Project 1 Report

Min Xiao, Tianyu Yang, Cheng Zhong

**Data Science Problem**

Music evolves from simple patterned sounds back in ancient time to a rich, well developed, and universal culture around the world. Meanwhile, the uniqueness and commonness long existed in the population lead to discrepancy in tastes and contributes to formations of music genres. One might accurately identify the type of a song with perception, yet such self-developed standard hardly ever tells the whole story about either the song or the genre. Nowadays, we are no longer restricted by perception based criteria. Instead, quantitative measures that describe music features and widely used in music analysis, could assist us in providing more insights in a song and the genre it belongs(Ridley and Dumovic, 2016). In this study, we hope to establish a relationship between the sound features and music classification. Meanwhile, by associating with music ranking, we would also like to explain music trending on a fundamental level.

**Data Collection and Potential Analysis**

We plan to use playlist and soundtrack data from Spotify. Spotify develops categories for playlists, which we will be using as proxies for music genre. Besides, Spotify also has detailed metrics for sound features, which quantitatively and thoroughly describe the sound tracks. Each song and playlist have unique ids, giving conveniences when merging the datasets. The Spotify data offers a well-constructed platform where we can conduct analysis for our research goal. For music ranking, we will collect billboard weekly music rankings including the overall rank and ranks by music genre, which provides channels for more detailed analysis involving trending. On the aspect of analysis, we are able to conduct either descriptive or predictive analysis with the versatile datasets obtained from Spotify and Billboard. A few topics we interested in includes:

· what are the similarities among audio features for sleeping time music and dining music?

· what is the most important feature given the music genre?

· what are the differences among R&B, Blues, and soul music on the fundamental level? Prediction of ranking position using sound feature

· what is audio feature’s role in music classification?

and so on.

**Variable Description**

Spotify Dataset (developer.Spotify.com/web-api/object-model/)

***Playlist ID (string)***: The Spotify ID for the playlist containing  
the sound track

***Sound Track ID (string)***: The Spotify ID for the track

***Album(string)***: The album on which the track appears

***Artist (string***): The artists who performed the track

***Available Market(string):*** A list of the countries in which the track can be played, identified by their ISO 3166-1 alpha-2 code

***Duration(integer):*** The track length in milliseconds.

***Song Name***: name of the song

***Category(string):*** The name of the category.

***Popularity (integer)***: The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.

***Acousticness(float):*** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0  
represents high confidence the track is acoustic.

***Danceability (float):*** Danceability describes how suitable a track is for dancing based on a  
combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

***Energy(float)***: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example,  
death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.

***Instrumentalness (float)***: Predicts whether a track contains no vocals using float ranging from 0 to 1. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental  
tracks, but confidence is higher as the value approaches 1.0.

***Key(integer)***: The key the track is in. Integers map to pitches using standard Pitch Class notation.  
E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. Integers from 0 to 11.

***Liveness(float)***: Detects the presence of an audience in the recording ranging from 0 to 1. Higher  
liveness values represent an increased probability that the track was performed  
live. A value above 0.8 provides strong likelihood that the track is live.

Loudness  
(float): The  
overall loudness of a track in decibels (dB). Loudness values are averaged  
across the entire track and are useful for comparing relative loudness of  
tracks. Loudness is the quality of a sound that is the primary psychological  
correlate of physical strength (amplitude). Values typical range between -60  
and 0 db.

Mode  
(integer): Mode  
indicates the modality (major or minor) of a track, the type of scale from  
which its melodic content is derived. Major is represented by 1 and minor is 0.

Speechiness  
(float):  
Speechiness detects the presence of spoken words in a track. The more  
exclusively speech-like the recording (e.g. talk show, audio book, poetry), the  
closer to 1.0 the attribute value. Values above 0.66 describe tracks that are  
probably made entirely of spoken words. Values between 0.33 and 0.66 describe  
tracks that may contain both music and speech, either in sections or layered,  
including such cases as rap music. Values below 0.33 most likely represent  
music and other non-speech-like tracks.

Tempo (float): The overall  
estimated tempo of a track in beats per minute (BPM). In musical terminology,  
tempo is the speed or pace of a given piece and derives directly from the  
average beat duration.

Time Signature (integer): An estimated overall time signature of a track. The  
time signature (meter) is a notational convention to specify how many beats are  
in each bar (or measure).

Valence (float): A measure  
from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks  
with high valence sound more positive (e.g. happy, cheerful, euphoric), while  
tracks with low valence sound more negative (e.g. sad, depressed, angry).

Billboard Dataset

Artist (string): Name of the artist.

Last Position (integer): Last week’s ranking position.

Peak Position (integer): Highest ranking position.

Rank (integer): Current ranking position.

Title (string): Name of the song.

Weeks (integer): The number of the weeks the song stays on the billboard rank.

Date (datetime): releasing date of the ranks.

Genre (String): Name of the ranking genre

Data Collection

To collect data from Spotify, we used the ‘spotipy’ package from github. Spotify started to uses Oauth2 authentication for every request. Using the package simplifies our workflow for extracting data. To avoid privacy issues, we only obtained information in the public scope. To start with, we identified the category names listed on Spotify. Then we used the category names as searching keywords to get related playlists. After, using the sound track ids contained in the playlists, we were able to acquire information on the sound tracks and their audio features. Finally, we merged the sound track information and audio features meanwhile assigning a new variable indicating what category name the songs belonged to.

To collect data from Billboard, we used the ‘billboardpy’ package from github. The package generalized the web scraping process to functions with given searching parameters. We picked three overall rankings: ”Hot 100”, “Billboard 200” and “Greatest Hot Singles 100’’, and one ranking in each of the nine categories Billboard has and acquired rankings in the past year (the most recent 51 weeks from Sept. 30 2017). At the end, we merged the all the rankings, meanwhile assigning a new variable indicating the genre.

Data Issues

For the Spotify dataset, given the collection methods, object models of Spotify and the nature of data recording, we anticipate issues rising from the following perspective: missing value, bad value, wrong data type, collapsing of dimensions, and irrelevant variables. To start with, the raw dataset contains variables irrelevant to the study, only for internal use. Such variables should be dropped. Then, since the dataset is multidimensional, making accurate projections in a reduced dimension according to axes requires extracting information from the higher dimensional object (json string). Next, we may have variables with wrong data type, missing values, and bad values (out of bounds). Under the current work flow, we will not be able to identify outliers, and inliers which requires further analysis of the data (Broeck, Jan Van den et al.,2005).

Data Cleanliness

To develop a criterion for the cleanliness of the data, we used the score composed by the sum of errors (missing values, bad values, wrong data type) percentages. Duplicates and irrelevant variables were deleted beforehand. 0 is the lowest score, representing no need to be cleaned, and the higher the score, the “dirtier” the variable.

Data Cleaning

Work Cited:

Ridley, Richard, and Mitchell Dumovic. “Classification of Artist Genre through Supervised Learning.” Oct. 05 2017

Broeck, Jan Van den; Cunningham, Solveig; Argeseanu Eeckels; Roger, Herbst. “Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities” Oct. 05 2017