Text Mining and NLP ON Star Wars Movie Scripts

Introduction

In analysis we are going to perform a statistical text analysis and sentiment analysis on the Star Wars scripts from The Original Trilogy Episodes (IV, V and VI), using word clouds to show the most frequent words

In this data analysis project I will be analyzing the dialogues of few characters from star wars episode 4, 5 and 6. Each of the dataset is in form of text file and has 2 columns the character name and the dialogue spoke by that character. To start things off I will first load the data from the text files into 3 different datasets.

Task#01

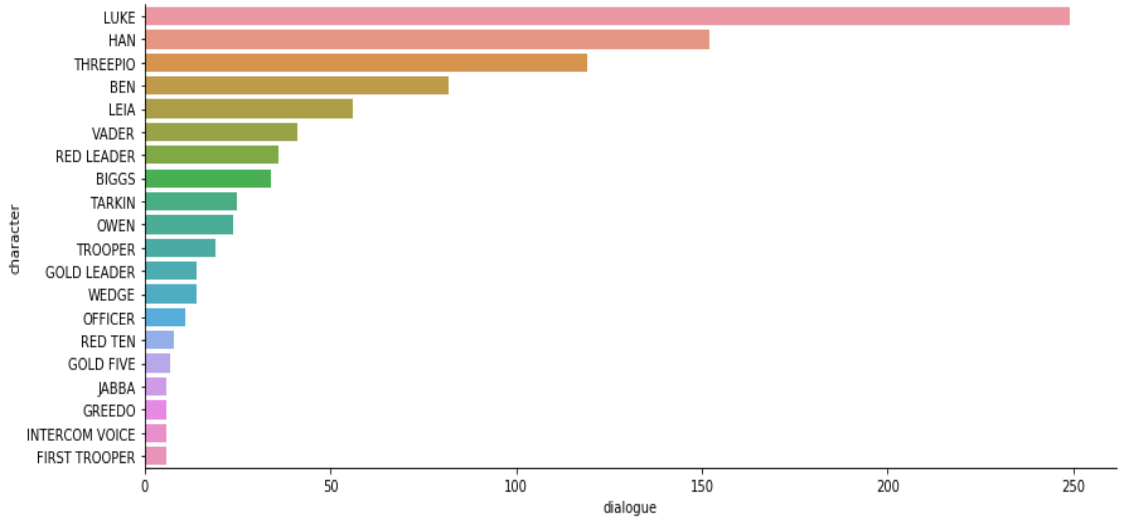
In this task we Query the characters with most dialogue in each of The Original Trilogy (episodes IV, V, VI).

- In Episode 4 “LUKE” have dialogue 249
- In Episode 5 “HAN” have dialogue 186
- In Episode 6 “HAN” has dialogue 124

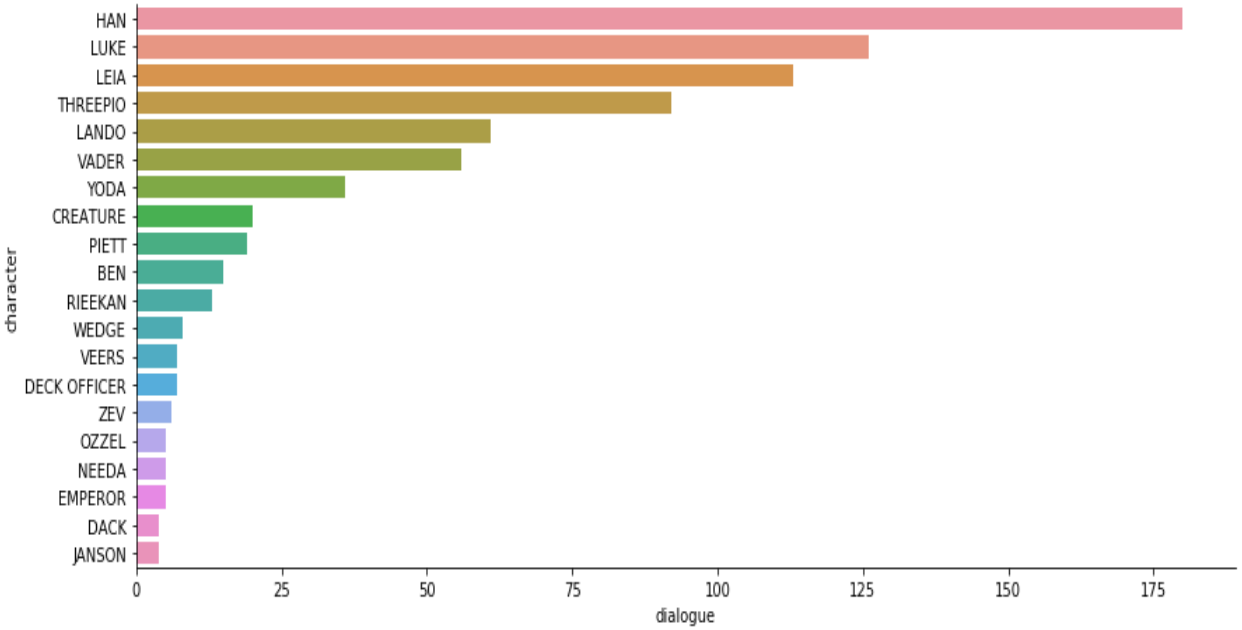
We analyze that in episode 4 have characters 60, episode 5 have characters 49 and episode 6 have characters 53 which is Visualize in Task#2

