Lyricology: The Hipster's Guide to Understanding Music Trends Through NLP

Rishita Shroff, Kush Ramchand Raimalani, and Urvi Bhojani
DS3500: Advanced Programming with Data
Homework 3: Report
February 27th, 2023

1. INTRODUCTION

In regards to Homework 3, this is a reusable NLP framework which can handle a variety of datasets for comparative text analysis. In particular, this project registers up to 10 CSV files containing various attributes about an artist - for example, their album names, song titles, and lyrics. Leveraging this data, this project **- Lyricology -** produces corresponding Sankey, Word Cloud, and Bar Chart visualizations to indicate differences in most common words, and sentiment score for song lyrics across all files.

2. MATERIALS & METHODS

In order to create our diagrams and visualizations, we utilized a Song Lyric Dataset from Kaggle containing 21 different CSV files. Each file is uniquely labeled by a different artist name, and includes 6 columns - Artist Name, Song Title Name, Album Name, Year Released, Release Date, and Lyric.

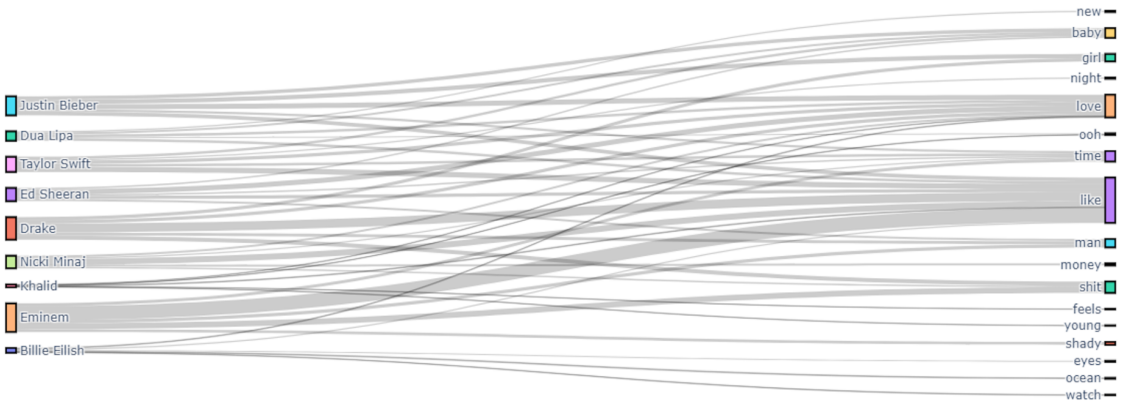
This framework is built using Python libraries - Pandas, Collections, WordCloud, Matplotlib, and TextBlob. We began by creating a class object, “Lyrics”, which creates an instance of a lyrics of songs from an artist. The purpose of the lyrics object is to perform data pre-processing, and analysis on data from files. From here, a customized parser is used which reads a CSV file, and performs steps such as removing null values, removing punctuation, converting all text to lowercase, and removing all stop words. Following this, the parser returns a dictionary containing the word count and a list of all cleaned words from a file.

Furthermore, the Lyrics class also provides a method to create a dataframe of the word counts across all artist files and sorts it to find the k most common words for each artist. This method is then used to create a Sankey diagram to display the most common words for each artist. Along these same lines, another method is created to generate a Word Cloud visual representation of the most common words in each Artist text file, split up into corresponding subplots. Finally, there is a method, “sentiment_analysis” which calculates the polarity scores of lyrics per each file which

range from -1 to 1 where -1 = Negative and +1 = Positive. This is shown by creating a corresponding bar chart.

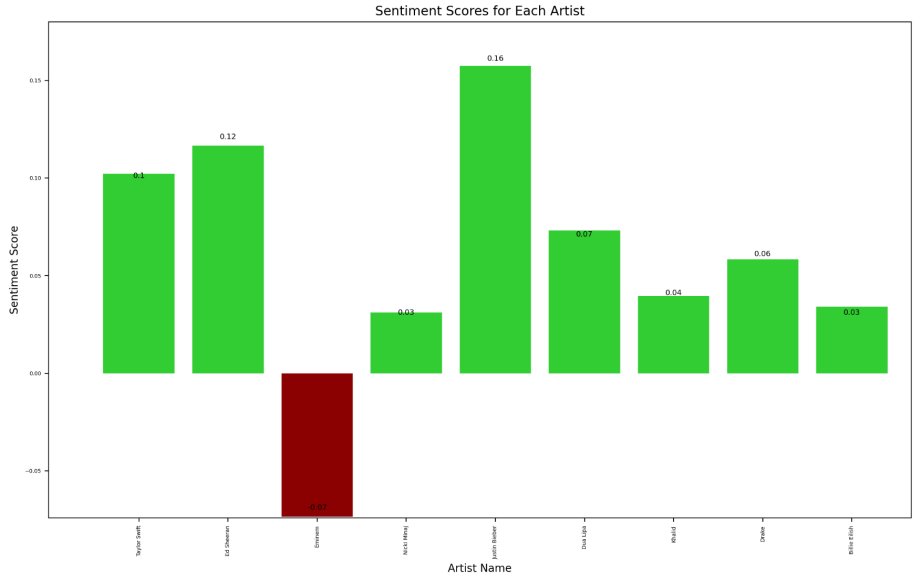
3. RESULTS & DISCUSSION

a. Word Count Sankey



b. Word Cloud Subplots

c. Sentiment Score Bar Chart



The Sankey and the subplot of Word Clouds above depict the most common words in each Artist file. For the Sankey specifically, we wanted to analyze the top 5 most common words, and found quite a bit of overlap between the other files. As seen, the word “like” is the most recurring word across all files, followed by the word “love”. From there, we can also see that the words “baby”, “shit” are quite common as well.

On the other hand, the Bar Chart illustrates the sentiment score for each artist, where the sentiment score is a metric for measuring emotional depth of text. According to the figure above, almost all artists we looked at (Taylor Swift, Ed Sheeran, Nicki Minaj, Justin Bieber, Dua Lipa, Khalid, Drake, and Billie Eilish) all had positive sentiment scores, while Eminem was the only artist with a negative sentiment score. Additionally, out of all the positive scores, Justin Beiber had the highest positive sentiment score.

4. CONCLUSION

In conclusion, our library provided us great insights into the emotional ton of several artists. Our framework produced Sankey diagrams and Word Cloud visualizations that revealed the most common words in each artist file. We found that the words "like" and "love" were among the most recurring words across all files, and that there was a significant overlap of common words between different artists. Additionally, our sentiment analysis Bar Chart showed that most artists

had positive sentiment scores, with Justin Bieber having the highest positive score and Eminem being the only artist with a negative score. These specific results provide valuable insights into the common themes and emotional tones present in song lyrics across different artists, and demonstrate the usefulness of our NLP framework for comparative text analysis.

5. ACKNOWLEDGMENT

The authors would like to express appreciation for the support of the Professor and Teaching Assistants. This work is also an equally combined effort of all authors. The paper was written and the code was developed together with all authors.

6. REFERENCES

Shah, Deep. "Song Lyrics Dataset." *Kaggle*, 8 Feb. 2021, https://www.kaggle.com/datasets/deepshah16/song-lyrics-dataset/code.

"W3Schools Free Online Web Tutorials." *W3Schools Online Web Tutorials*, https://www.w3schools.com/.

"Where Developers Learn, Share, & Build Careers." *Stack Overflow*, https://stackoverflow.com/.

Additional Resources:
- Code from class
- HW pdf
- Python Documentation