2023

Report - Lab 4

TEXT MINING AND NATURAL LANGUAGE PROCESSING

TWAN GERJO VELDHUIS, H22TWAVE@DU.SE

The Star Wars franchise has been a cultural phenomenon for over four decades, captivating audiences with its epic storyline, iconic characters, and memorable dialogues. In this text mining and natural language processing exercise, we will analyze the script dialogues of the first three movies (Episodes IV, V, and VI) of the Star Wars series, also known as The Original Trilogy. By analyzing the frequency distribution of words and performing text-mining operations using the Natural Language Toolkit (NLTK), we aim to uncover insights about the trilogy's most used words, relevant words, and overall sentiment. We will also create word clouds to visualize the most used words for two of the most popular characters, Darth Vader and Yoda.

As a first step towards analyzing the Star Wars movie scripts, we imported the dialogues into a data frame and added an extra column to identify the episode in which each character spoke their respective lines. To achieve this, we compiled a list of the character names and counted the number of times each character's name was used in each episode/movie.

We then plotted the character names against their corresponding counts to create a bar chart visualization, which we refer to as "Figure 1". This bar chart helps us to identify the most prominent characters in each episode/movie, as well as the frequency of their appearances in the script dialogues.

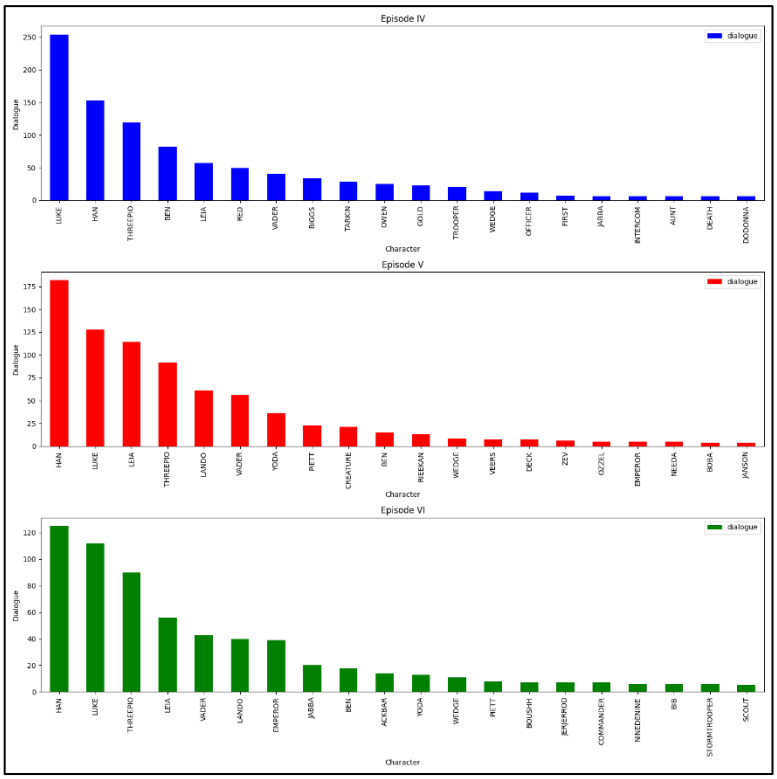

Figure 1 - character dialogue count per episode

After the initial analysis, we took a deeper dive into the word usage of the dialogues, these have first been executed without the use of any cleaning. This resulted in a word count that was done without any cleaning (only tokenization and lower casing). It resulted in a word frequency distribution that didn't contain a lot of words.

