SC1015 Project

Movie Genre Predictor

Presented by Ang Wai Kit, Anson and Chew You Chun
DSF3 Group 1

Problem Definition

- **To predict the genre of the movie from its overview**

Motivation

- Explore more about natural language processing
- Test whether machines can interpret human language

Exploratory Data Analysis and Data Cleaning

- Dataset was taken from Kaggle (tmdb_5000_movies) and the link to it is in our Github readme
- Cleaned genres and overview of our dataset as it will be used for machine learning later on
- Explored both numeric and categorical data

Data Cleaning

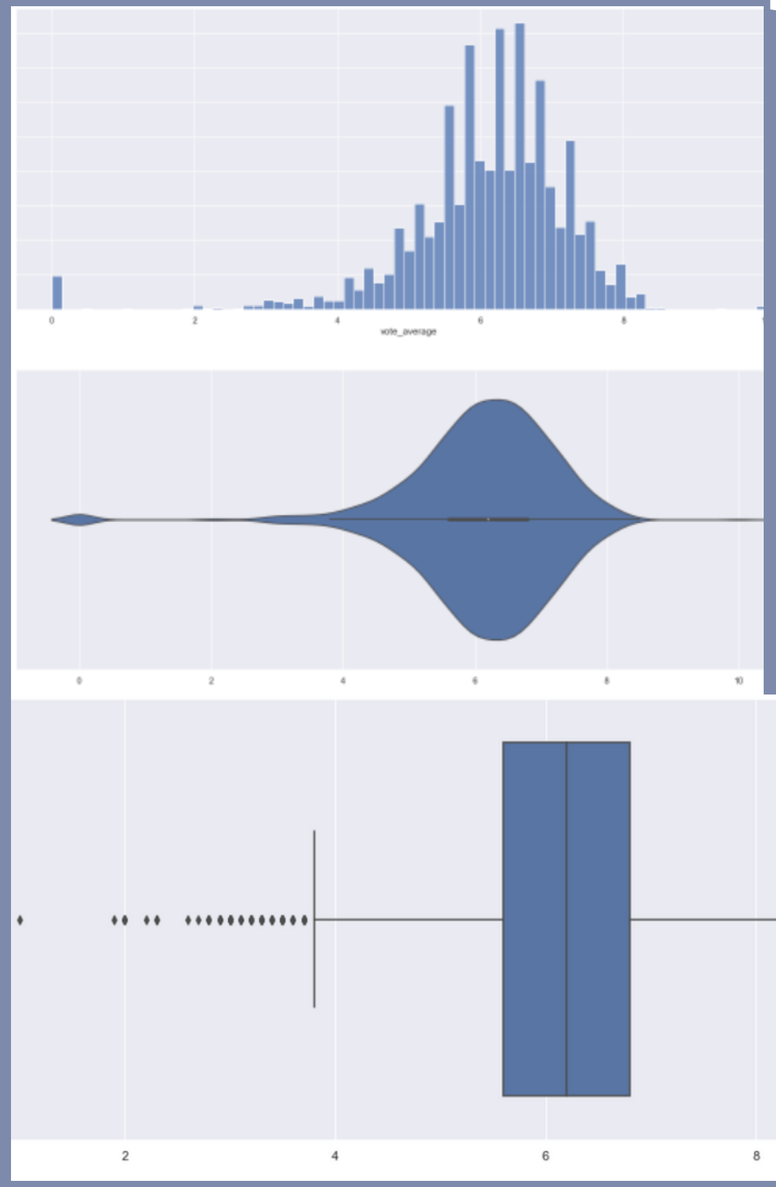
- For genres, our data consists of few dictionaries
- Split the data into several parts by delimiters
- Concatenate only the names of the genres back to the main dataframe

Data Cleaning

- Hard to analyse multiple genres based on one overview
- Further cleaned the data by only keeping the first genre
- Converted the release dates to years.

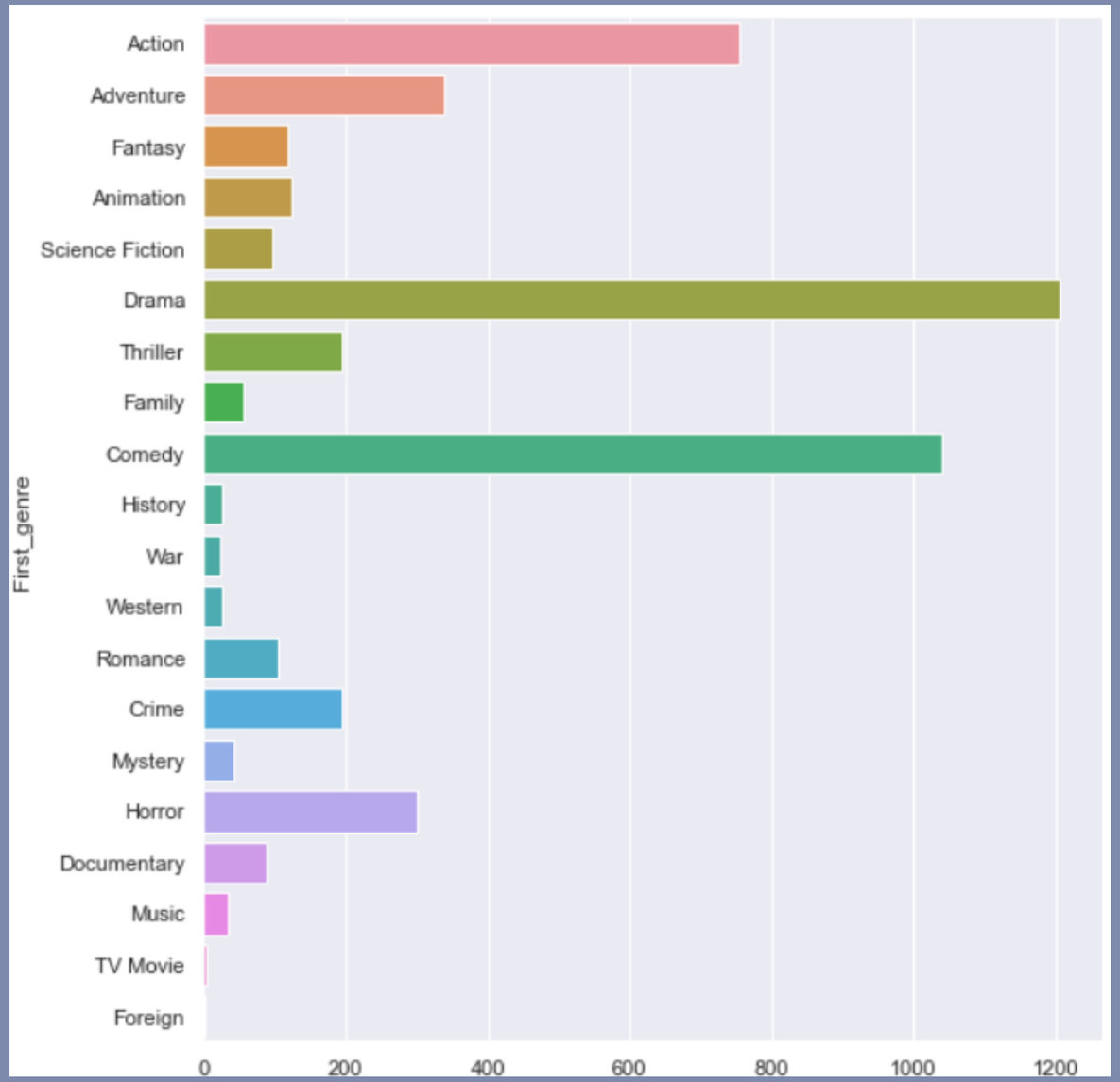
Exploratory Data Analysis

- First explored numeric data
- Used box-plots, histograms and violin plots to explore Vote average
- Negatively skewed



