ORIE4741: Project Proposal

Stephanie Zhou (sz244), Danny Yang (dzy4), James Chen (jzc8)

Due September 22, 2017

Objective

Given metadata of a song, we would like to classify the genre of the song.

Motive

The ability to provide relevant music suggestions tailored to each user is a major part of the value that streaming services provide. This functionality takes advantage of the large music library available, while also allowing users to discover new music that they will enjoy. Major streaming services like Pandora, Apple Music, and Spotify have all attempted this feature with varying degrees of success, and we feel like music recommendation systems can still be improved upon. A big part of developing an effective suggestions system is understanding exactly what makes two songs similar, and what characteristics define music from different times, genres, and subgenres.

Dataset

The dataset that we are using is the Million Song Dataset, a 280GB dataset that includes information on one million popular songs. It contains information about the features of the music such as tempo and key, and also has labels for artist, genre, popularity (as determined by another source), and other metrics that we can use to train our model.
Link to dataset: https://labrosa.ee.columbia.edu/millionsong/
Link to genre labels: http://www.tagtraum.com/msd_genre_datasets.html

Approach

Before we work to build a classifier for genre, it would be prudent to perform some exploratory data analysis first. There are several questions we can choose to explore. We would like to know which features in our dataset are most helpful in determining genre. We could perhaps hand select features or just use a dimensionality reduction algorithm. We would like to visualize the data somehow and see if the songs naturally fall into several different clusters. In that case, a distance based classification metric might produce good results. That also requires us to define a robust way to vectorize songs such that songs of a specific genre are close in vector space while songs of different genres are further away. Other interesting insights we can obtain from the data would be determining if popular songs tend to be under a certain genre along with the features that make a song popular as well as determining similarities between songs of a particular artist.