Florence, Jameson, and Zongrui
CS 5002, Spring 2021
Final Project Proposal

1 PROJECT CONTEXT

Jameson: I've been involved with music for most of my life, whether it be middle school band, high school choir, or singing in the shower. The opportunity to combine my love of music with my academic studies in computer science is fantastic! I've always been interested in looking into the different ways that we classify music, genre being one of them. While song lyrics and topics definitely play into the definition of a genre, I think that instruments, musical structure, and other unquantifiable things such as the way a song "sounds" play a more significant part in how we view a given song's or artist's genre. However, lyrics still play a large part in determining genre: read the lyrics of a typical pop country song and a typical hip-hop/rap song and you will see a marked difference. I am therefore very curious to see how accurately we can predict the genre of songs!

Florence: My love affair with music started as a very young age as I learned the piano and later on study the flute traversiere for a few years. As a child, I wanted to play the flute so badly that I accumulated year after year, all the money my grandmother gave me for birthdays and Christmases, until I could buy one. I also inherited my grandfather violin later in life (I was not aware of it until I was an adult as he died a few months after my birth) and without surprise I am sharing my life with a professional musician which instruments are the viola and the violin. I have a passion for a broad variety of musics and I am always listening to different sounds and genre from the world. I am paying attention to the quality of the sound and the melody as much as the way to produce it... I have done some foley sound effects in my animation projects and I have a particular interest in haptic sounds applied in XR. Being able to manipulate images and sounds in my future work would be ideal and this is why this project is of particular interest to me.

Zongrui: I started to play cello when I was seven years old. When I was grade 9, I passed the highest level of cello certification exam. I also attended many orchestra music competitions, and I also performed many times in my high school and university. During the summer break from high school to university, I was interested in playing violin. Since I have a solid foundation in cello, I can perform some simple songs in a short time. But I haven't played cello after I go to university because I was busy in my homework. Now, when this music topic of final project comes to my minds, I feel it very interesting to combine music with computer programming. That will bring me a unique experience with music.

2 PROJECT QUESTION

There are many qualities of a song that can be analyzed to determine its genre, but we want to focus on the lyrics. Therefore, our question is:

Can a computational model predict the genre of a song by analyzing the lyrics?

3 PROJECT SCOPE

Due to the short timeframe of this project, the scope will be restricted in several ways.

First, we will set an initial goal of creating a successful binary classifier. This means that we will pick a genre, and our classifier will determine if a particular song is a part of that genre or if it is not. This will increase the accuracy of the classifier and simplify the data scrubbing process. If there is enough time left over once the binary classifier is built, we could expand the classifier to two or three genres; this would increase the usefulness of the classifier while still remaining relatively simple and accurate.

We also intend to use a “bag of words” model for our song lyrics. This treats each word in the song as an independent token, and predictions are made based purely off token frequency. This does not take word order into account, which will reduce the accuracy of our model. However, it will simplify our model and make it easier to tokenize our set of song lyrics.

Another restriction is that we will be analyzing only the lyrics. It is possible to build a machine learning model that takes into account multiple types of data, such as the lyrics plus the audio waveform of a song. However, this would increase the complexity of the model and is not feasible to build in such a short time frame.

4 PROJECT WORK-TO-DATE

Our team first did some research into natural language processing as well as classification predictive modeling problems. We learned that the Naïve Bayes Classifier is a common machine learning algorithm for predictive classification of text, which would work well with song lyrics. There is a Python module, ScyPi, which includes functions and methods for building and running a Naïve Bayes Classifier. We will read more into programming such a classifier to work with our song lyrics dataset.

We also looked into different datasets to be used for developing a model. While there are many sets of data available on sites such as Kaggle, many of them only include songs from a specific artist or genre, and most are limited to around 10,000 songs. The most promising dataset appears to be Million Song Dataset’s musicXmatch dataset, which includes the 5,000 most common words from 210,519 songs (for the training set, at least). Each song contains the lyrics already converted to a bag-of-words format, which will simplify our data processing. It also contains a separate test set of lyrics. However, each song is only listed with a Million Song Dataset track tag. To get the genre of each song, we would need to compare the track tag to another dataset to get the artist, and then compare the artist to yet another dataset to get the genre. We will continue to look for a more user-friendly dataset that is still large and accurate.

We also researched different ways to process the lyrics into tokens that can be used by a machine learning algorithm. The NLTK (Natural Language Tool Kit) module for Python provides several useful libraries for text processing, including functions to tokenize words as well as libraries of stop words (commonly-used words that are not useful for categorizing). We will continue to look for more useful modules and methods to help us clean and feature engineer our dataset.