Task#2

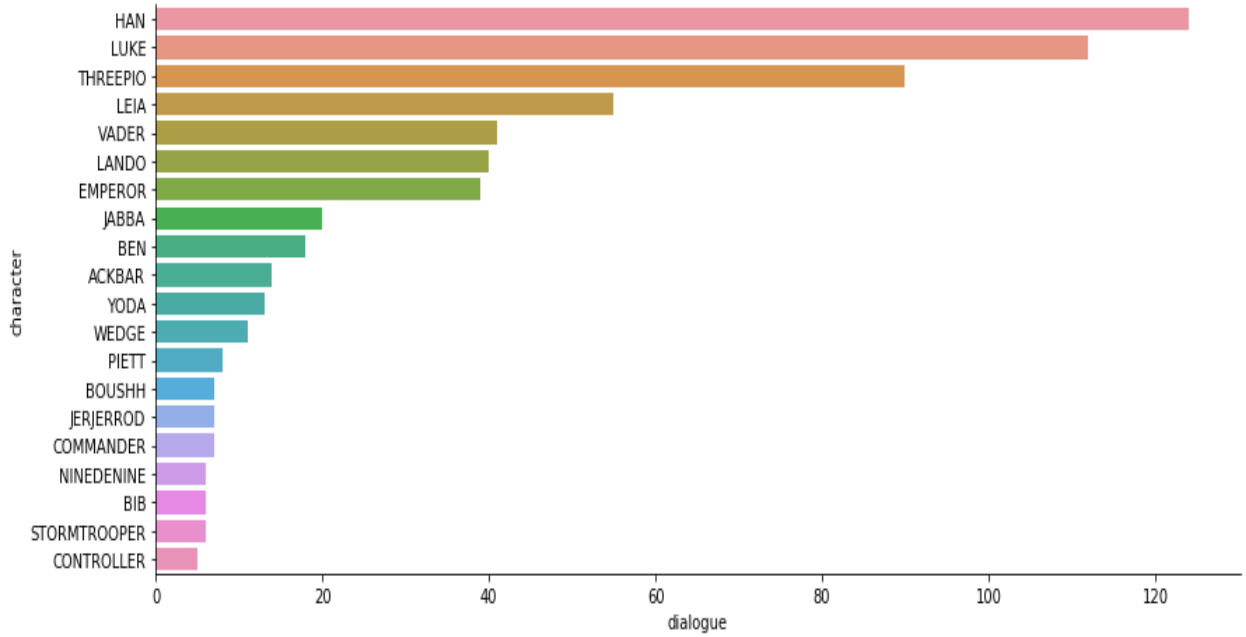
- Top 20 characters with most dialogue of episode 4.



- Top 20 characters with most dialogue of episode 5



- Top 20 characters with most dialogue of episode 6



Task#3

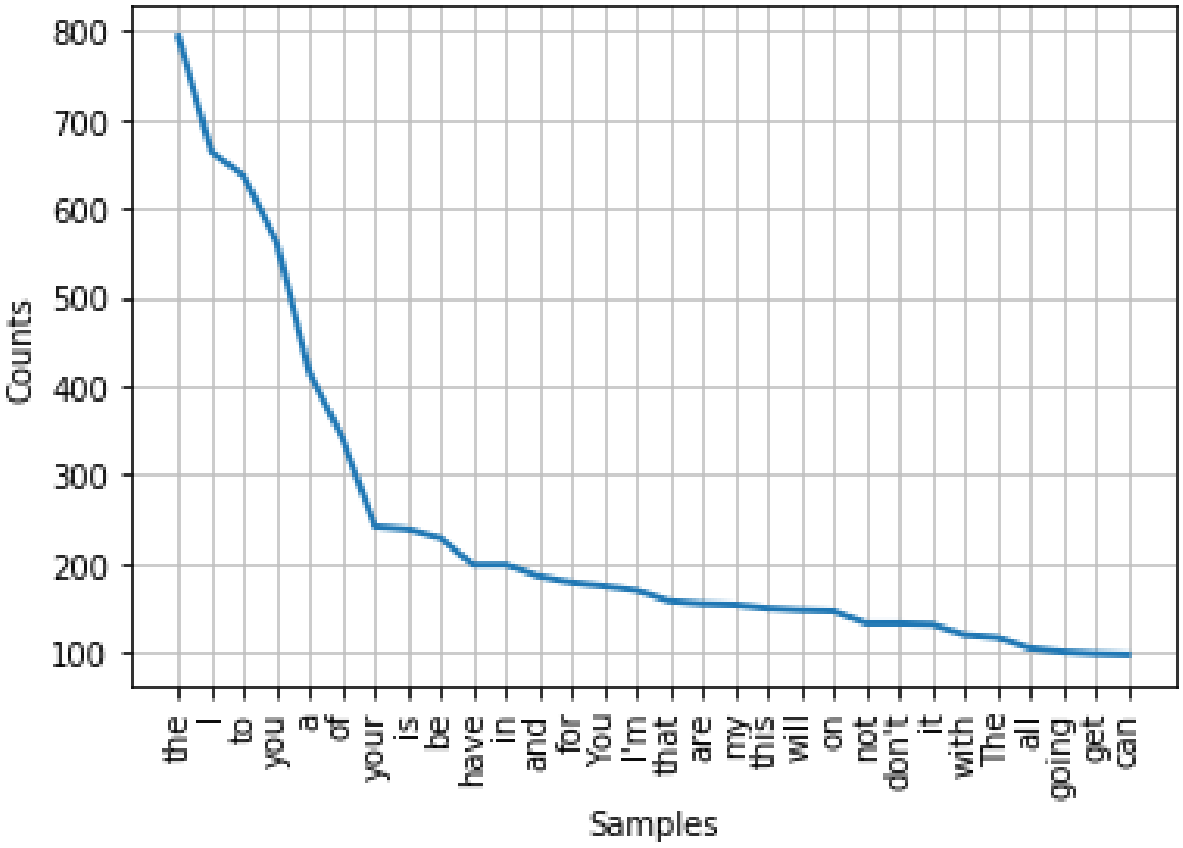
- In this task we add new column “episode” to the three datasets (to distinguish between the three episodes) and concatenate them into one dataset.