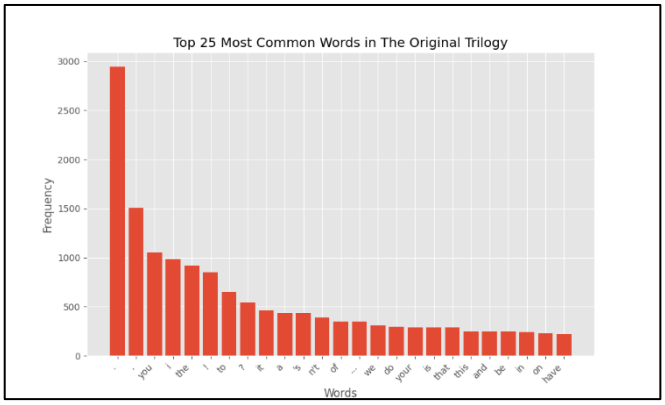

Figure 2 - top 25 words in The Original Trilogy before data cleaning

To prepare the Star Wars movie scripts for further analysis, we conducted a series of data-cleaning steps. For each dialogue, we followed a five-step process that involved:

1. Lower-casing: We converted all text to lower-case to ensure consistency and simplify future text analysis operations.
2. Removing punctuations, special characters, and numbers: We removed all non-alphabetic characters such as punctuations, special characters, and numbers from the text. This step is important as it allows us to focus on meaningful words and phrases rather than irrelevant symbols.
3. Word tokenization: We then split each dialogue into individual words to create a list of words, which is known as word tokenization.
4. Stop word removal: We removed stop words, which are commonly used words in the English language such as "the," "and," and "is." Removing these words helps to reduce noise and improve the accuracy of our analysis.
5. Word lemmatization: Finally, we performed word lemmatization to reduce each word to its base form, or lemma. This step helps to further simplify the data and make it easier to analyze.

After completing these steps for each dialogue, we added the cleaned sentences to a new column in the data frame called "new_script." By cleaning the data in this manner, we were able to standardize the text and prepare it for further analysis, such as creating a frequency distribution plot and using the TF-IDF model.

With the data now cleaned and prepared, we proceeded to create a word frequency distribution of the text using the newly added "new_script" column. This time, the frequency distribution would be more meaningful and provide greater insight into the word usage in The Original Trilogy.

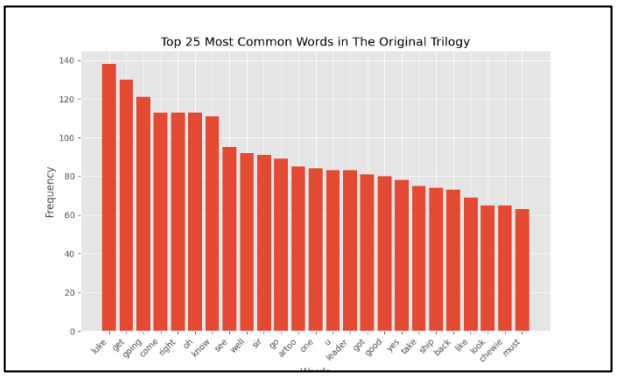

Figure 3 - top 25 words in The Original Trilogy after data cleaning

To create the frequency distribution, we used the Python Natural Language Toolkit (NLTK) library. NLTK provides a number of useful tools for text analysis, including the ability to count the occurrence of words in a text and create a frequency distribution. We first tokenized the words in the "new_script" column using NLTK's word_tokenize() function. This allowed us to break down each dialogue into individual words and create a list of words for each dialogue. Next, we created a frequency distribution using NLTK's Counter() function. This function counted the occurrence of each word in the list of words and created a dictionary with the word as the key and its count as the value.

By analyzing the frequency distribution, we were able to identify the most commonly used words in The Original Trilogy. These words can provide insight into the themes and motifs that appear throughout the movies, as well as the language and vocabulary used by the characters.

After analyzing the frequency distribution of words in The Original Trilogy, we proceeded to create word clouds for two iconic characters in the Star Wars universe: Darth Vader and Yoda. The purpose of these word clouds was to visualize the most frequently used words by each character in a more creative and engaging way.

To create the word clouds, we first extracted the dialogues spoken by Darth Vader and Yoda respectively from the cleaned script. Then, we used a Python library called "wordcloud" to generate the word clouds in the shape of the characters' respective images.

