INTRODUCTION TO TEXT MINING

TERM PROJECT

Fahrettin Çetin
Bilge Urçuk

A.1.

These are the preprocesses methods applied in the project :

- Removing new line character
- Removing punctuation, numbers, double white spaces
- Stopword removal
- Tokenization
- Reduce lengthening
- Dictionary check
- With spacy library doing lemmatization

First six preprocesses are predefined methods from libraries. These applications must be done for clean data shortly. Furthermore, the significant improvement about cleaning doing lemmatization with scapy. Instead, TextBlob.correct() can be chosen as an alternative but this method is not a good choice because of the running time. It is too long to run. The most important advantage of scary is that it is a very fast algorithm.

Before lemmatization, stemming is used. The results were more meaningless and inefficient. Since stemming has a sharp cutting procedures.

B.1.

Number of words: This feature shows how long text is. By length of texts in a row, it can be estimated how meaningful the sentence is and how much information is given or whether it contains unnecessary sentences or not.

Polarity: Polarity refers to the attitude or sentiment expressed in a text. Polarity can be positive, negative, or neutral. In this project, it is used to understand whether movie reviews are positive or negative.

Polarity is often used as one of the features in sentiment analysis, which is the process of extracting subjective information from text. Sentiment analysis is often used to determine the overall sentiment of a piece of text or to identify specific opinions or emotions expressed in the text.

Subjectivity: In natural language processing (NLP), subjectivity refers to the degree to which a text is based on personal opinions, feelings, or beliefs. In this project, it is used to understand whether movie reviews are objective or subjective under is_emotional statement.

Language: It is used for check whether all text have same language category or not. If they are not, that means Translation should be made. Luckily all text in this project have same language category.

B.2.

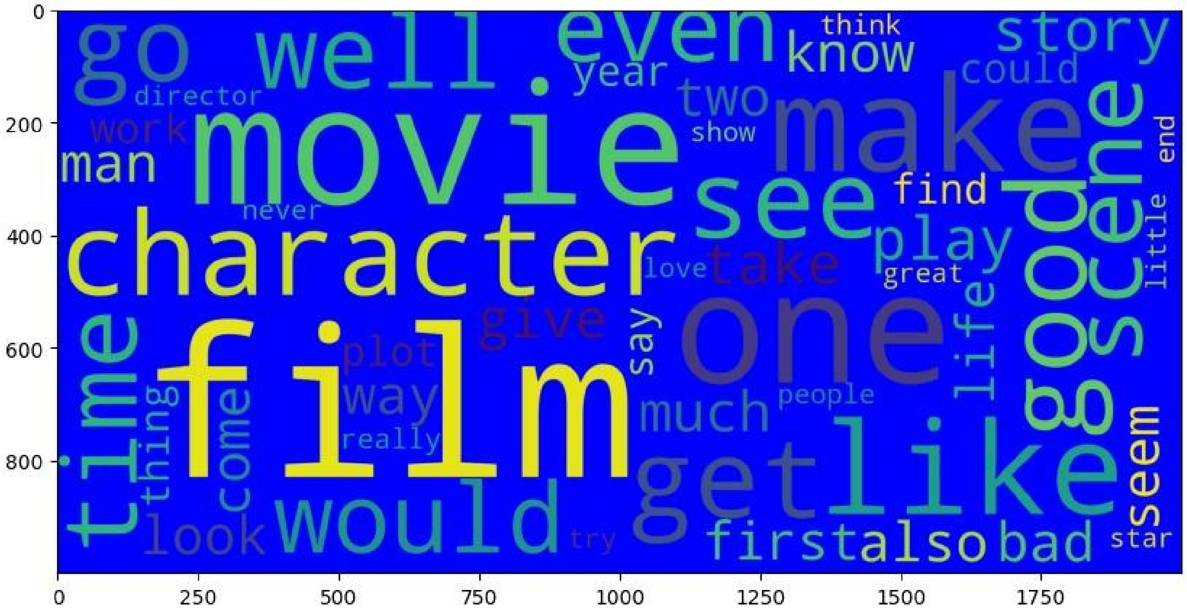
Bag of Words is a simple approach that represents a text as a bag of its words, ignoring grammar and word order. You can then represent each document as a numerical vector, with each element of the vector representing the count of a specific word in the vocabulary.

