Text Classification of Movie Plot Summary to predict Movie Genre

WONUOLA ABIMBOLA

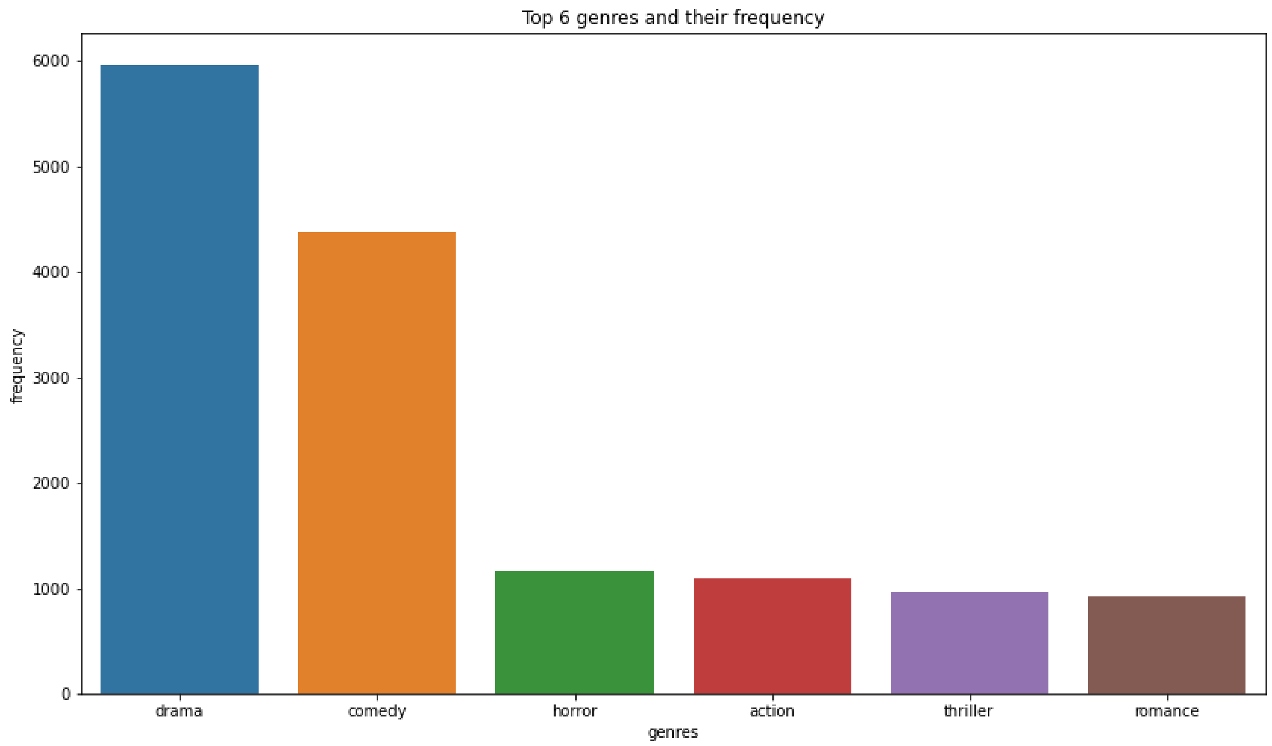
Overview

- In this project, NLP techniques were applied to movie plot summaries and used to predict movie genres

- The movie dataset used is from Kaggle . It contains plot summary descriptions scraped from Wikipedia.

Overview

- Large volumes of movie data online
- Popular means of entertainment
- Automated genre generation on movie streaming companies or websites like Netflix or Hulu