Task#04

- By discovering the Frequency Distribution of words in The Original Trilogy without any preprocessing.
- we analyze that the Frequency of stop word is high

Task#05

- **we can see in the following plot the Frequency of stop word is high**



Task#06

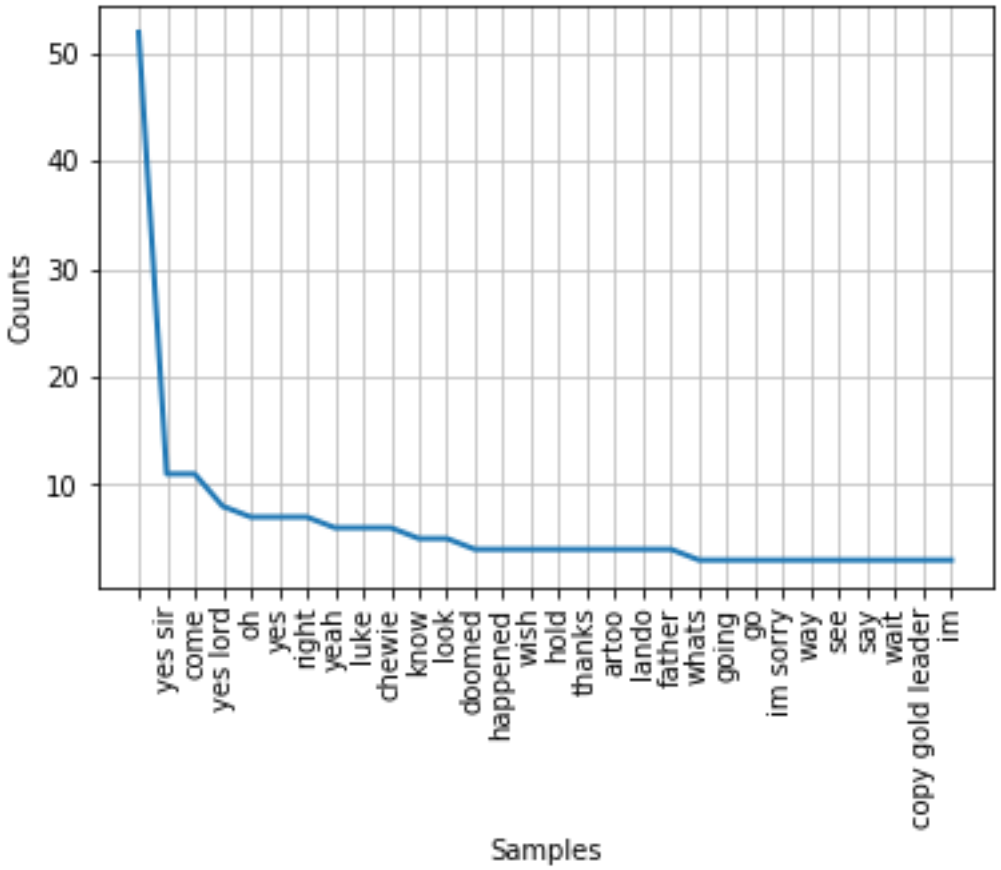
We Appling text-mining operations to prepare our dataset for further text analysis.

