LAB – 4 REPORT

Balaji Vijayaraj

v22balvi@du.se

Introduction:

In this lab assignment, we will be using text mining and natural language processing techniques to analyze the dialogue of the Star Wars movies' characters from episodes 4-6, also known as The Original Trilogy. We will be exploring the sentiments of the characters, their frequently used words and phrases, and creating word clouds to visualize their most common words. Through this analysis, we aim to gain insights into the personalities and interactions of the iconic characters from the Star Wars universe. We will be using Python and several libraries, including NLTK, pandas, and matplotlib, to perform our analysis.

Method:

1. For finding the characters with the most dialogues in each episode of The Original Trilogy, the script dialogues were parsed and processed using Python's Pandas and NLTK libraries to count the number of dialogues spoken by each character in each episode.
2. The findings from step 1 were plotted using Python's Matplotlib library to visualize the number of dialogues for each character in each episode.
3. A new column "episode" was added to the three datasets to distinguish between the three episodes and then concatenated into one dataset using Python's Pandas library.
4. The frequency distribution of words in The Original Trilogy was discovered using Python's NLTK library.
5. A Frequency Distribution plot of the most repeated words in The Original Trilogy was created using Python's Matplotlib library.
6. Text-mining operations, such as converting to lower case, word tokenization, removing stopwords, and lexicon normalization (lemmatization), were performed on the script using Python's NLTK library. The resulting array list was added to the dataset as a new column "new_script".
7. Steps 4 and 5 were repeated, but this time the frequency distribution of the "new_script" was checked.
8. Word Clouds were created using the provided word cloud masks to visually represent the most repeated words for Darth Vader and Yoda using Python's WordCloud library.
9. The TF-IDF model was used to discover the most relevant words in The Original Trilogy script.
10. Sentiment analysis was performed on the movie scripts using Python's NLTK library to assign sentiment scores to each episode and character.

Result and Discussion:

```
Sentiment scores for Episode 4:
neg_score          0.085021
neu_score          0.820175
pos_score          0.094794
compound_score     0.004970
dtype: float64
Sentiment scores for Episode 5:
neg_score          0.083298
neu_score          0.772295
pos_score          0.144409
compound_score     0.067095
dtype: float64
Sentiment scores for Episode 6:
neg_score          0.083307
neu_score          0.788723
pos_score          0.127976
compound_score     0.051601
dtype: float64
```

Based on the sentiment analysis, it appears that the three episodes have a generally neutral sentiment, with slightly more positive sentiment in Episode 5. The compound scores for all three episodes are close to zero, indicating that the sentiment is balanced and not strongly positive or negative.

```
Sentiment scores for YODA:
neg_score          0.114163
neu_score          0.779796
pos_score          0.106020
compound_score    -0.014833
dtype: float64
Sentiment scores for LUKE:
neg_score          0.077992
neu_score          0.806638
pos_score          0.115368
compound_score     0.038590
dtype: float64
Sentiment scores for VADER:
neg_score          0.086086
neu_score          0.778357
pos_score          0.135564
compound_score     0.029026
dtype: float64
Sentiment scores for EMPEROR:
neg_score          0.123977
neu_score          0.685636
pos_score          0.190386
compound_score     0.102123
dtype: float64
```

Based on the sentiment analysis results, we can observe the following insights:

Yoda:
Yoda's sentiment leans more towards the negative side, indicating caution and criticism in his dialogues.

Luke:
Luke has a balanced sentiment score, indicating a mix of positive and negative emotions. His sentiment is more neutral overall.

Vader:
Vader has a mostly positive sentiment, suggesting he speaks with authority and power.

Emperor:
The Emperor's sentiment is mainly positive, reflecting his manipulative and charismatic nature. His dialogues may involve persuasive language, emphasizing his control and dominance.

In conclusion, the sentiment analysis aligns with the characteristics associated with the Dark Side and Light Side characters in the Star Wars universe. The Dark Side characters (Emperor and Vader) tend to have more positive sentiments, reflecting their assertiveness and use of power. On the other hand, the Light Side characters (Luke and Yoda) exhibit a range of sentiments, with Yoda leaning towards caution and criticality, while Luke portrays a balanced mix of emotions.