CS505 Project Milestone

Team Members

Asianna Haughton, Alex Miller, Yunqi Ni

Problem Description

Brief Description of the Project: The project aims to explore the relationship between song lyrics and their sentiment using Natural Language Processing (NLP) methods. This exploration is rooted in our shared interest in music and aligns with the NLP skills we've learned, such as tokenization, classification, and language generation.

Problem to Solve: We're investigating how the emotional sentiment of a song is reflected in its lyrics. This problem is significant because it blends the subjective art of music with the objective analysis of language, offering insights into how words in music convey emotions.

Relation to Interests and Classwork: This project directly relates to our interests in music and leverages our NLP skills developed in class. Techniques like textual data cleaning (removing punctuation, stemming, and removing stop words) and our experience with classification models (such as the one used on the Enron email dataset) are particularly relevant. Our work on lyric generators in previous homework also forms a foundation for this study.

Resources

Research Paper
MuSe: The Musical Sentiment Dataset
Contains 90,001 Entries; User-generated tags, artist, title, genre metadata (valence, arousal, dominance) MusicBrainz ID and Spotify ID
Will be used as our dataset
https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.33
This paper explores the use of the Last.fm API to create three sentiment tags – valence, arousal, and dominance. Valence is described as the pleasantness of a song, arousal is described as the intensity of a song, and dominance is described as the control of a song. This paper explores the creation of the dataset, how to use the dataset, and what it could/should be used for. We will use it to tag our songs with sentiments.

Research Paper
DEEP LEARNING BASED MOOD TAGGING FOR CHINESE SONG LYRICS
https://arxiv.org/pdf/1906.02135.pdf
Index Terms— Natural language processing, Sentiment analysis, CNN, Lyrics
This is a detailed paper about sentiment analysis using RNN, LSTM, CNN and the author makes experiments on which one has a higher accuracy. It also explored some text processing methods like TF-IDF and LIWC which we can test.

Blog Post
Spotify API and Genius Lyrics
https://medium.com/swlh/how-to-leverage-spotify-api-genius-lyrics-for-data-science-tasks-in-python-c36cdfb55cf3
This article discusses what libraries to use to work with Spotify's API. The URI in the API data allows us to access the individual song metadata. It then attaches the song lyrics from Genius.com using BeautifulSoup to scrape from the website.

Book Chapters
Natural Language Processing with Transformers (*Chapter 2- Text Classification*)
https://buprimo.hosted.exlibrisgroup.com/primo-explore/fulldisplay?docid=ALMA_BOSU151104957164000116 1&context=L&vid=BU&lang=en_US&search_scope=default_scope&adaptor=Local%20Search%20Engine&tab=default_tab&query=any,contains,Natural%20Language%20Processing%20with%20Transformers&sortby=rank&mode=Basic
This chapter discusses text classification, specifically sentiment analysis. The chapter walks through the classification of Twitter messages about a product and utilizes sentiment analysis to classify feelings in the text. It assigns the input to one of several possible labels (i.e anger, joy, fear). This is useful for us as we are trying to make our model assign an input (in our case song lyrics) to a sentiment label.

Data Set
https://www.kaggle.com/datasets/vatsalmavani/spotify-dataset/data
Data from Spotify API placed into csv.

Project Plan

Problem
The focus of the project is to explore the correlation between a song's lyrics and its emotional tone/ sentiment. We seek to investigate the connection of the interpretative artistry of music and the analytical scrutiny of linguistics to provide a deeper understanding of the emotional impact conveyed through song lyrics.

Project Flow:
Data Wrangling (Lyrics to Song)> Lyric Preprocessing (Tokenization/ Lemmatization)> Feature Engineering (Combining Metrics to Processed Lyrics; Finding Relevant Features (TD-IDF/word embeddings)> Model Selection (Neural Network)> Training (Combined Feature Set)> Tune Hyperparameters> Evaluation (Cross-Validation for Performance; accuracy, precision, recall, F1-score)> Optimization (Optional)

Data Sources: We intend to use data collected from the Spotify API found here on kaggle with various sentiment tags, the MuSe dataset of songs tagged with sentiment, and song lyric data we will scrape from the genius api using BeautfiulSoup.

Methods/Algorithms:
For the generation of lyrics: LSTM neural network
For the classification of lyrics: Feed-Forward Neural Network

Data Wrangling:
1. Limiting data size to be able to attach lyrics to songs
2. Use BeautifulSoup parser to scrape Genius lyrics
3. Attach the lyrics onto the data frame

Running on:
Google Colab (we will try SCC if Google Colab isn't fast enough)

Evaluation Strategy:
For the sentiment classification by lyrics: Cross-validation for Performance; accuracy, precision, recall, F1-score
For the generation of lyrics: coherence of probing with various inputs

Division of Tasks:
Asianna:
Data Wrangling (Beautiful Soup, TD-IDF, etc.)
Alex:
Creating a generative model to generate lyrics
Yunqi:
Creating a classification model to do sentiment analysis based on its lyrics