We apply the following data cleaning and preprocessing technique

- Remove all punctuation marks
- Converting all characters lowercase
- Remove some common English stop words
- lemmatization

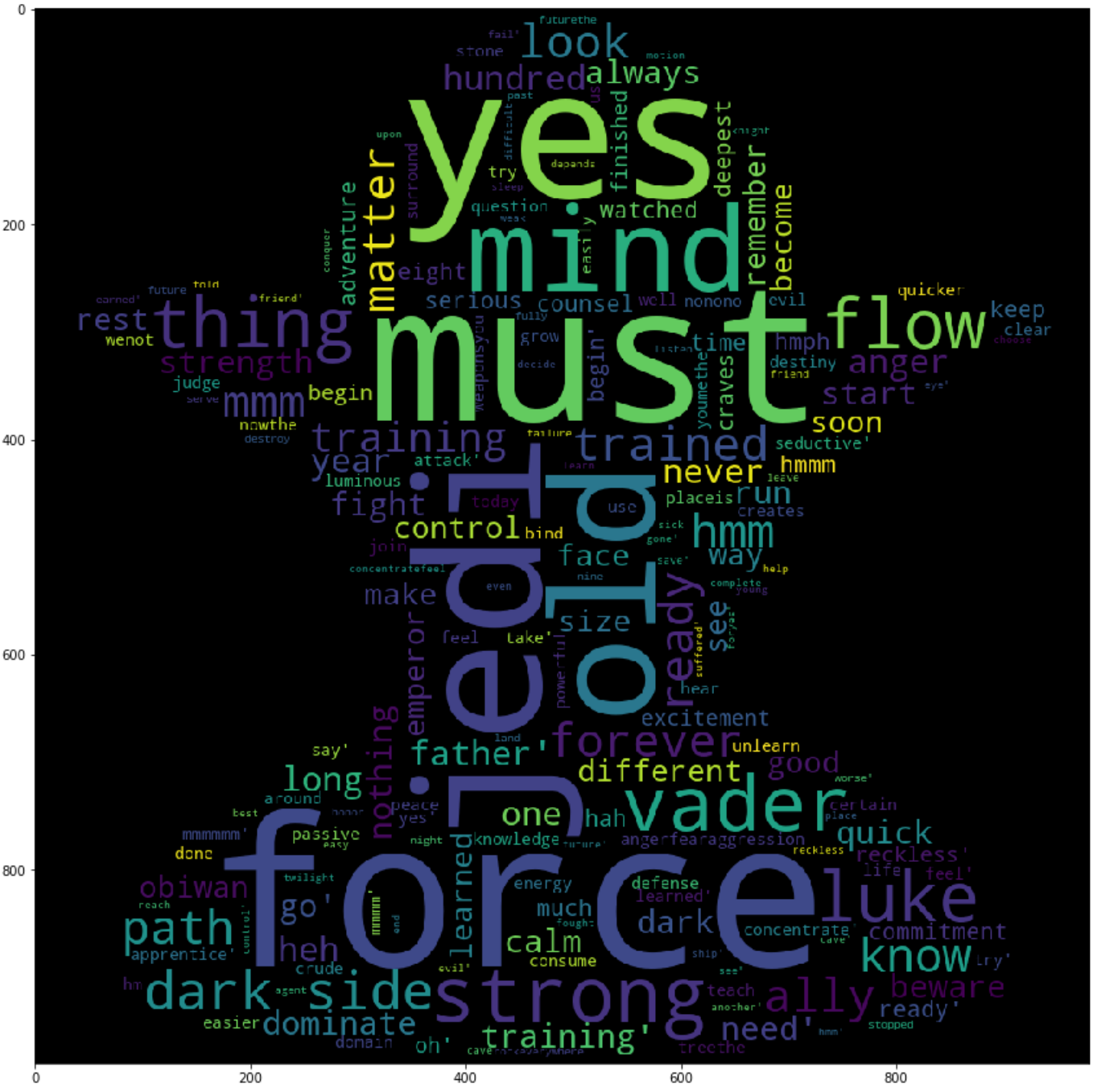
Task#07

By discovering the Frequency Distribution of words we show following plot of the Frequency Distribution in The Original Trilogy after the data cleaning and preprocessing

