Proposal for Genre Categorizer

October 2, 2020

Intro

If you go on Spotify and listen to any number of songs, the Spotify algorithm will begin to recommend new music based on what you have previously listened to. This in itself is machine learning: where, based on existing data, you are able to make predictions as to what a user may want to listen to. Following this theme of music, we want to bridge the gap between lyrics and genres, in which given some lyrics, we can identify the genre of the song.

Problems

With the large number of genres that currently exist, it can sometimes be challenging to categorize a particular song. As a team, we want to create an application that can understand the relationship between the lyrics and genres, with the lyrics being any meaningful text one can input. More specifically, we are going to first utilize an NLP technique2 to analyze components of lyrics in respect to its genre and train a classifier model to predict genre of the song. Combining the identified genre and component of lyrics, we then identify the sentiment level of the song.

Method

The primary intended dataset for our project will be The Million Song Dataset1 (Bertin-Mahieux, 2011). The AWS dataset consists of various features of a song, with key features being pre-defined genre, lyrics, and energy level. If more songs are needed, we plan to scrape lyrics from an online platform using the Genius API. The first stage of the project is data parsing and cleaning. Data will be collected and stored in a customized database. Any irrelevant information will be excluded, and we will mostly focus on the lyrics(input) and the genre(output). The lyrics are filtered and selected by processing the language (tokenization, deletion and lemmatization).

Potential Results

The genre types will be represented as clusters. The raw results would be the assignment and correlation between a song and a chosen genre set. Analyzing the various features we collected, we hope to be able to visualize the outputs that are relevant to our study. For the training stage, we plan to use regression methods and Gaussian Naive Bayes to calculate the probability of a lyric belonging to a genre given its features. We then make the prediction using the MLE method to see the most likely genre the song would belong to. We will cross compare the result between different methods in order to check for accuracy. Key classification metrics such as precision, recall, and F1-score will be used to run the accuracy test.

Discussion

After genre prediction, a correlation index can be assigned given the genre, and based on the frequency of keywords, we can refine the index to reflect a more accurate sentimental level. To sum up, our model will be a supervised learning model, and potential results will be a translation to genre.

References:
1.Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.
2. Chu Claudia, Kumala Enoch, 2020, Mapping song lyrics to musical feature space, https://suraj-masand.github.io/cs4641-project/index.html
3. Tsaptsinos, Alexandros, 2017, Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network

Links: https://jeremyhua18.github.io/MachineLearning/

Youtube Video