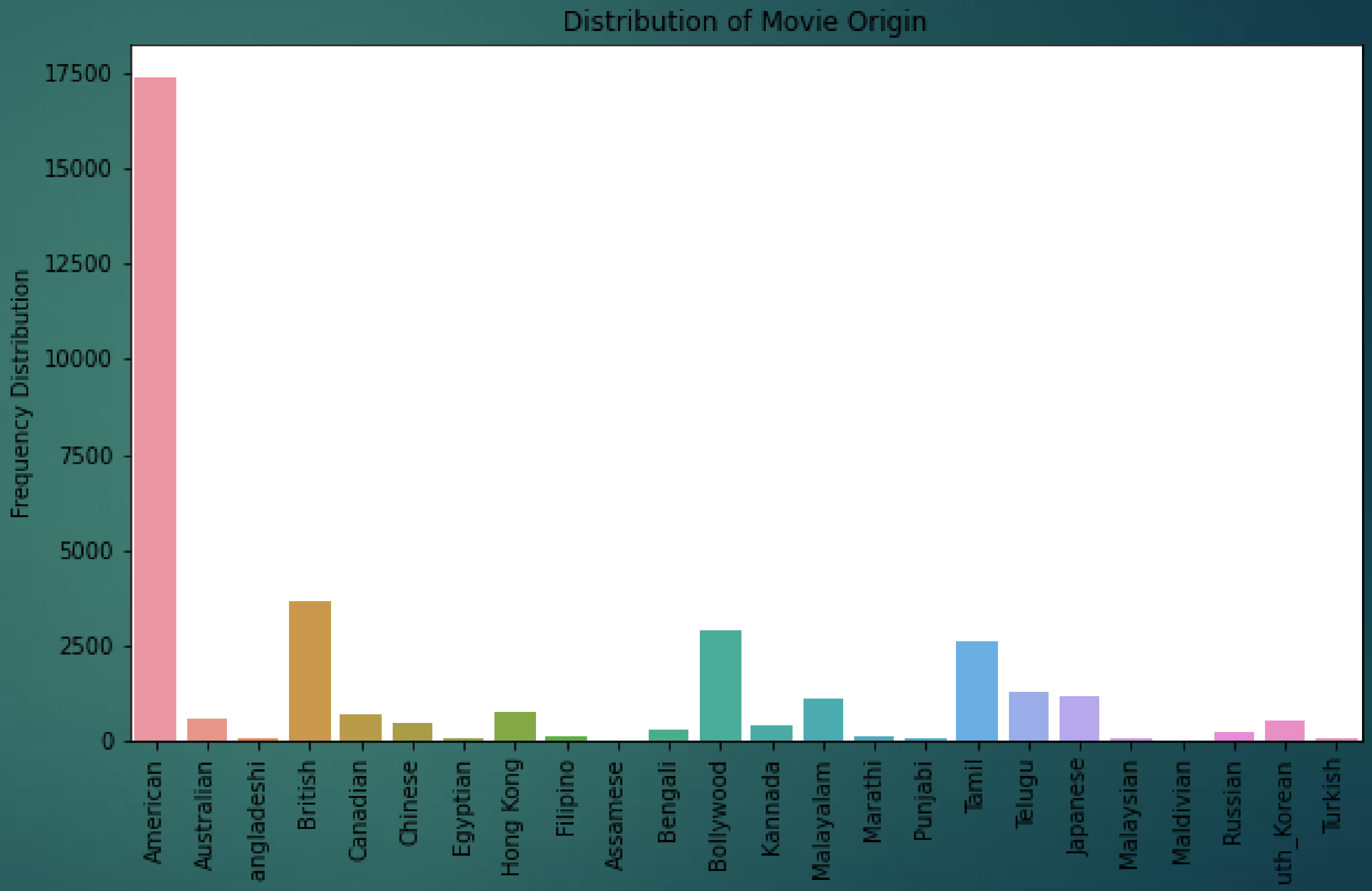
Exploratory Data Analysis

FREQUENCY DISTRIBUTION OF GENRE

Distribution of Movie Origins

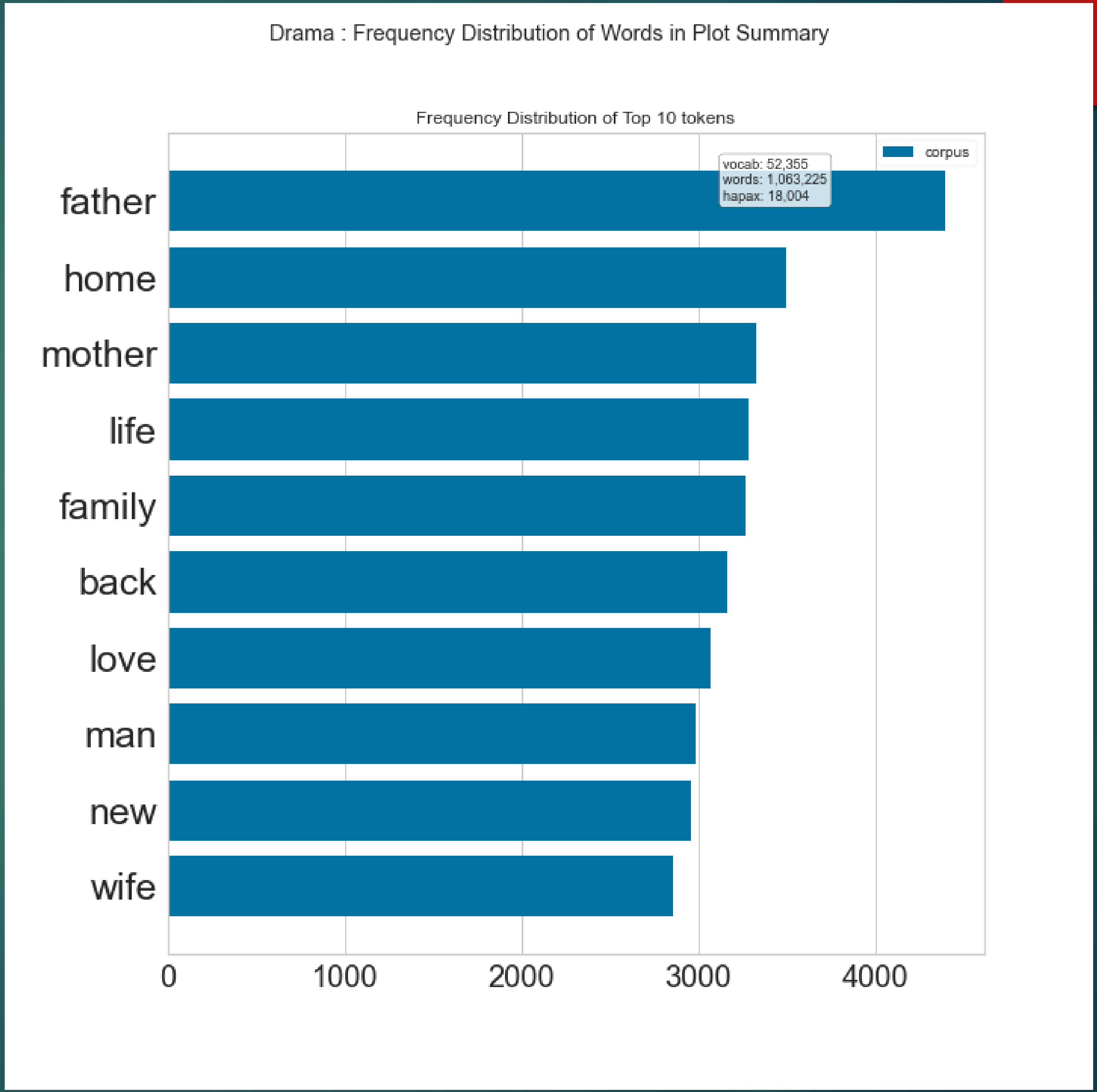
Most of the movies in the dataset were American movies