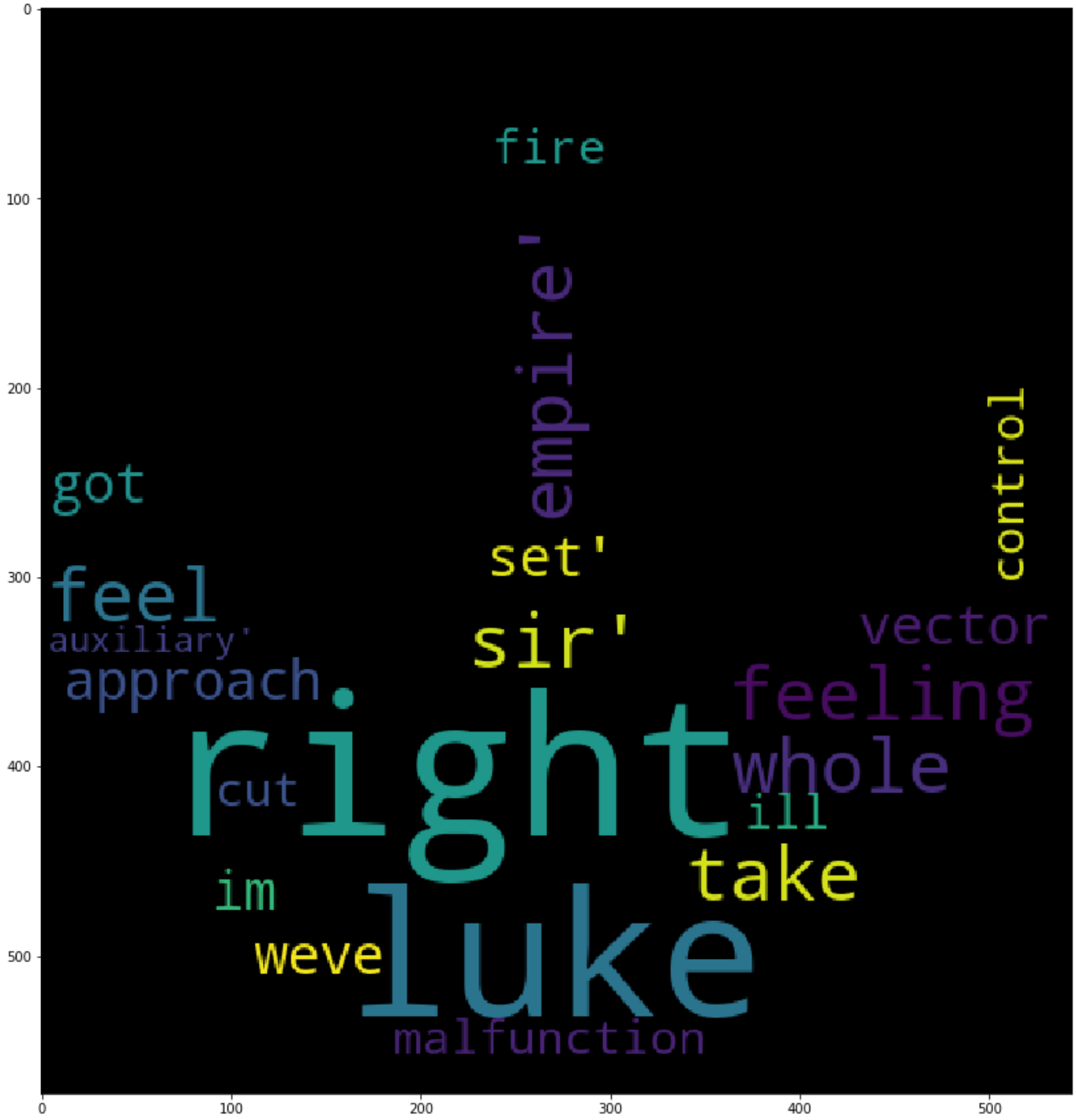


Task#08

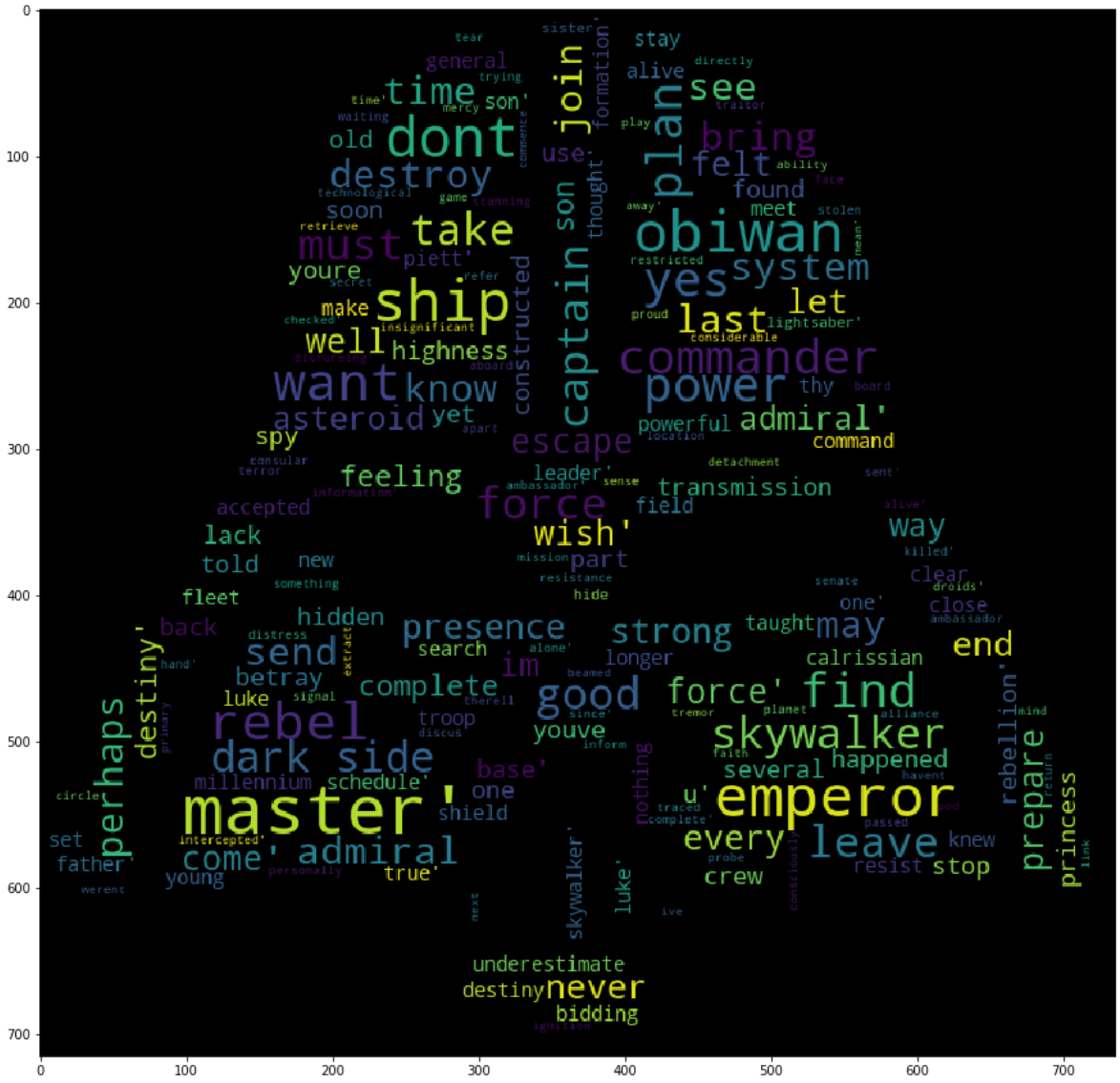
- **Visually represent the most repeated words of Yoda character using word cloud**