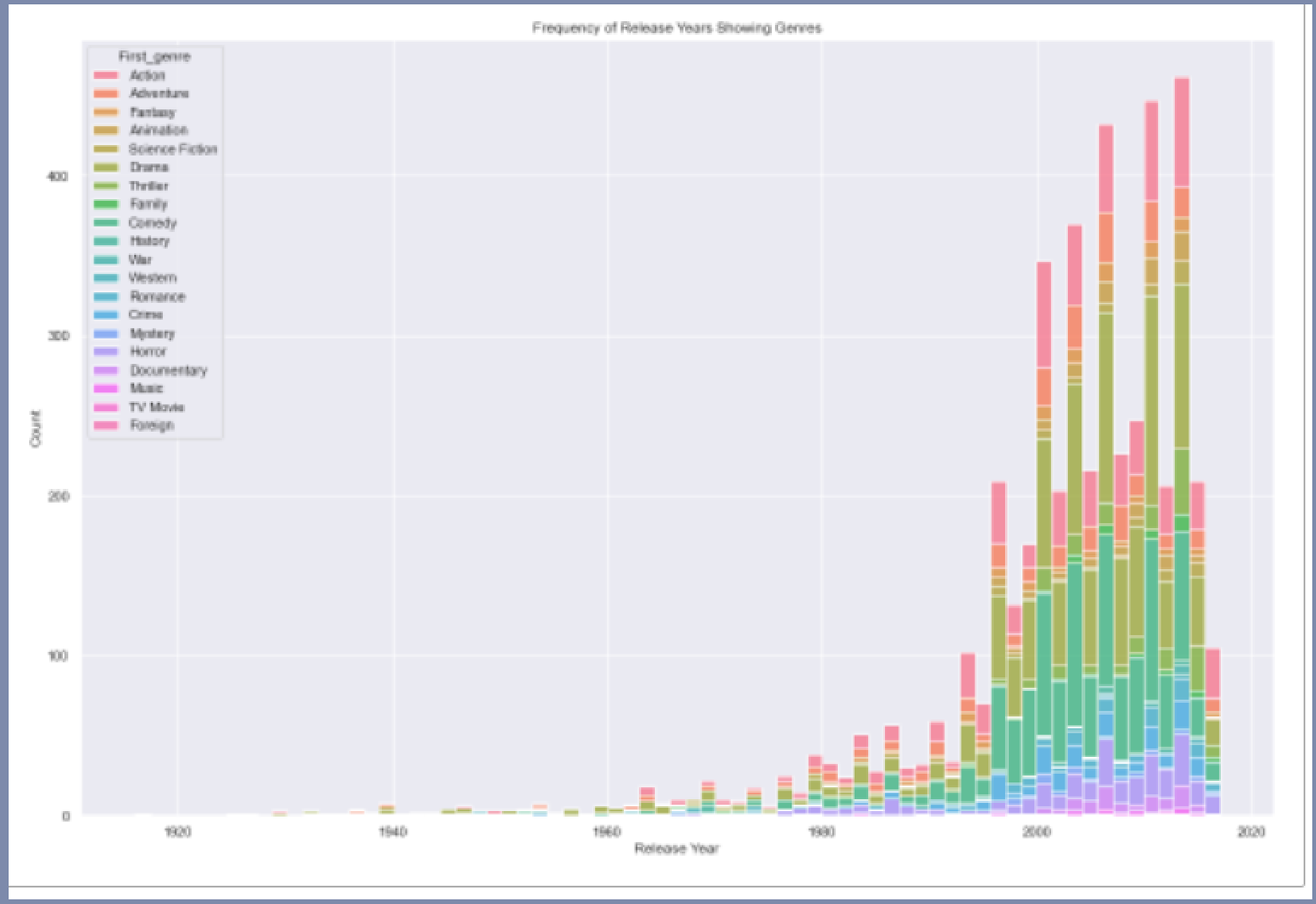
Exploratory Data Analysis

- Categorical Data used were Genres and Original Language
- Barplot used to visualise the number of movies with each different kinds of genres