The movie origin with the lowest frequency is Maldavian



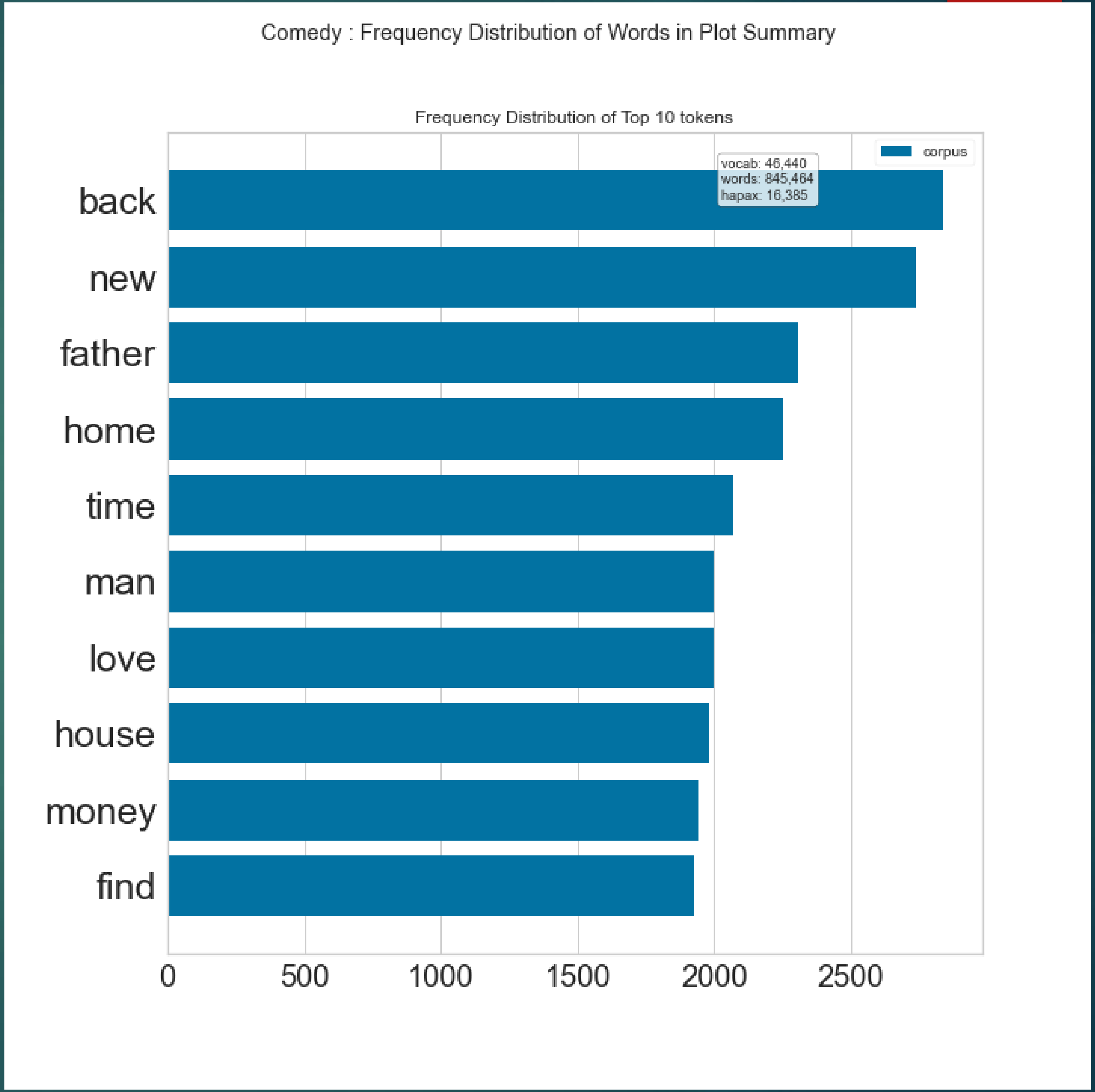
Most Frequent words in Plot summary for Drama