TF-IDF is a more sophisticated approach that takes into account the frequency of a word in a document as well as its frequency across all documents in a corpus. The goal of TF-IDF is to give more weight to important words that occur frequently in a specific document but not in many documents in the corpus.

Bag of words is chosen for Wordcloud and TF-IDF is used in machine learning algorithms because of its sophistication.

In the results of the most important 10 features, both algorithms have shown the same results. One situation where BOW and TF-IDF might produce similar results is when all the documents in a corpus have similar lengths and contain a similar mix of words. In this case, the term frequencies (TF) of the words in each document would be similar, and the inverse document frequencies (IDF) of the words would also be similar, leading to similar TF-IDF values.

B.3.



B.4.

K-means Clustering: Distinctively, this algorithm has 5 clusters. The most important five words are selected according to the TF-IDF and it is implemented in k-mean. The reason of that is to understand the most popular topics in text. These are movie, character, love, good, and scene. In this way, inferences can be made about which subject the comments about the movie.

Logistic regression: Logistic regression was used because it would provide higher accuracy for classification than the linear regression model.

Random forest: Random forest is an ensemble kind of machine learning model and that's why no overfitting problem is observed and it has higher accuracy than other tree models. In order to obtain higher accuracy, random forest was applied in all tree models.

SVC: Support Vector Machines are a type of machine learning algorithm that can be used for classification, regression, and outlier detection. In natural language processing (NLP), SVMs are often used for classification tasks, such as spam filtering, sentiment analysis, and topic classification.

Logistic Regression Score

```
 precision    recall  f1-score   support

         0       1.00      0.02      0.03        59
         1       0.85      1.00      0.92       341

   accuracy                          0.85       400
  macro avg 0.93      0.51      0.48       400
weighted avg 0.88      0.85      0.79       400
```

Random Forest Score

```
        precision    recall  f1-score   support

         0       0.85      0.25      0.38       157
         1       0.67      0.97      0.79       243

   accuracy                          0.69       400
  macro avg       0.76      0.61      0.59       400
weighted avg       0.74      0.69      0.63       400
```

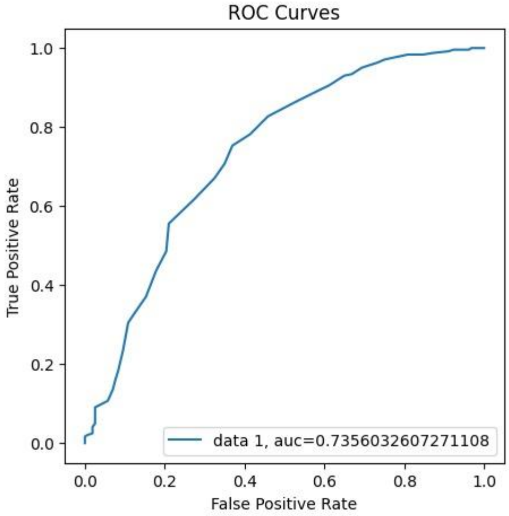
SVC Score

```
 precision    recall  f1-score   support

         0       0.88      0.24      0.37        59
         1       0.88      0.99      0.94       341

    accuracy                         0.88       400
macro avg      0.88      0.62      0.65     400
weighted avg  0.88      0.88      0.85     400
```