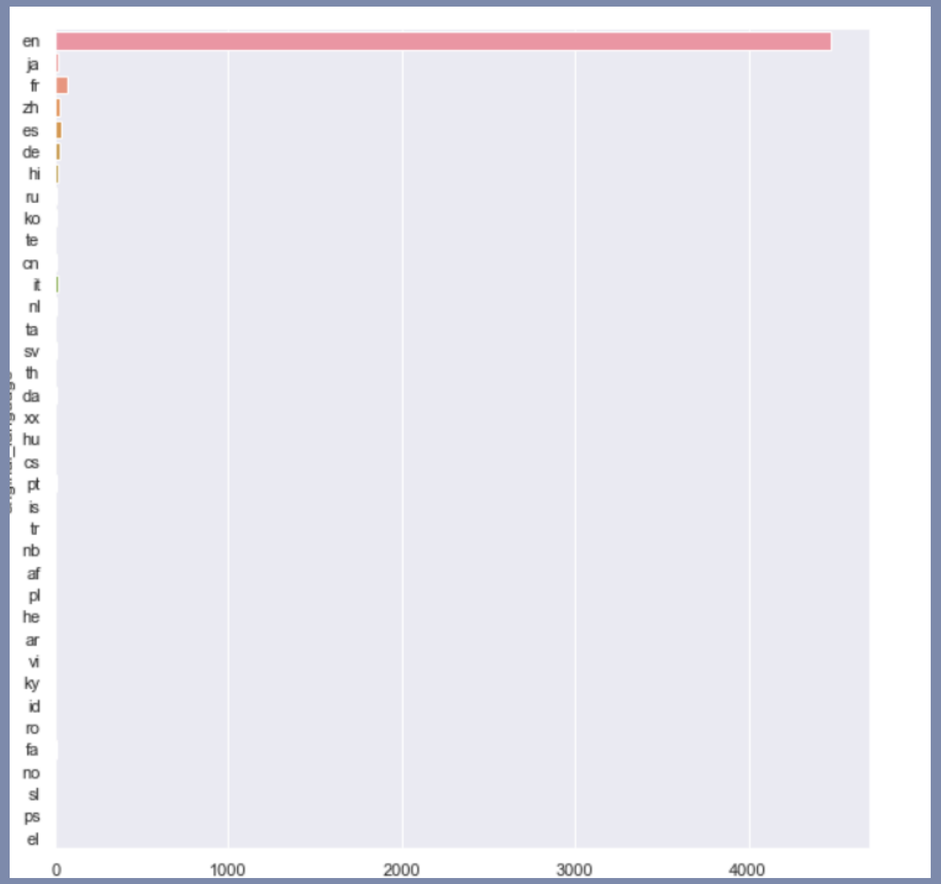
Exploratory Data Analysis

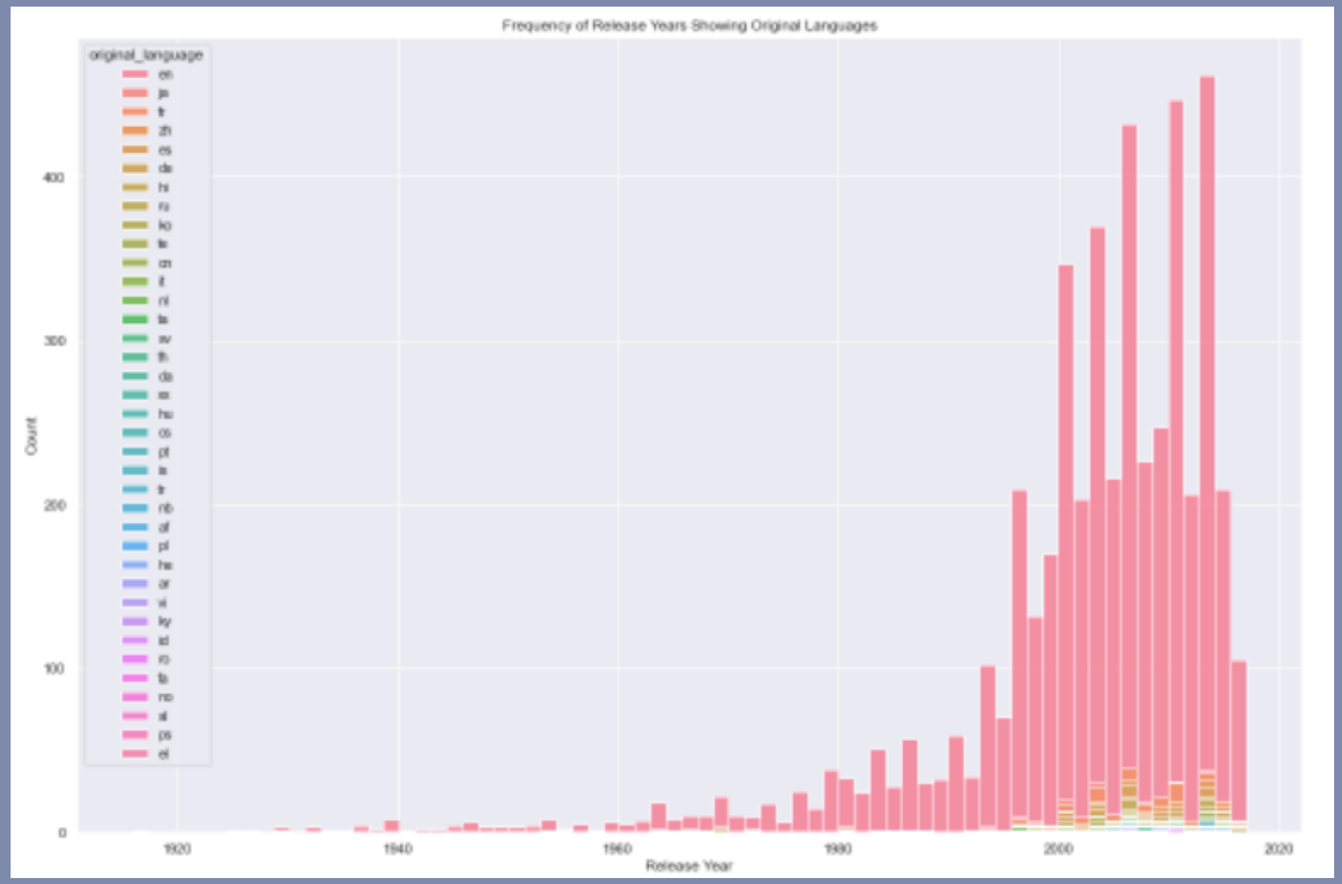
- Used stacked bar charts to compare genres across the years