- **Visually represent the most repeated words of Dack character using word cloud**



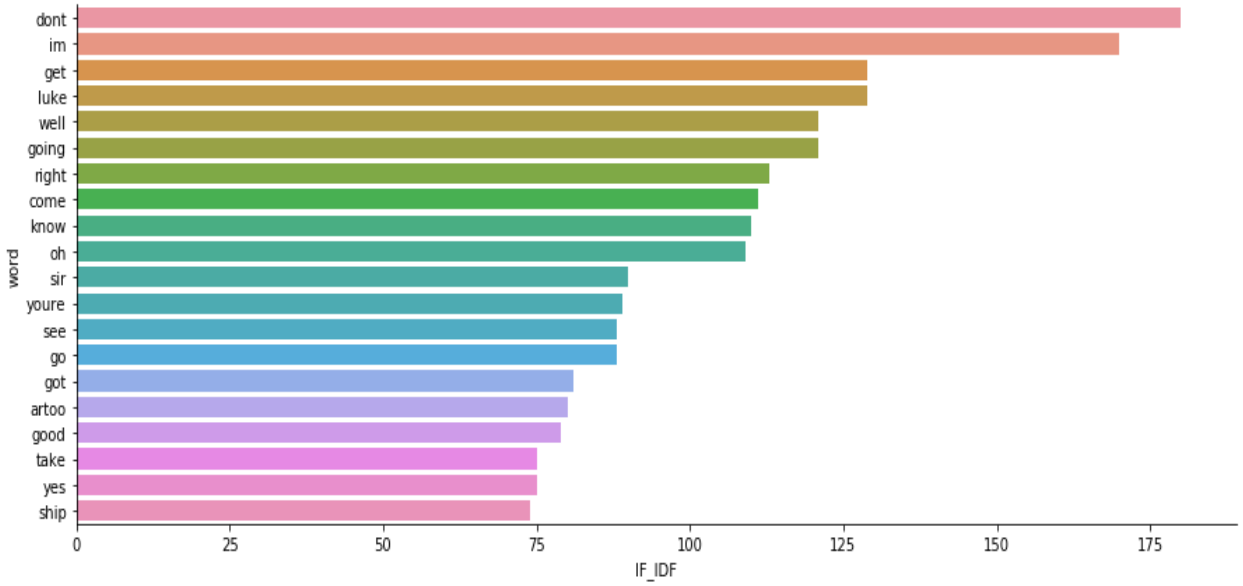
- **Visually represent the most repeated words of Varder character using word cloud**