- Looking at the plot on the right, we see that drama genres commonly rev olve around family, life



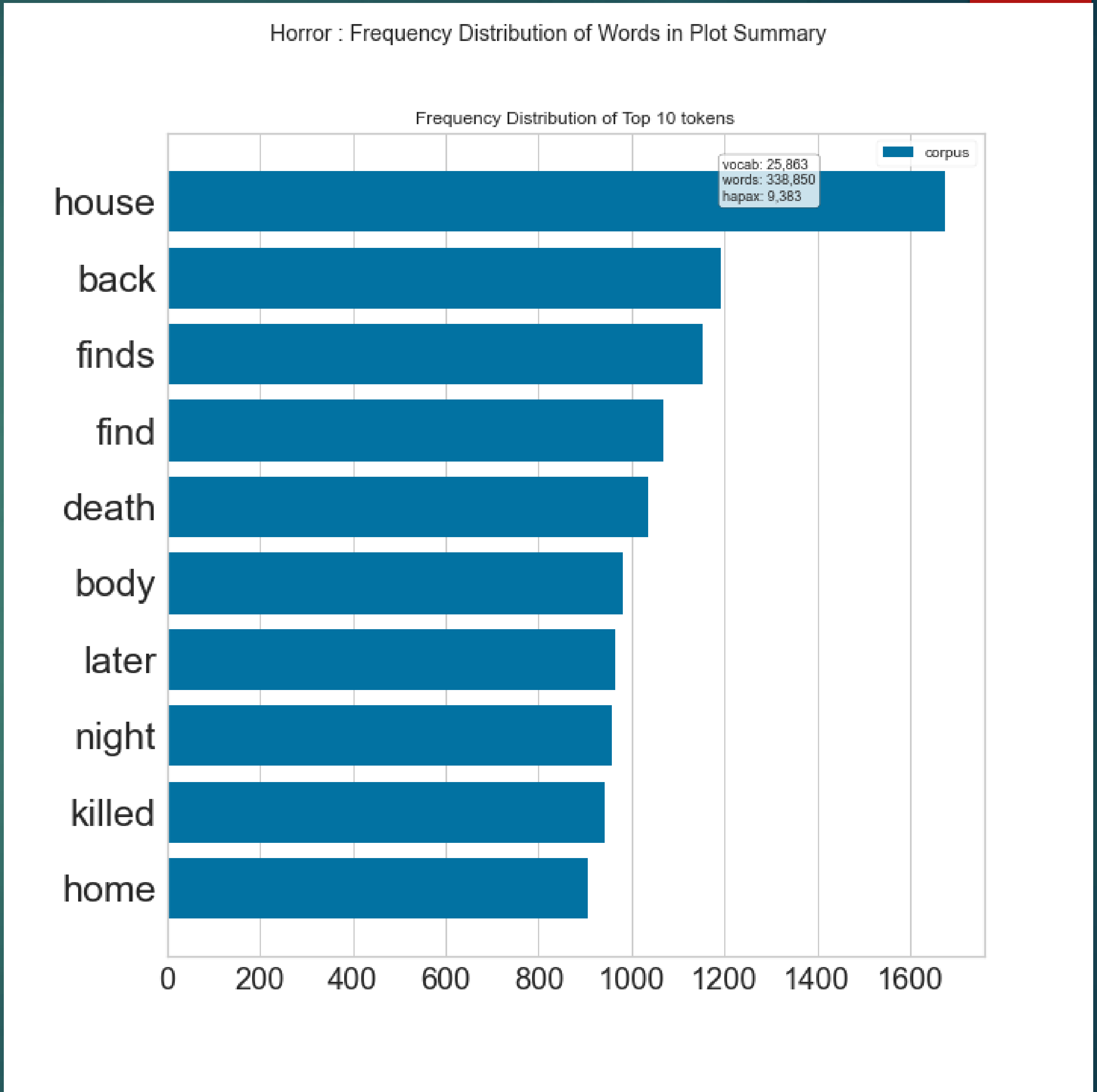
Most Frequent words in Plot summary for Comedy