Exploratory Data Analysis

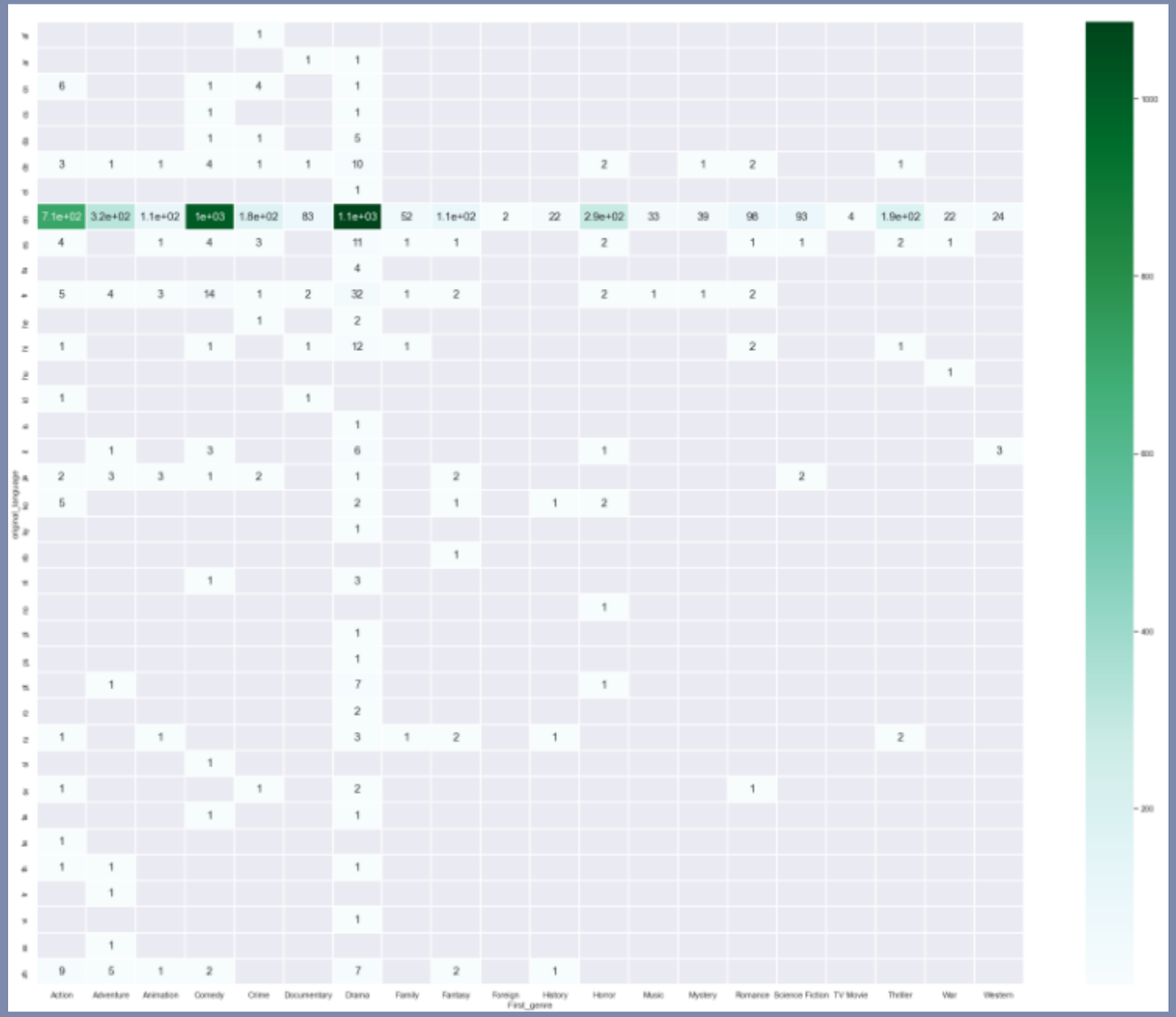
- Used the same approach as genres





Exploratory Data Analysis

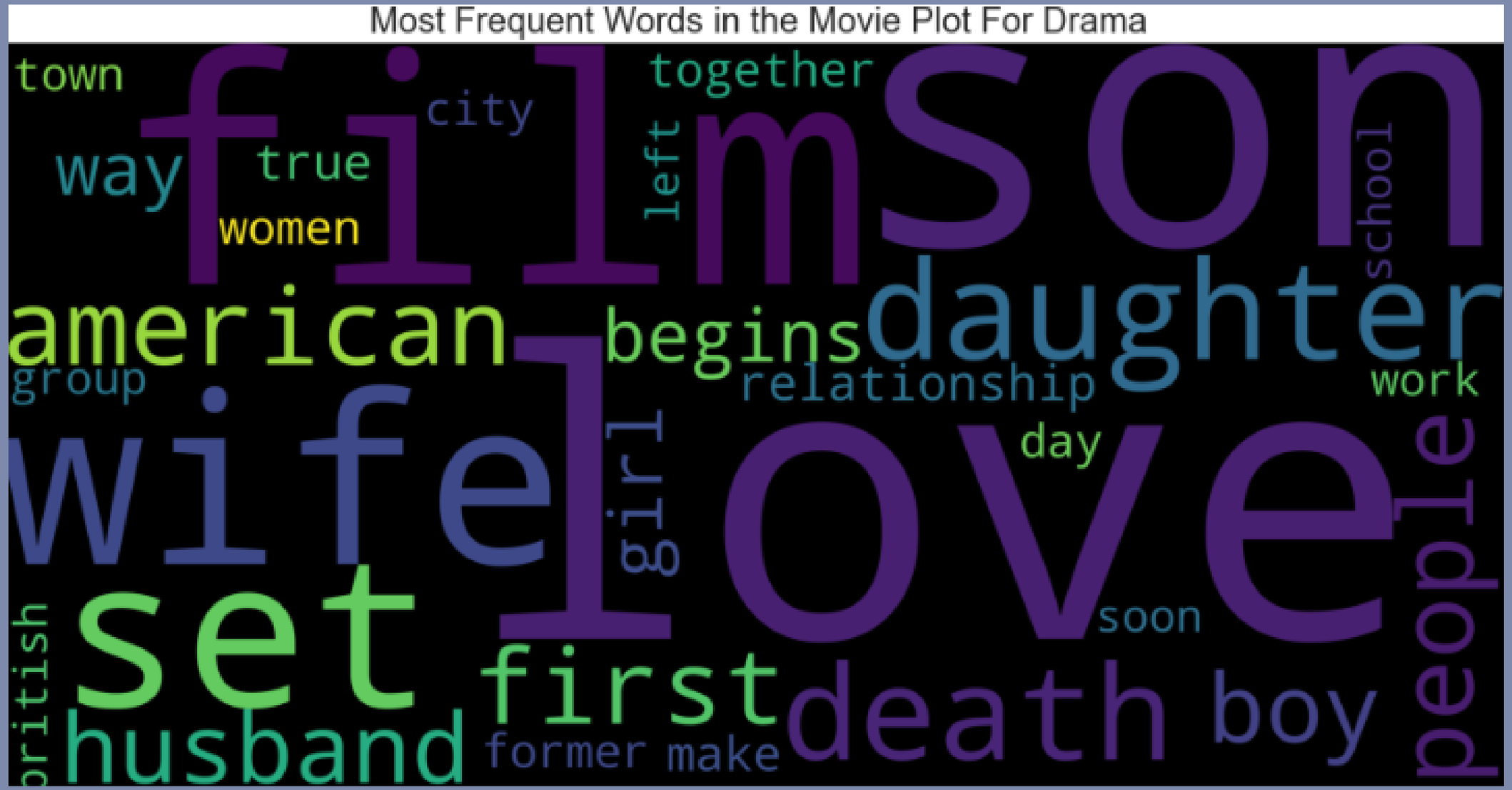
- Compared Genres against original languages using a heatmap



