Project Proposal

Trever Andrus, Conner Greenhalgh, Sam Johnson

1. What research question are you trying to answer

Our group is trying to determine if there is a relationship between lyrical content and music genre – can a machine learning model determine the genre of a song solely based on its lyrics? In short, we want to try and build a model that will take in a song’s lyrics, then classify that song in to one of a specified number of genres.

If we end up doing that fairly quickly, below are some other potential questions we may attempt to answer:

* Can we identify an artist by their lyrics?
* How do sentiments compare between genres?
* Is one genre more positive than the rest?
* Have songs gotten more explicit over time?
* How do genres rank in terms of explicitness?

2.  What data are you planning on using and where are you going to get the data?

To train and test our model, we need to acquire quite a bit of labeled data. Specifically, a database of song lyrics, where each song is identified by its genre. (If time permits, we also might try and identify a song by artist). To acquire this data, we hope to access an API furnished by Genius – a platform that boasts a wide variety of music data. There are a couple of python packages that were created to access these API’s – one of them being lyricsgenius. It appears that we may have to apply for an API key to get access to this data. (If we can’t get that key in time, we may have to look for other options to get the data –scraping may be required).

3.  How much data wrangling or cleaning will be required.  Are you planning on making new features?

It’s hard to tell at this point what cleaning will need to be done, chances are most of the data from the API will come fairly clean. Ideally it will come with genre labels already, but if not, we may generalize and add those features based on artist. We will also probably have to clean the lyrics of the songs themselves to remove filler words – words like “the is are” will distract the model from the actual identifying content of the song, but we will make that decision after looking at the data and building some models. (We will most likely follow general cleaning procedures consistent with NLP models).

4.  What types of EDA are you planning on doing?

We will need to look at the proportion of genres represented by the data we collect, to make sure we are feeding the model enough information to make accurate classifications. We will also try to identify some of the words that are most commonly used in each genre, and see if there is much overlap between those. If not, we may be able to apply some weight to these “buzz words” to help our model correctly identify genres.

With all this music data, (and if time permits), we may increase the scope of our project to include some more insights (like if music lyrics have significantly changed over time, or if lyrics are becoming more explicit over time).

5.  What kind of machine learning model will you most likely use?

We will likely try a variety of classification algorithms, (but because most NLP models are built around neural nets, chances are we will have to work with those). Before jumping to those more computationally expensive measures, however, we will apply some less expensive models – decision trees, random forests etc.)

6.  What are your anticipated results?  Do you have a hypothesis?

We hypothesize that there will be a significant distinction in lyrics by genre. We hope that our model will be able to predict the genre of a song based on its lyrics with a reasonable amount of accuracy. (We have yet to define what we consider reasonable, but we will probably make that decision after we play with the data for a bit).

7.  Outline a plan of how everyone in the group is going to participate in this project

Connor- API

Sam-Feature Engineering

Trevor- Model Basics

8.  Who's GitHub account are you planning on using to host this project?

Sam Johnson (smbjohnson)