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



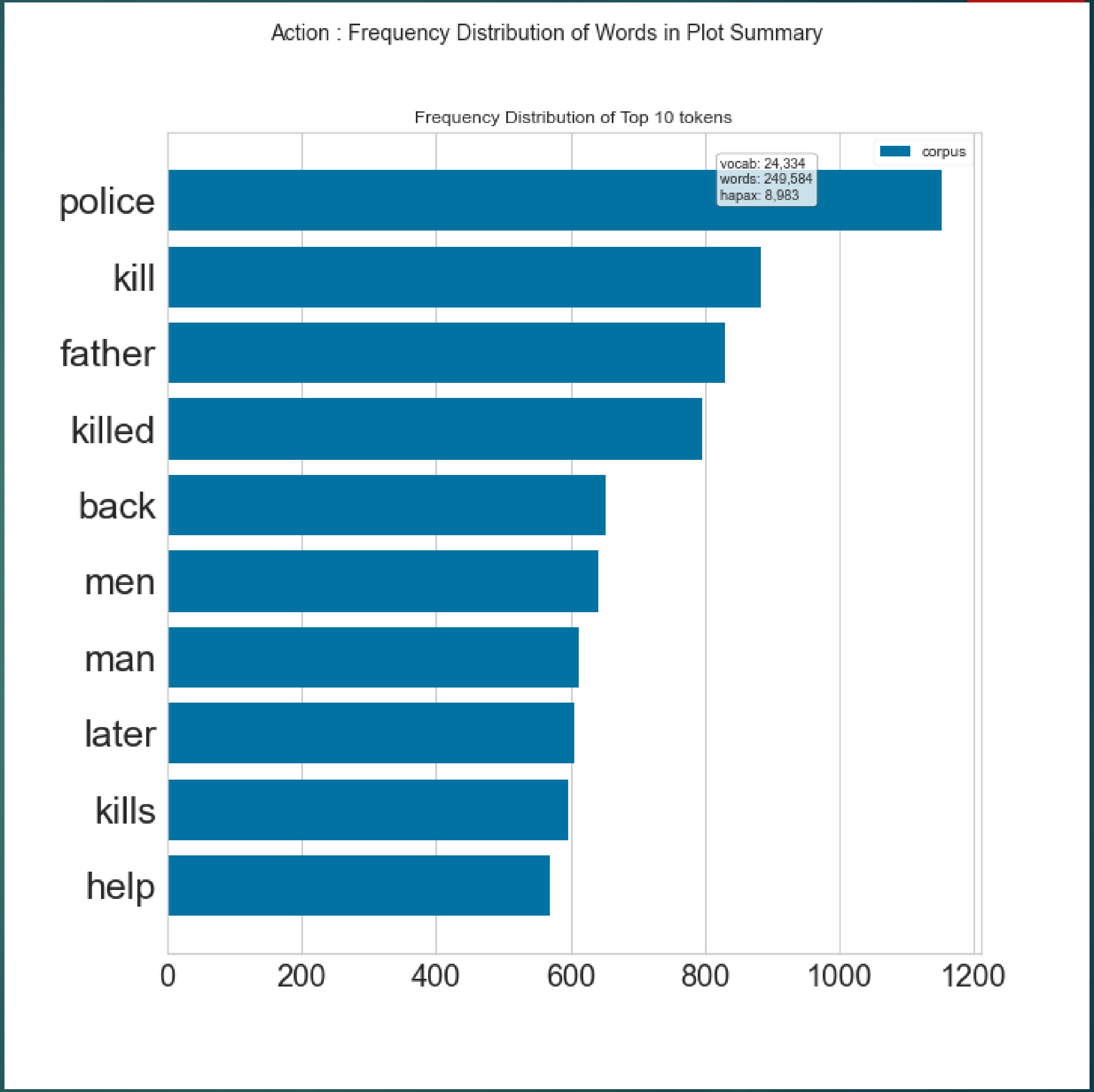
Most Frequent words in Plot summary for Horror