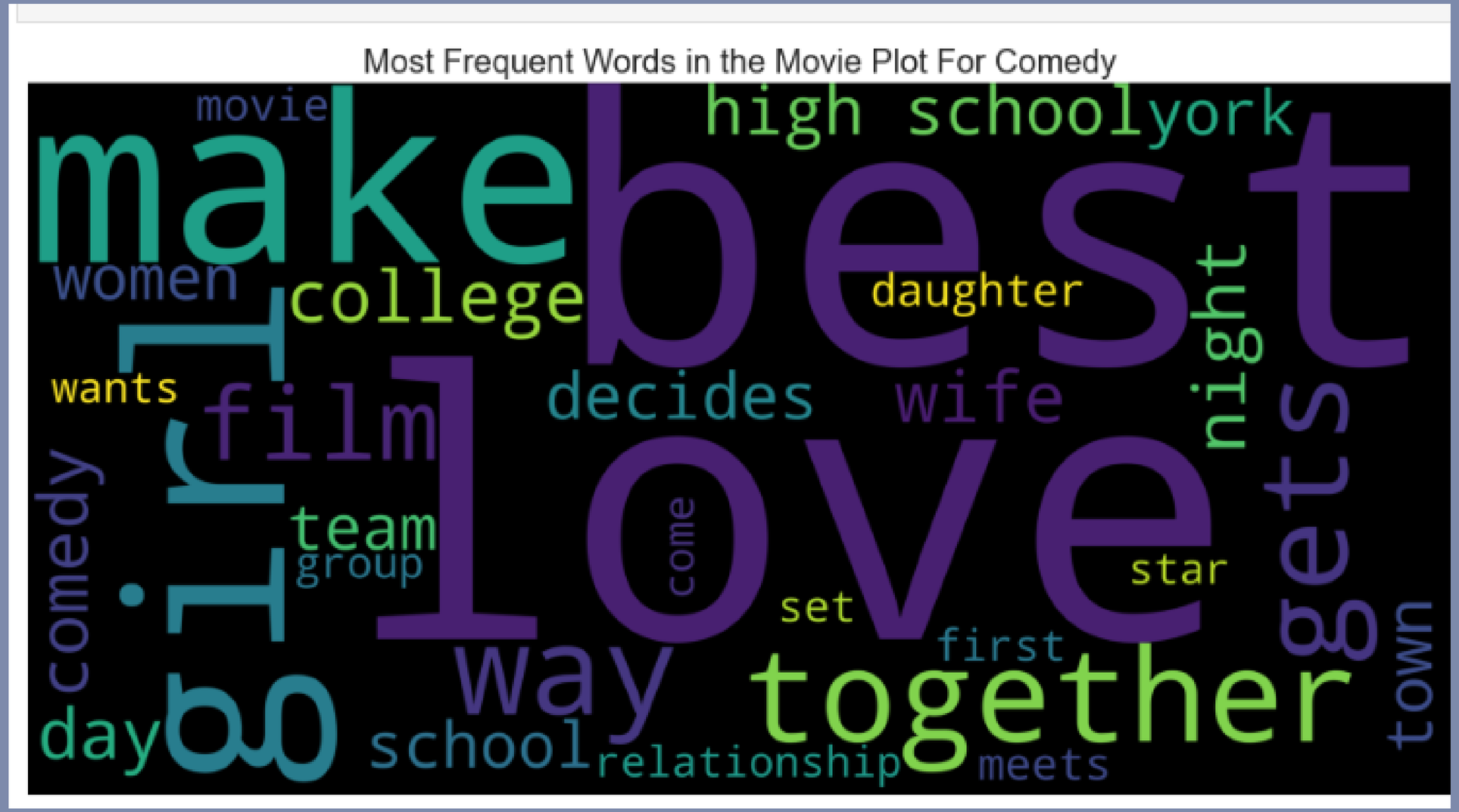
Preparation for machine learning

- Kept the top 5 genres where each genre have number of movies more than 5% of the total movies
- Dropped all columns except genre and overview
- Clean meaningless words in overview with STOPWORDS from nltk library
- Generate WordCloud for the most frequent 30 words for each genre

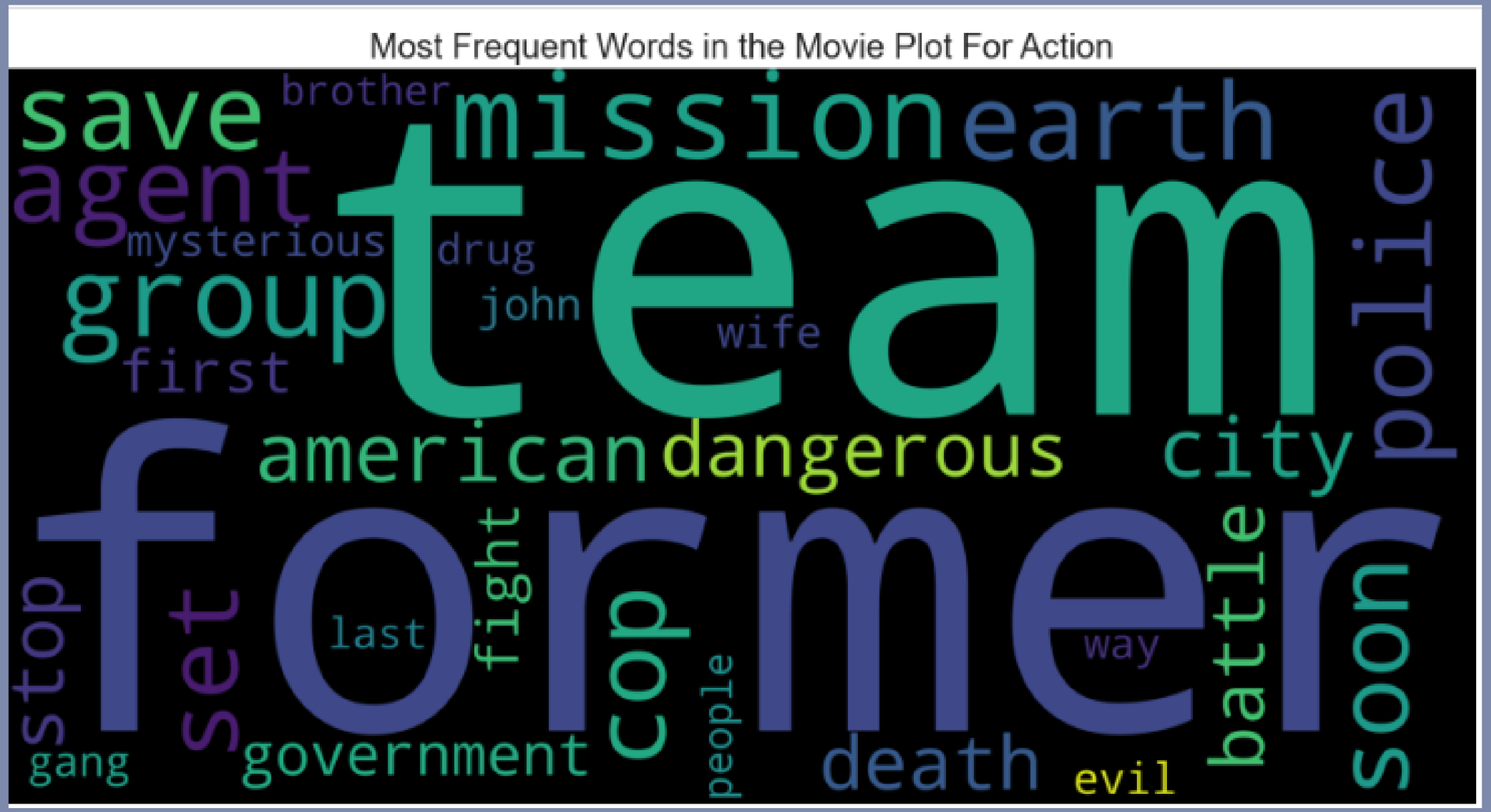
Drama


Most Frequent Words in the Movie Plot For Drama
town
city
together
left
film
son
way
true
women
school
american
begins
daughter
group
relationship
work
wife
love
day
girl
people
set
soon
british
first
death
boy
husband
former
make