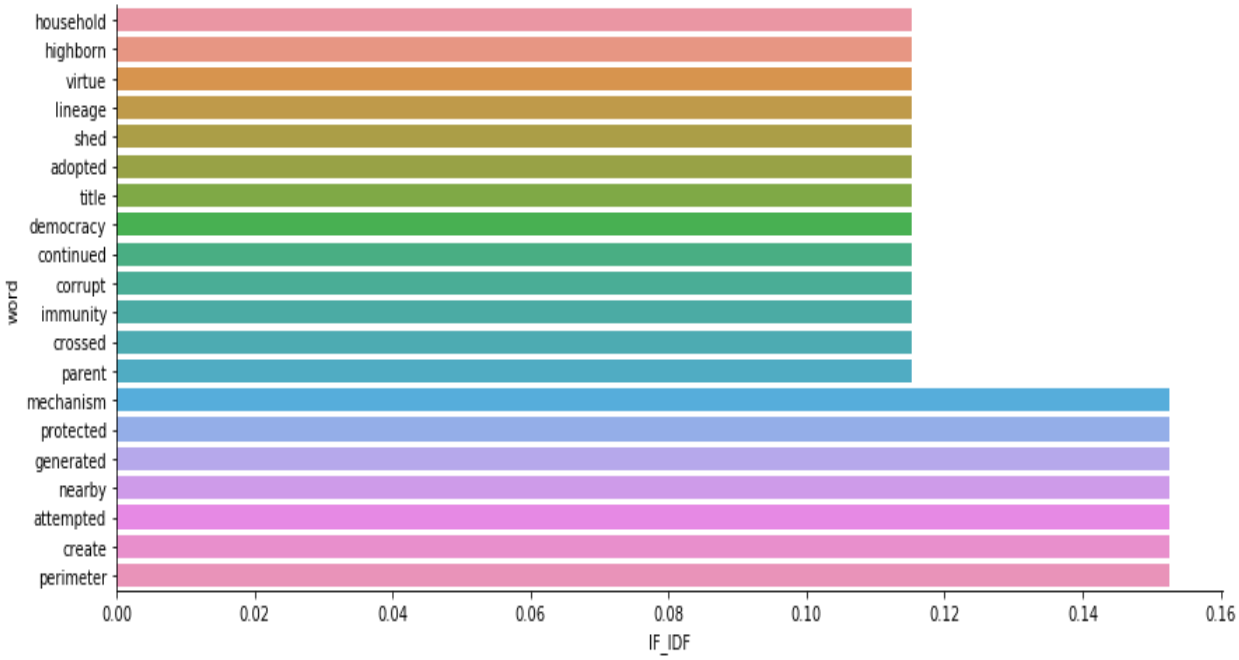
Task#09

- We discover the most repeated words and the most relevant words
- We discover most common words using Bag of Words/count vector which just creates a set of vectors containing the count of word occurrences in the document

- We can visualize most repeated word



- We discover most relevant words using TF-IDF model which contains information on the more important words and the less important ones as well
- We can visualize most relevant word



Task#10

Sentiment analysis

Sentiment analysis is refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

We performing a sentiment analysis by using TextBlob library in python we categorizes dialogue in a binary fashion into positive and negative categories.