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



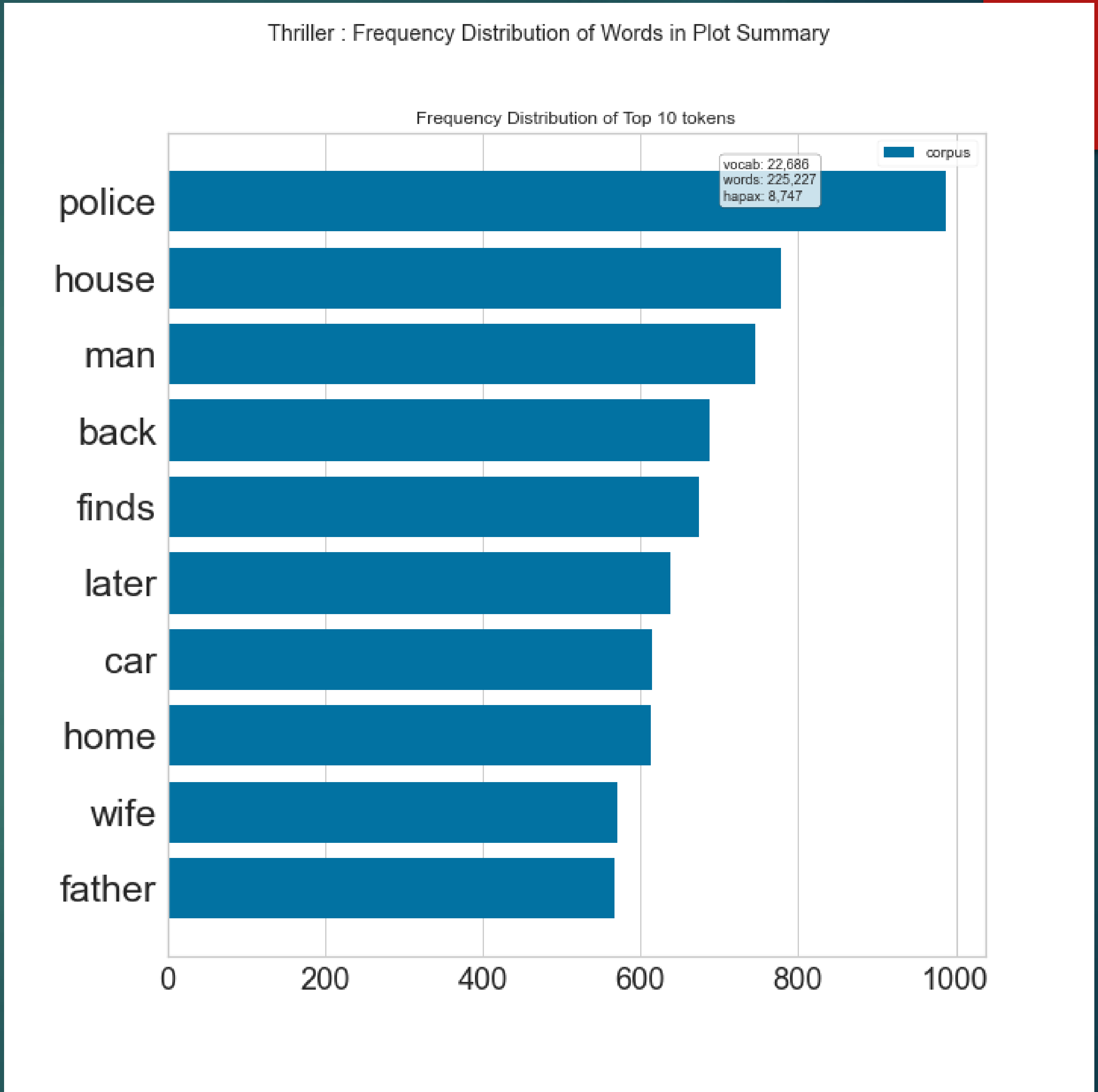
Most Frequent words in Plot summary for Action