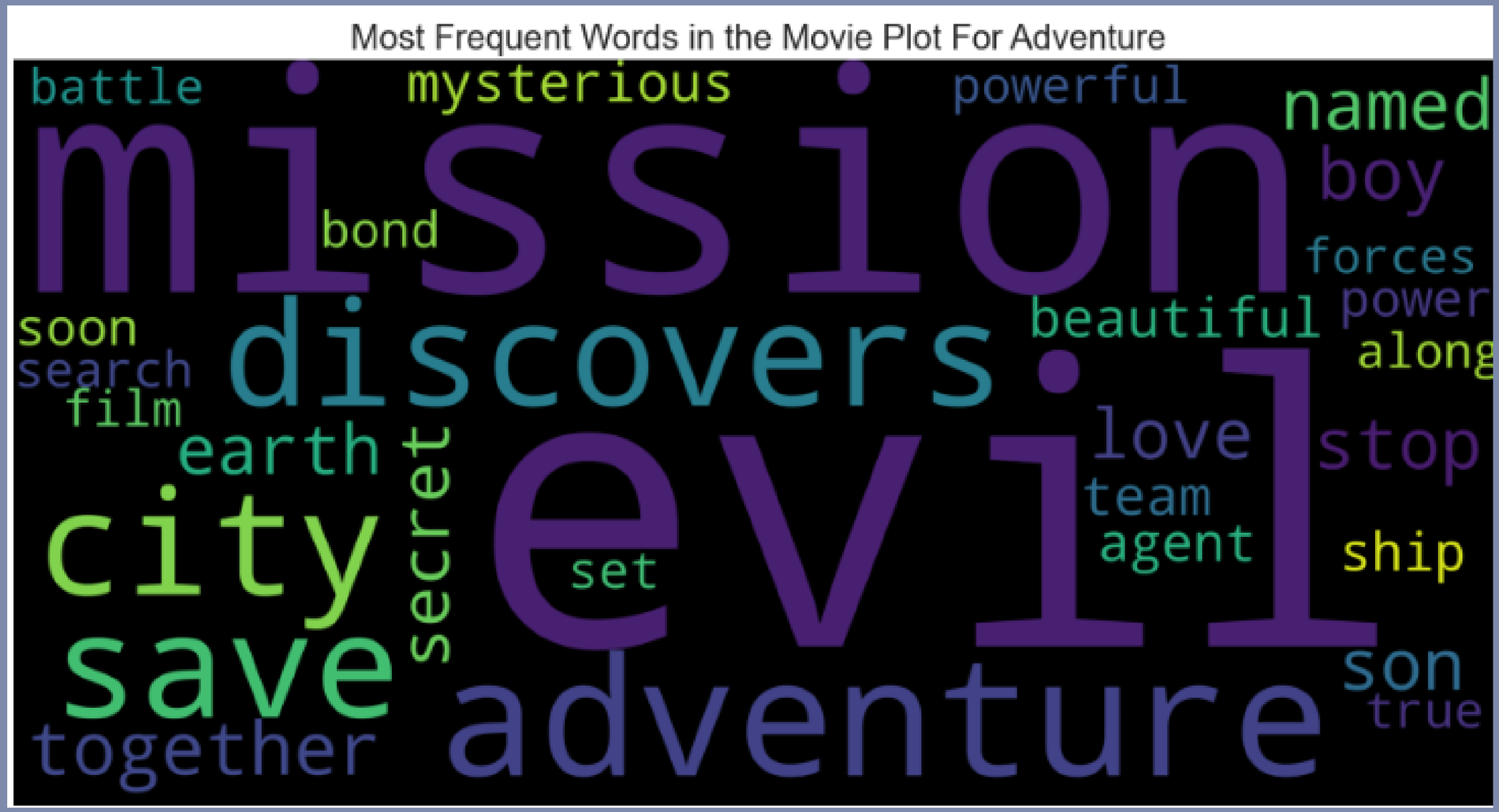
Comedy


Most Frequent Words in the Movie Plot For Comedy
movie
high schoolyork
make
best
love
girl
women
college
daughter
night
wants
film
decides
wife
gets
comedy
team
group
come
star
set
first
way
together
town
day
boy
school
relationship
meets

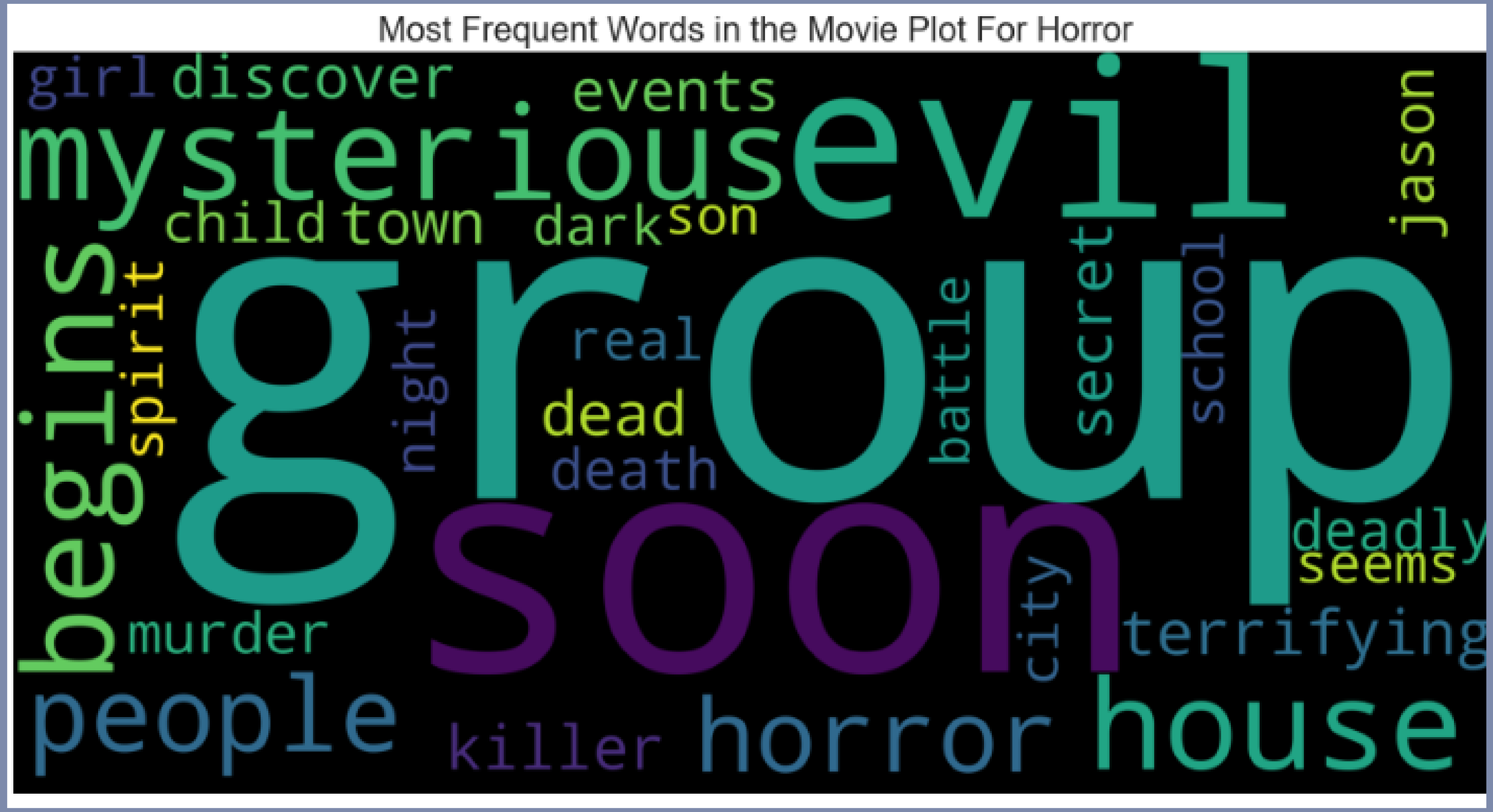
Action


Most Frequent Words in the Movie Plot For Action
save
brother
mission
earth
agent
team
police
mysterious
drug
group
john
wife
first
former
american
dangerous
city
stop
set
last
fight
cop
battle
soon
way
people
gang
government
death
evil

Adventure


Most Frequent Words in the Movie Plot For Adventure
battle
mysterious
powerful
named
mission
boy
bond
forces
power
soon
beautiful
discovers
search
film
along
evil
earth
love
stop
secret
team
city
agent
ship
set
save
son
adventure
true
together

