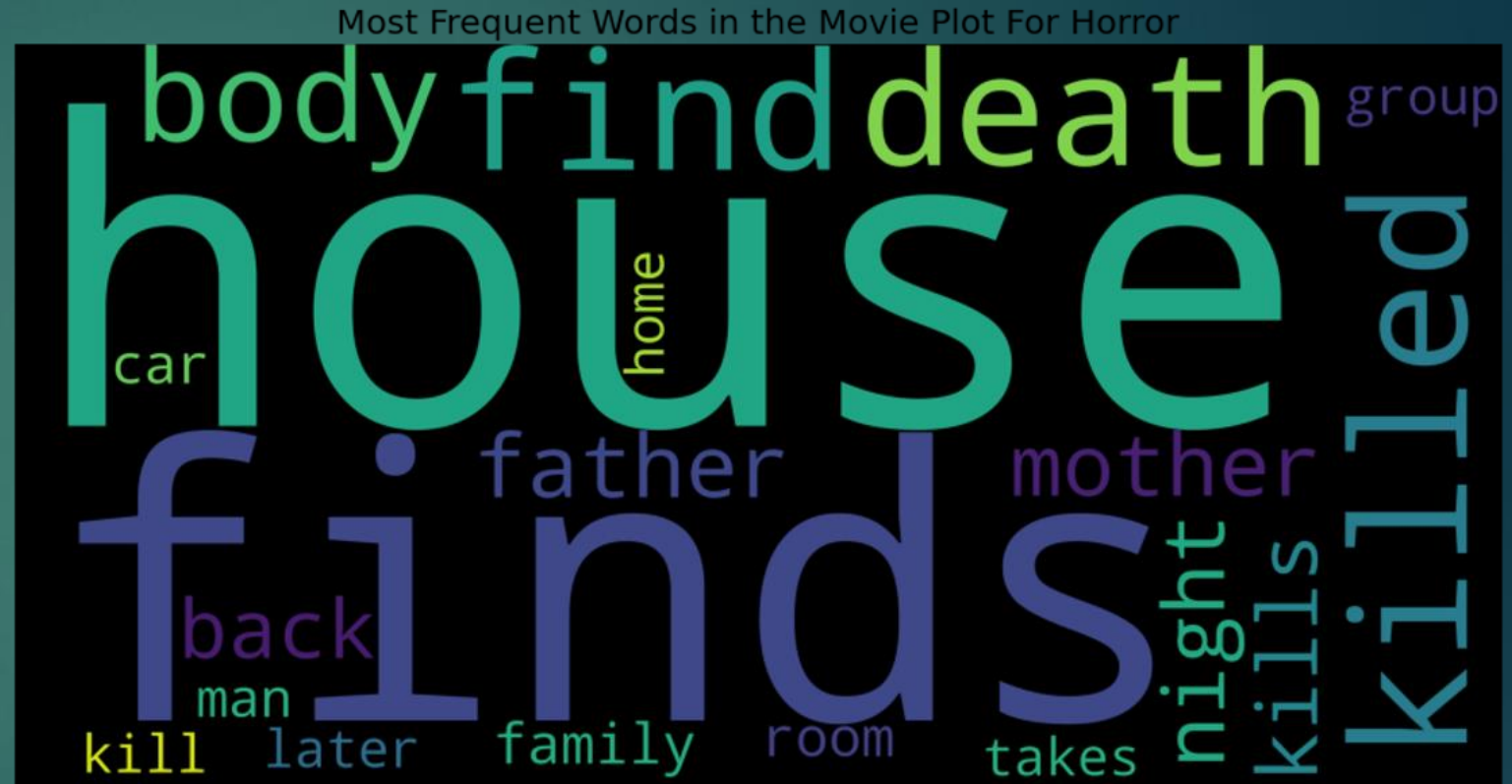# Most Frequent words in Plot summary for Comedy

- Most common words here are mostly similar to those in drama e.g. family, life, father.

Most Frequent Words in the Movie Plot For Comedy

Most Frequent words in Plot summary for Horror

- The horror genre seems to have words like 'kill'/'killed', 'death' and 'body commonly mentioned in the plot summaries

Most Frequent Words in the Movie Plot For Horror

Most Frequent words in Plot summary for Action

- The words 'killed', 'police' appear commonly in the plot summaries of this genre.