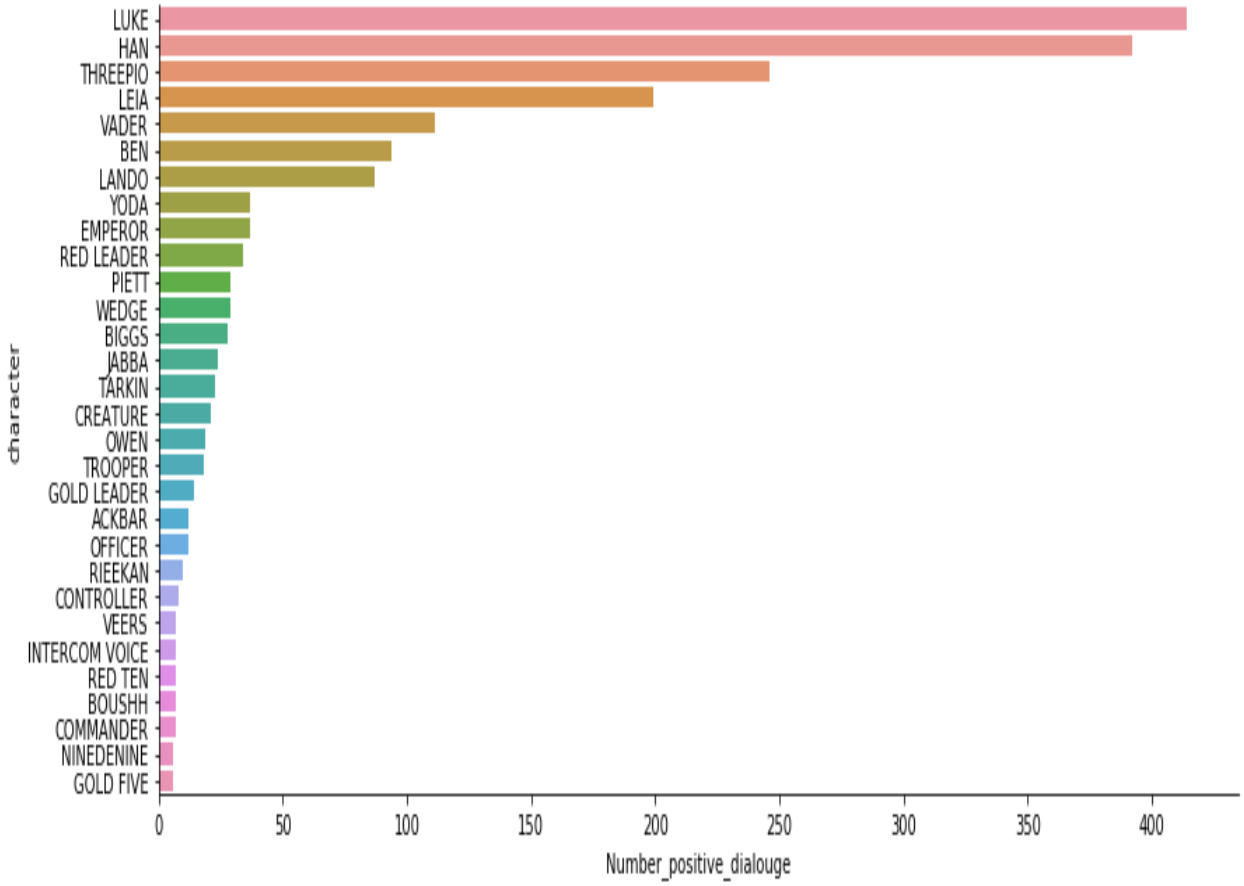
Results of sentiment analysis

After performing a sentiment analysis we categorizes dialogue in a binary fashion into positive and negative categories.

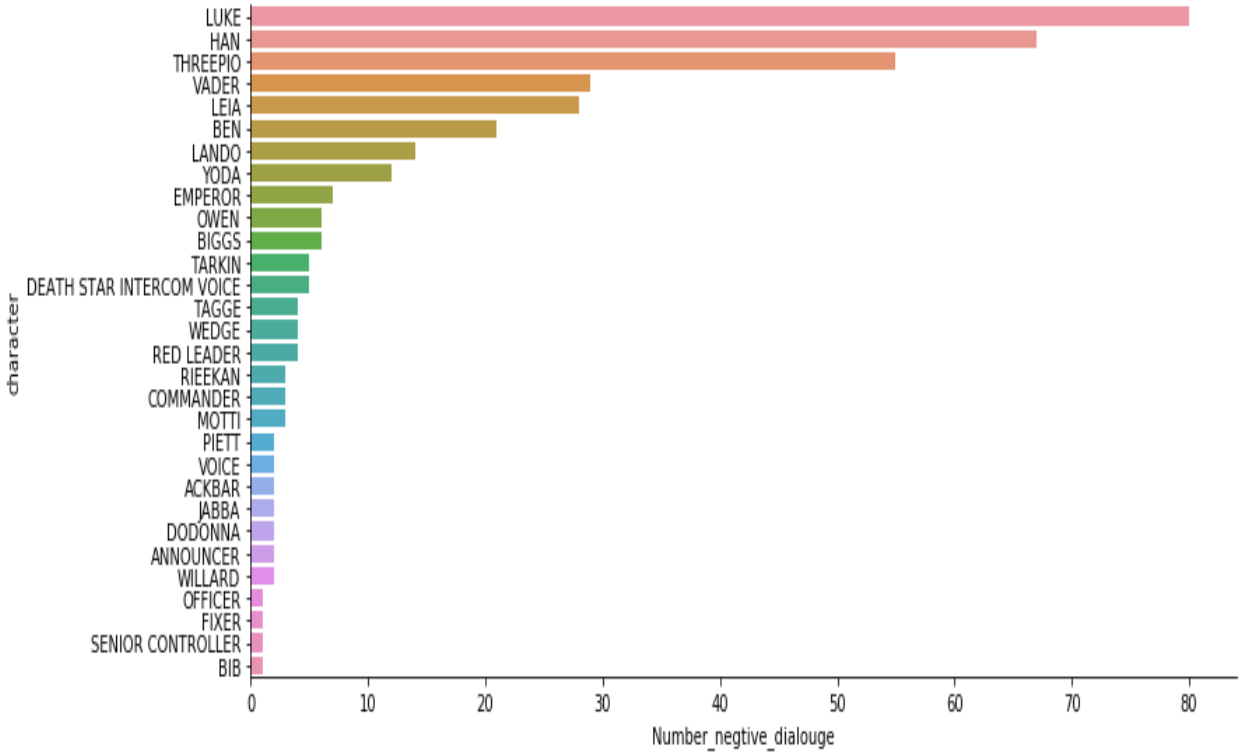
We analyze that number of positive and negative dialogue as following

Positive	**2126**
Negative	397

Analyze the character with positive dialogue



Analyze the character with negative dialogue



Analyze the episode with negative and positive dialogue

- EPISODE 4 has negative and positive dialogue respectively 174 and 836
- EPISODE 5 has negative and positive dialogue respectively 137and 702
- EPISODE 6 has negative and positive dialogue respectively 86and 588