Horror

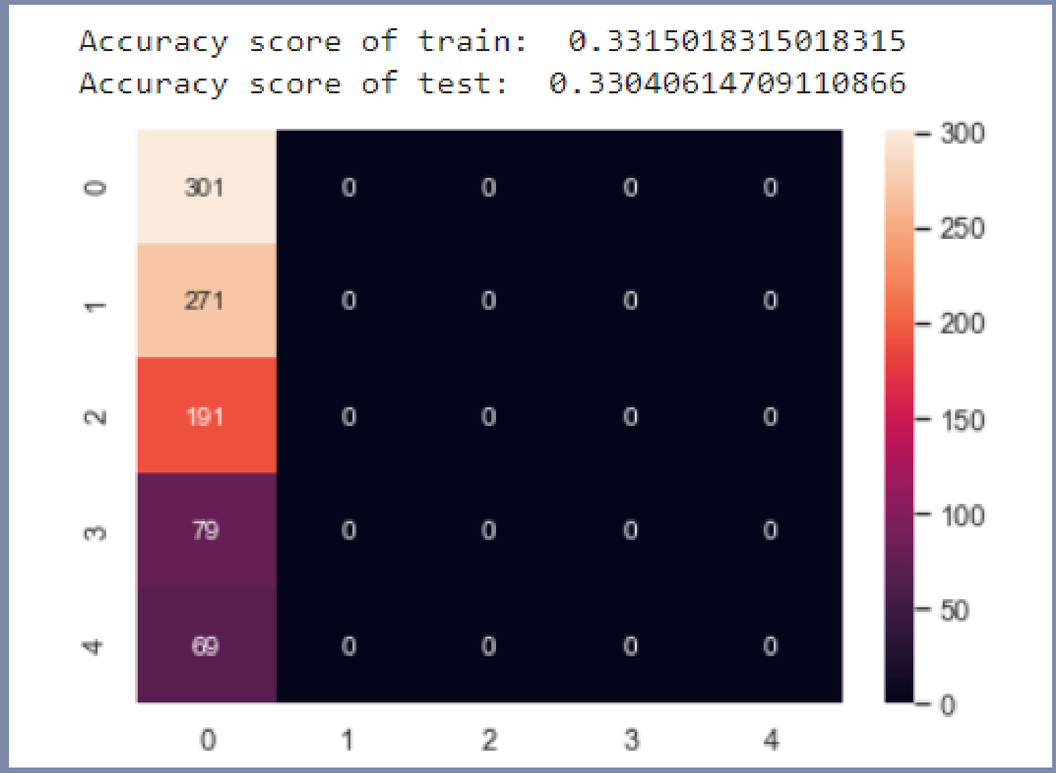

Most Frequent Words in the Movie Plot For Horror
girl
discover
events
mysterious
evil
child
town
dark
son
jason
begins
spirit
group
night
real
dead
death
battle
secret
school
soon
deadly
seems
city
murder
terrifying
people
killer
horror
house

Machine Learning

- Dummy classifier, Gaussian Naive Bayes, Logistic Regression
- Helps us to predict categorical data , allowing us to predict the genre of movies

Dummy Classifier

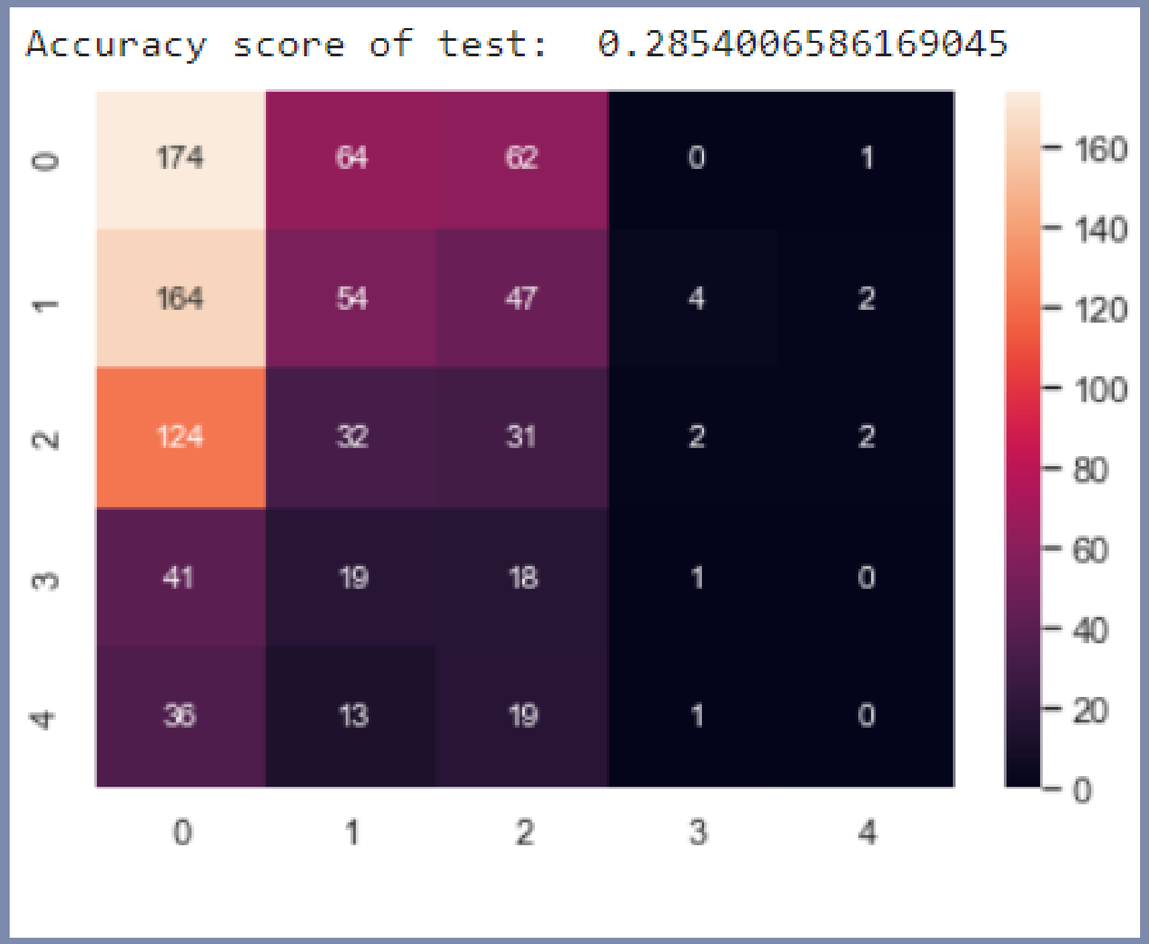
- Makes predictions without finding the trend of the data
- Predicts most frequent class in the dataset
- Accuracy score for test: **33.04%**