- There is also similarity in most common words between action and horror

Most Frequent Words in the Movie Plot For Action

Most Frequent words in Plot summary for Thriller

Thriller has words in common with the last two genres we've seen I.e. action and horror
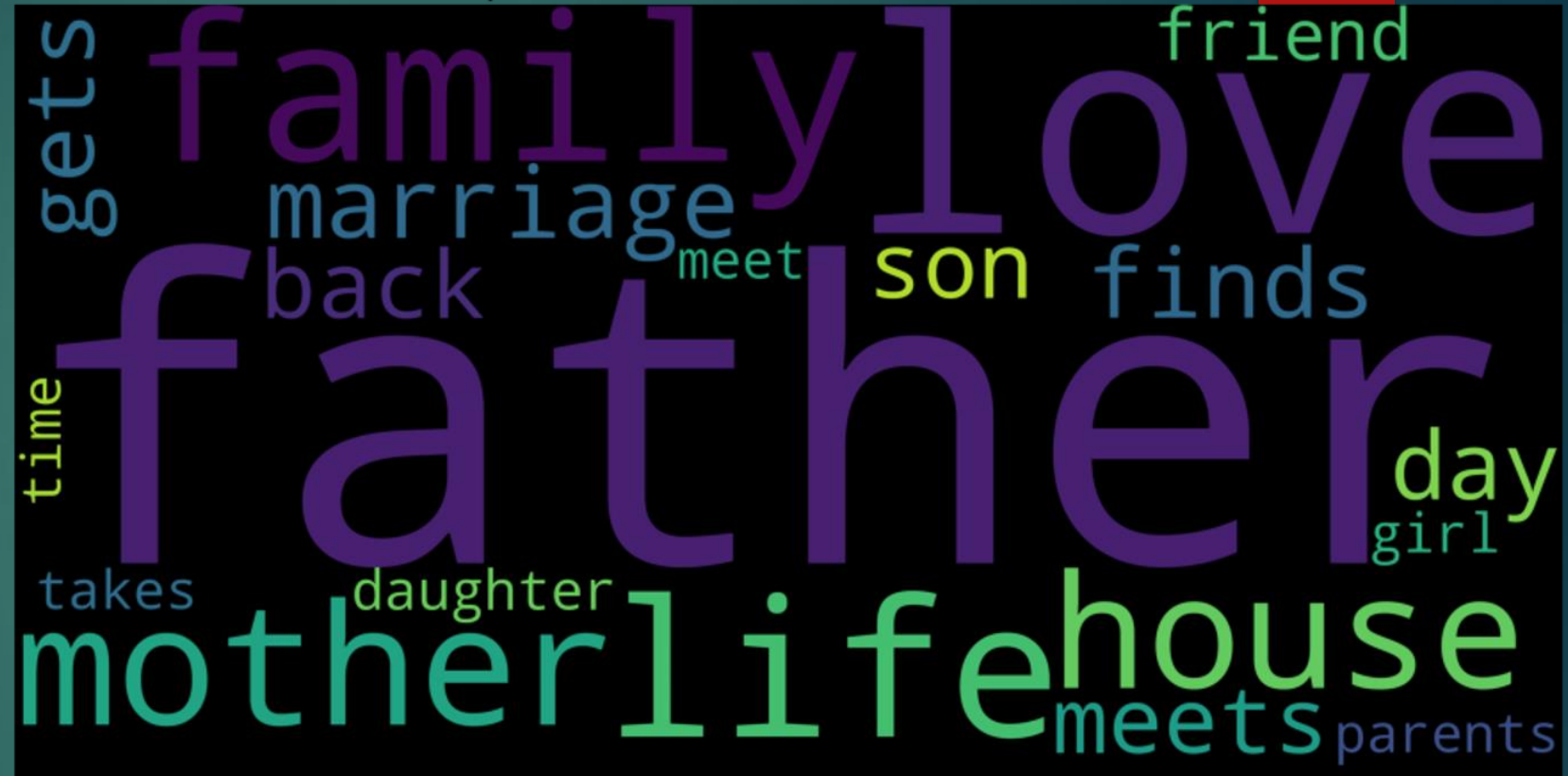

Most Frequent Words in the Movie Plot For Thriller

Most Frequent words in Plot summary for Romance

- Looking at the words in romance genre. It's no surprise that some of the most common words in the plot summaries is 'love', 'life', 'marriage'

Most Frequent Words in the Movie Plot For Romance

Modeling and Results

▶ The first simple model which predicted the most frequent class gave us an accuracy score of 41% which served as the baseline

▶ After performing grid searches with the following models:

▪ Multinomial Naïve Bayes

▪ Logistic Regression

▪ Decision Tree

▪ Random Forest Classifier

Modeling and Results

- When evaluating all the models, the decision tree and random forest models performed least favorably

- The Logistic Regression Model performed best when used on the tfidf-transformed X variable with an accuracy score of ~63%; a significant increase from our baseline score!

- When this model was applied on the test set, the accuracy score was approximately 64%

Conclusion and Next Steps

- This is was very interesting project that was fun to tackle.

- I would like to try apply Neural Networks in this problem to possibly get more favorable results.