The word clouds represented the most commonly used words by each character, with the size of each word representing its frequency in the script. By examining these word clouds, we were able to gain a deeper understanding of the language used by these characters and their personalities.

For instance, the word cloud for Darth Vader contained words like "master," "force," "rebel," "emporer," and "obi," which are consistent with his role as a Sith Lord and his allegiance to the dark side of the Force. On the other hand, the word cloud for Yoda featured words like "Jedi," "mind," and "flow," which align with his role as a wise and powerful Jedi Master.

Overall, the word clouds provided a unique and visually appealing way to analyze the language used by these iconic Star Wars characters and gain insight into their personalities and motivations.

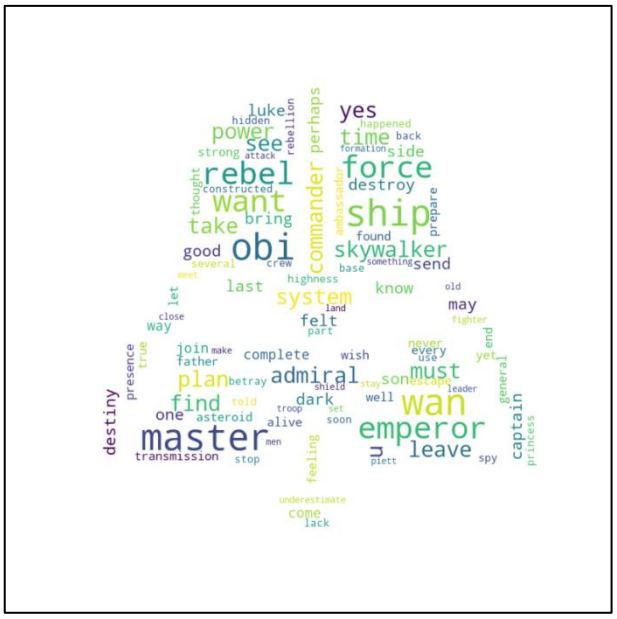

Figure 4 - darth vader wordcloud with mask

Figure 5 - yoda wordcloud with mask

The TF-IDF model was utilized on the pre-processed script of The Original Trilogy to pinpoint the most significant words in the documents. This model rates words based on their frequency in a document, inversely proportional to their frequency in the corpus of documents. The top 20 most relevant words in the trilogy, according to this model, include "luke," "right," "come," "going," "oh," "know," "yes," "leader," "artoo," "sir," "chewie," "got," "good," "let," "look," "like," "think," "ship," "help," and "father."

It's surprising that the character of Darth Vader, one of the most iconic characters in the Star Wars universe, doesn't appear in the top 20 most important words list. Instead, the list is dominated by common words like "luke," "right," "come," "going," "oh," "know," and "yes." These findings imply that the trilogy's primary focus is on the actions and conversations of the protagonists, rather than on

the antagonists like Darth Vader. Moreover, words like "leader," "artoo," "sir," and "chewie" are also highly rated, indicating the supporting characters' importance in the narrative.

In the final stage of our project, we conducted sentiment analysis on the Star Wars movie scripts to explore the contrast in sentiment between the Dark Side and Light Side characters. To do this, we utilized the SentimentIntensityAnalyzer() function from the Natural Language Toolkit (NLTK) library to compute the sentiment score for each dialogue in the dataset, which we then classified as positive, negative, or neutral.