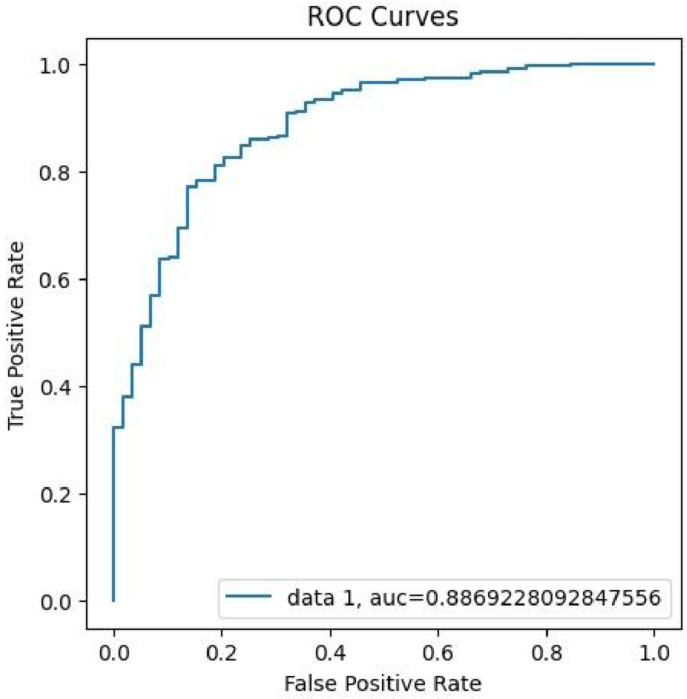
K-Means Score

```
  precision    recall  f1-score   support

         0       0.58      0.07      0.12       277
         1       0.04      0.20      0.07        20
         2       0.00      0.00      0.00         1
         3       0.06      0.13      0.08        55
         4       0.03      0.04      0.03        47

    accuracy                          0.08       400
   macro avg 0.14      0.09      0.06       400
weighted avg   0.41  0.08      0.10       400
```

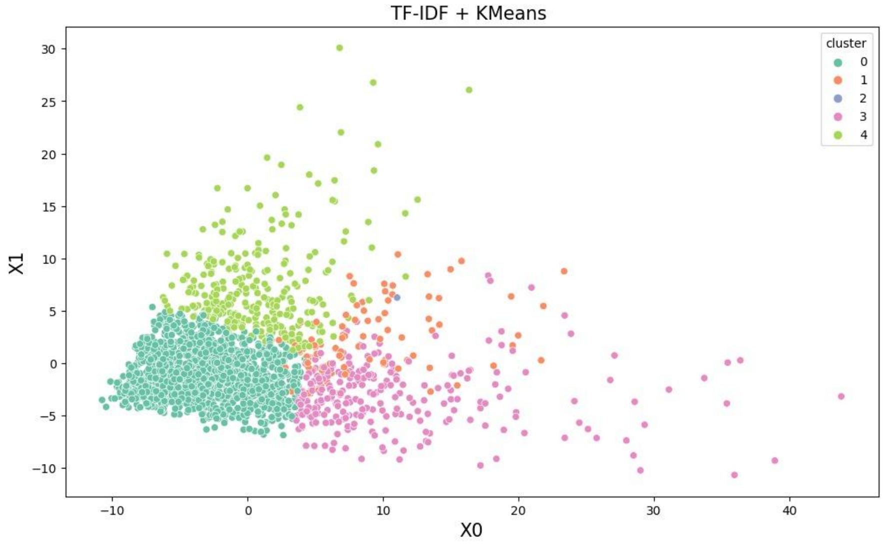
Random Forest Roc Curve



Logistic Regression Roc Curve



K-Means



C.

As mentioned before, According the this algorithm and TF-IDF scores most important topics are character, movie, love, good, and scene. According the elbow method the calculated k value can be chosen as 4 or 5. In this algorithm it is chosen as 5.
As it turns out, both algorithms have both a k-mean and an lda agreement each other

E.

The logic of the chosen algorithm aims to create a set of sentence vectors by counting how many words are in each sentence. From the combination of these vectors, how important which word is and its sentence structure can be understood. The general logic of the applied summarization method is as follows. However, due to the very long time and resource consumption of this algorithm, the sample summarization in the sent file was obtained from a small part of the text data, not all. Despite this short text amount, it contains positive and negative meaning, which should also contain two different summaries