- 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



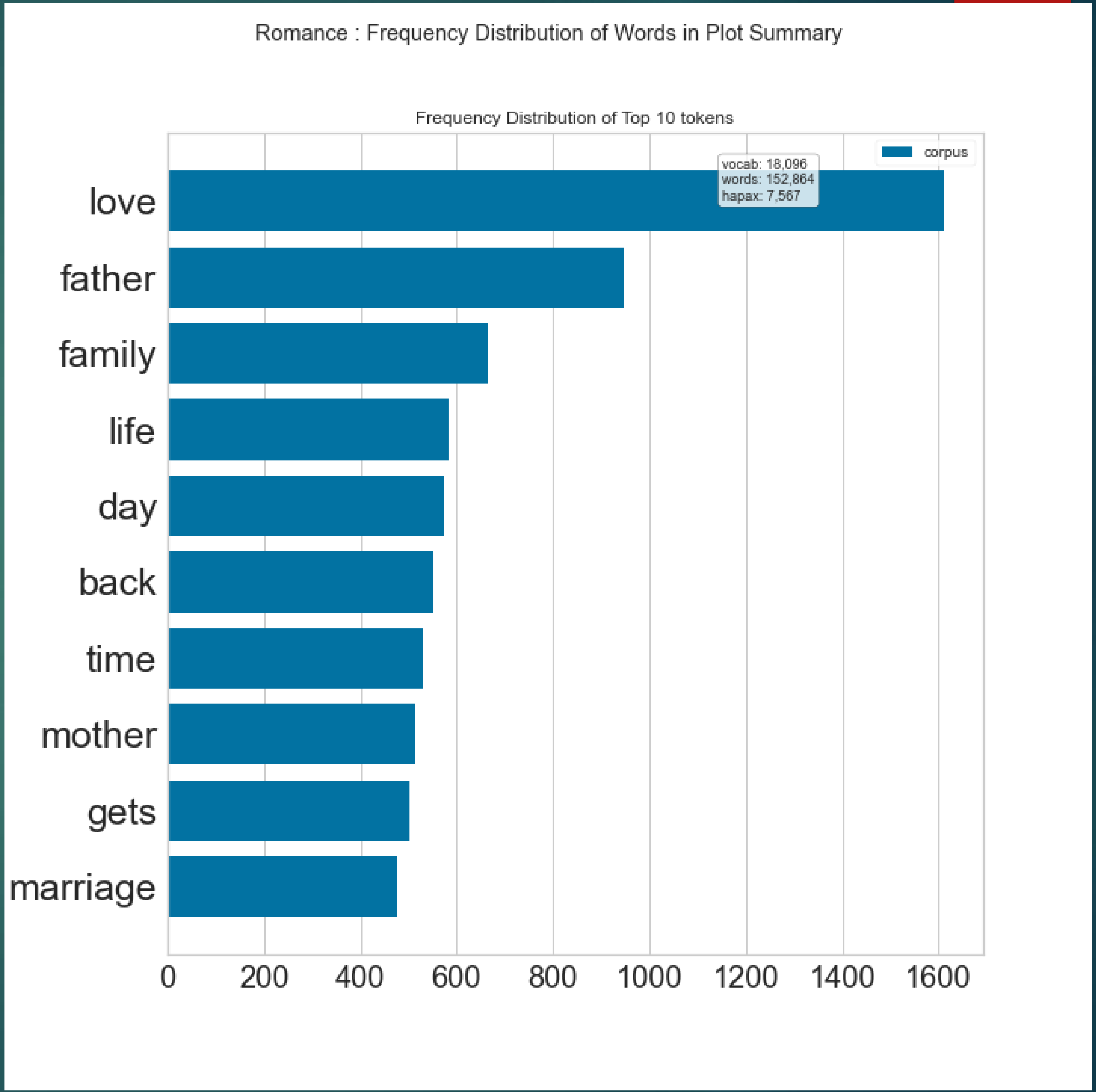
Most Frequent words in Plot summary for Thriller

