CDS Geeks "Predicting Song Popularity on Spotify" Project Update

As of now, each of team members is working on a different part of data and analyses. Our team has met with Prof. McPhee. He advised us that the audio features we obtained from Spotify API can only provide limited information about a song, incapable of predicting popularity. While we are still going to work on optimizing the baseline model with these features, we believe only using the original data we received from Warner Music Group is not sufficient for the prediction. Based on this conversation, our team is exploring a few different data sources: 1) We have obtained song lyrics, producer information and songwriters through Genius API. Currently, we have about 200k English lyrics that can be matched to the main database 2) Genre from Spotify API, grouped by Time Warner classification 3) We are waiting for WMG to provide audio samples (15,000 songs) of mp3 and midi files. We have the code ready to decode midi files as soon as we receive the new data.

With audio data we will be able to apply deep learning for forecasting song popularity.

With the 200K lyrics data, there are some NLP techniques we would like to try: topic modeling (using gensim's LDA models), repetitiveness in the lyrics, etc. We're working on developing these models and got some preliminary results on the lyrics scraped from the Genius API. Also, we have a list of the top artists from Warner Music, so we are planning to do some similarity calculation to see if a given song is similar to a top artists' song, and if high similarity score can predict the high popularity score.

Challenges:

1) data storage: midi file is highly compressed. Once a file is decoded into a pandas data frame, it typically needs about 2MB. We are anticipating 15k songs from WMG, which means an additional 23G data will be added for the study. Besides, each mp3 file is about 5MB, so the 15k mp3 files will take up 57G space. Totally we need about 80G. We are facing the challenge to efficiently store and process it.
2) computational resources: in addition to 23 G midi and 57 G mp3 data, we also have an NLP database that requires running CNN and intensive computational models. For topic modeling, the LDA models from gensim took around 2 hours to train on the lyrics data on a local machine. It is difficult to install these libraries on Dumbo without the admin permission. We are currently planning to use Google's free open source Colab as well as internal Dumbo server for other applications.
3) Sample size: with only 15K audio samples available and given several sources of data, matching the datasets will result in 15K sample size, unless we can obtain more audio files.