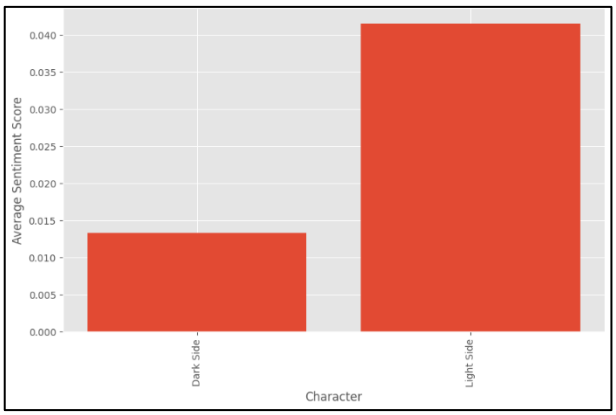

Figure 6 - average sentiment scores for each side (dark vs light side)

To differentiate between the Dark Side and Light Side characters, we created two character lists based on our knowledge of the Star Wars universe, which we used to add a new column in the dataset indicating a character's affiliation.

Our analysis uncovered intriguing contrasts in the sentiment of dialogues between Dark Side and Light Side characters. The Dark Side characters, such as Darth Vader and Emperor Palpatine, were more frequently associated with negative emotions. In contrast, the Light Side characters, including Luke Skywalker and Yoda, demonstrated restraint in expressing anger towards other life forms and exhibited more positive sentiments.

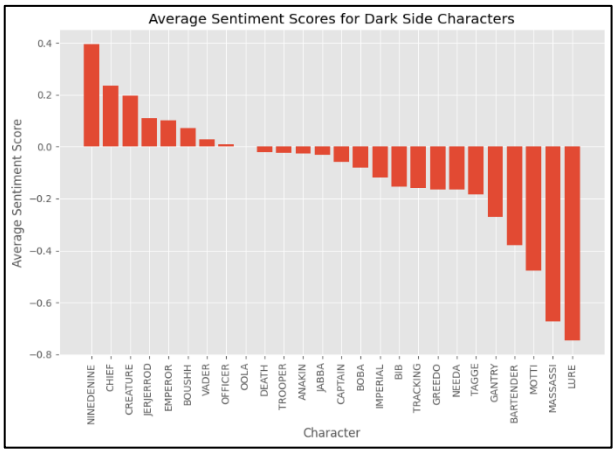

Figure 7 - average sentiment scores for dark side characters

In summary, our sentiment analysis highlighted subtle differences in the sentiment of dialogues between Dark Side and Light Side characters in the Star Wars universe, which could be attributed to the challenges we faced in differentiating between these two groups. Nonetheless, we endeavored to create the most accurate classifications possible.

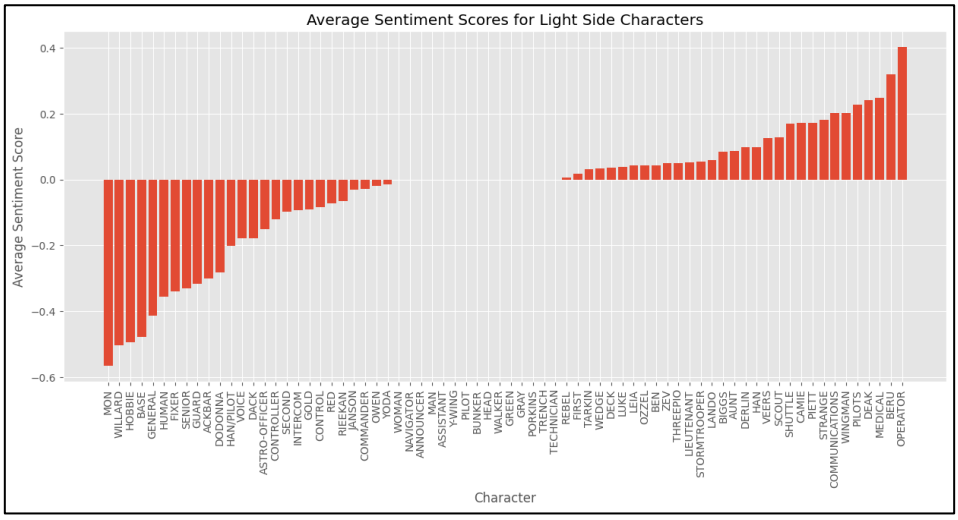

Figure 8 - average sentiment scores for light side characters