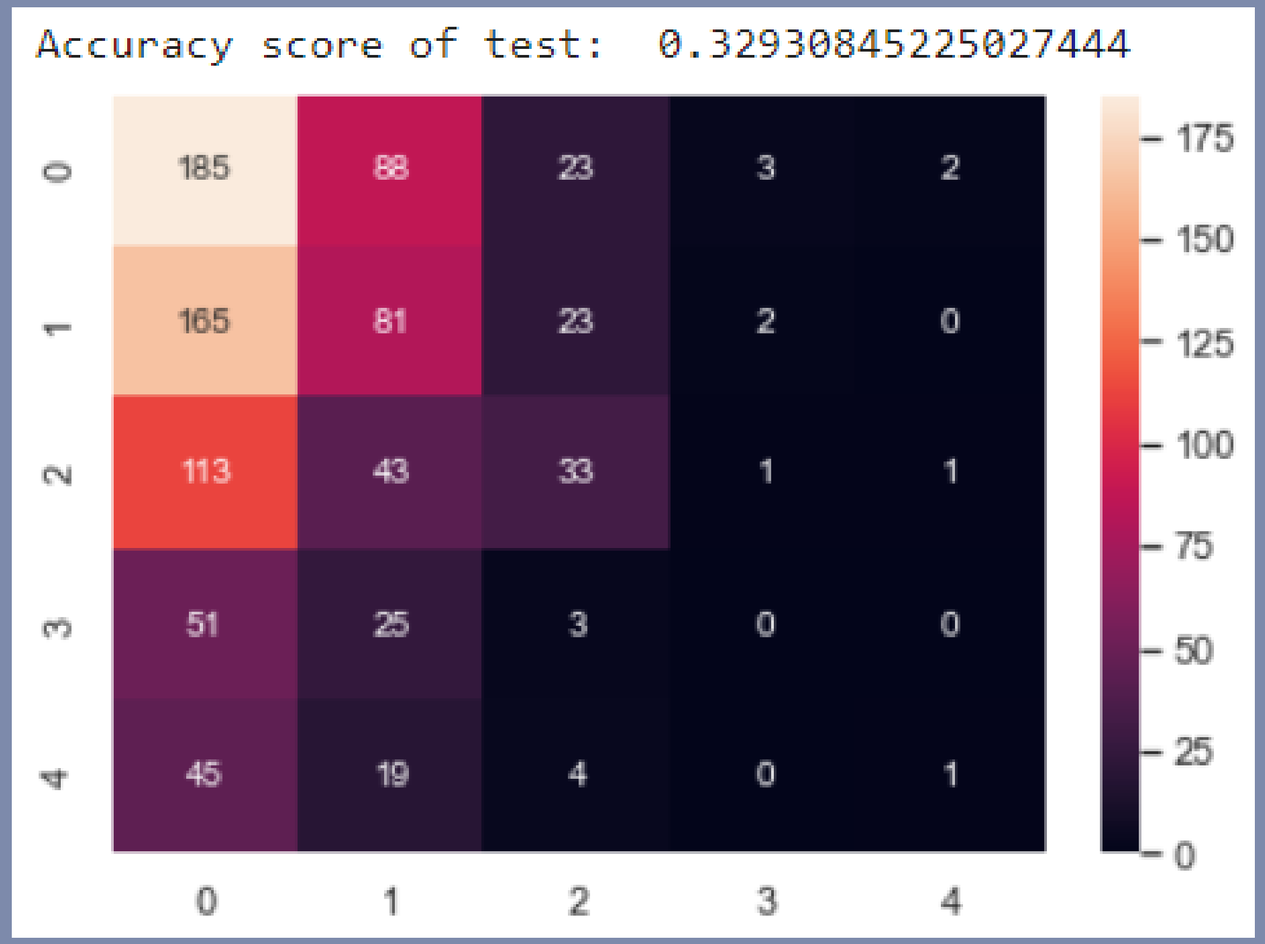
Gaussian Naive Bayes and Logistic Regression

- Change the the bag of words into an array using **CountVectorizer** from sklearn
- The overview was first converted into array before the machine learning process
- The train set is then fit to the model
- Accuracy score for Naive Bayes & Logistic Regression: **28.54%** & **32.93%**

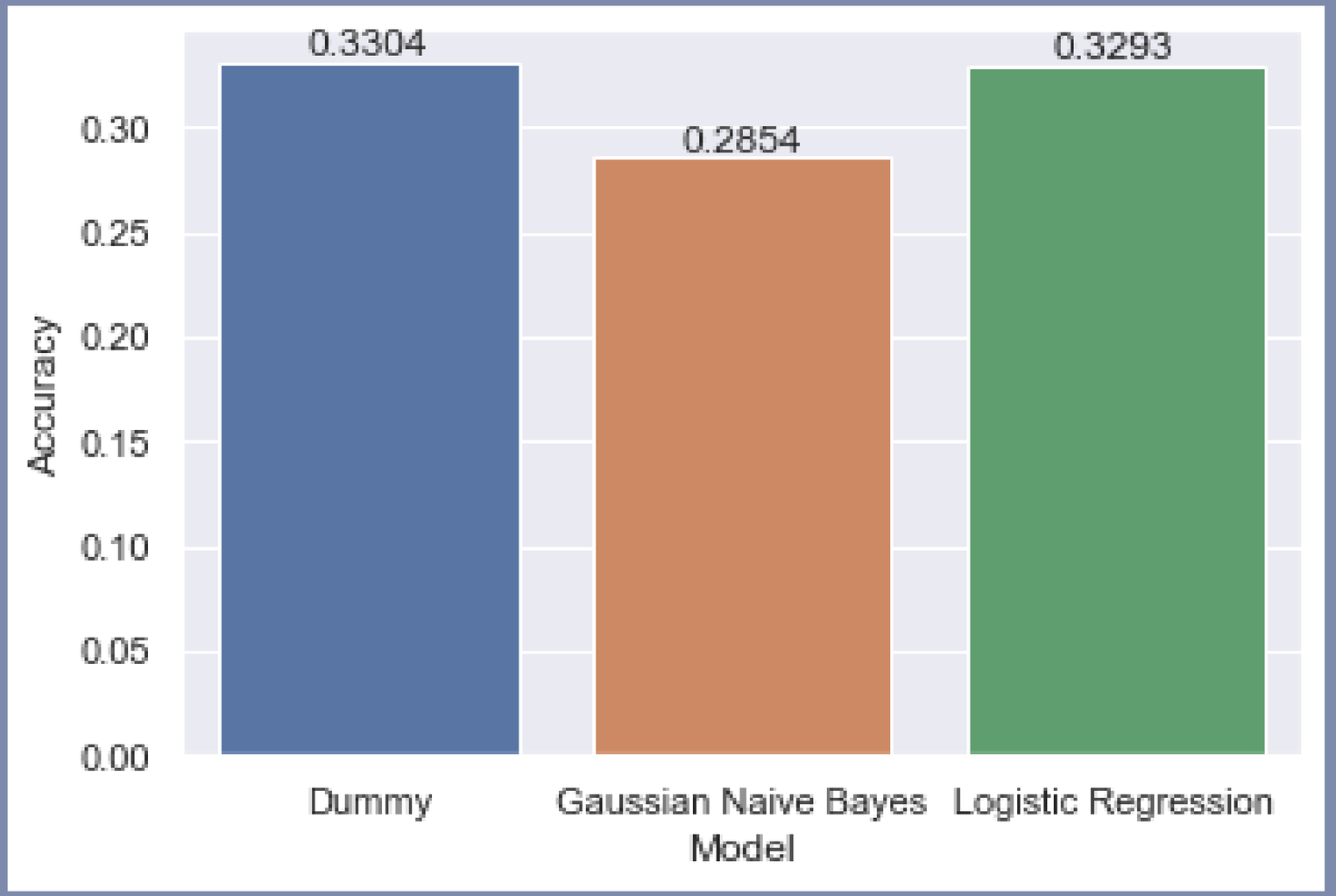
Gaussian Naive Bayes



Logistic Regression



Accuracy Plot of the Three Model



Conclusion

- Accuracy is low for machine learning (33.04%)
- Dummy Classifier performed the best followed by Logistic Regression and finally Gaussian Naive Bayes
- False classification occurred on genres which had most words in common
- Recommend using a dataset of a larger scale to improve the accuracy of machine learning
- Although the accuracy is low, it is still higher than the probability of randomly classifying a movie into one of the genre (20%).

What we have learnt

- Learnt how to generate wordclouds
- Learnt the use of NTLK library and how to handle unclassified data
- Learnt 3 new machine learning techniques to predict categorical data

Thank you !