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



Most Frequent words in 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 'lov e', 'life', 'marriage'


Romance : Frequency Distribution of Words in Plot Summary

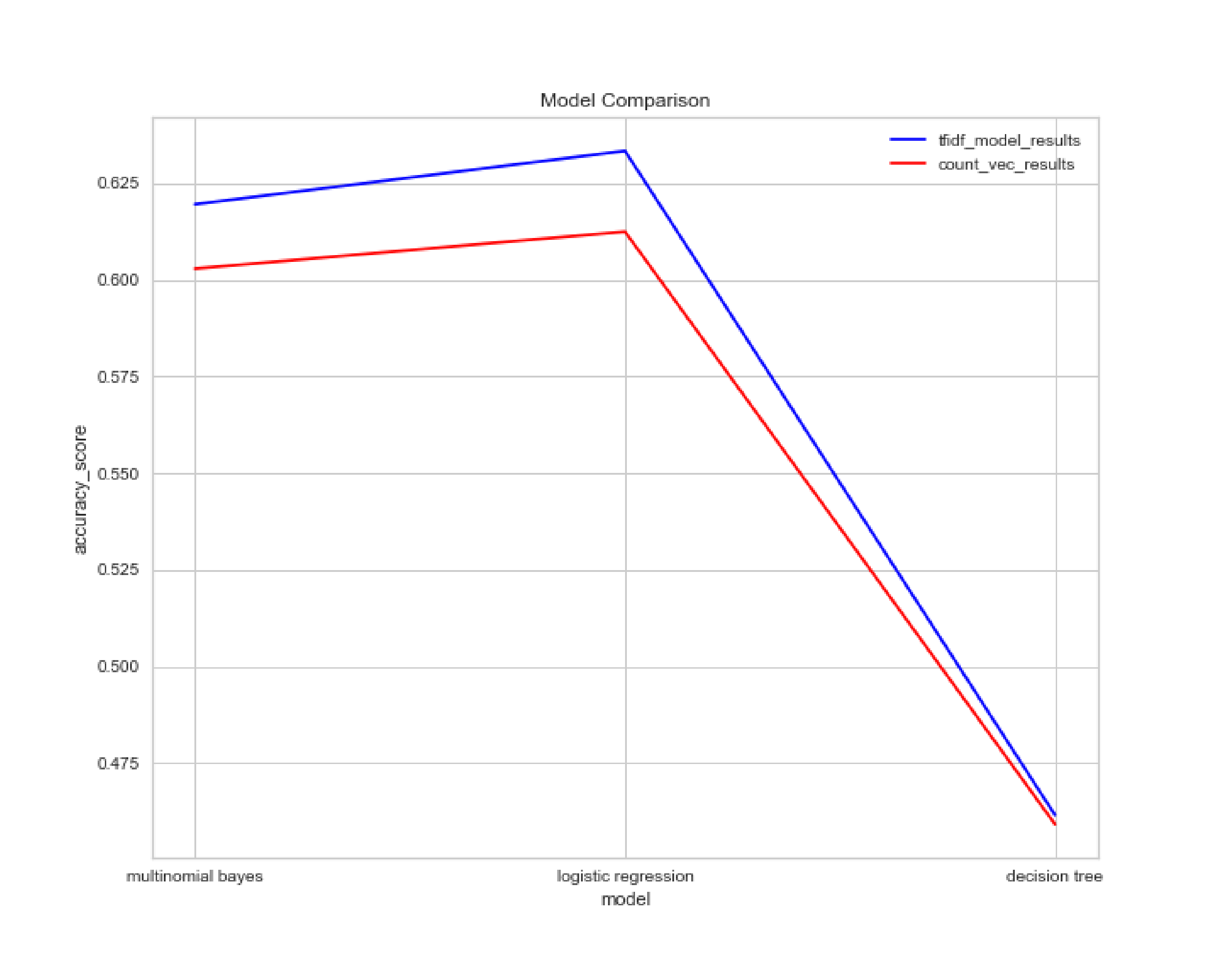
Modeling and Results

- The first simple model had an accuracy score of 41% which served as the baseline.

- Grid Searches were performing on the following models:
  - Multinomial Naïve Bayes
  - Logistic Regression
  - Decision Tree

Modeling and Results

- Decision tree model performed least favorably
- The Logistic Regression Model performed best with an accuracy score of ~63%;
- Test set accuracy score was ~64%



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.