CSCI 1951A: Data Science - Spring 2019

Final Project Pre-Proposal:

Group members:

- Nikhil Pant - npant - nikhil_pant@brown.edu
- Deniz Bayazit - dbayazit - dbayazit@brown.edu
- Athiwat Thoopthong - athoopth - athiwat_thoopthong@brown.edu
- Cancan Huang - chuang25 - cancan_huang@brown.edu

Blogpost Link: https://cancan233.github.io/MusicalTribes/

Capstone: None of us are taking this course as a capstone.

Vision:

- **Idea:** There is a lot of data on music ratings ranging from sites like Yahoo or APIs like Spotify. This is largely due to the fact that companies like Spotify and Apple Music have been working on how to make better music recommendations with data science techniques. However, although the data is out there, the analysis behind these ratings are not public for business reasons. To the extent of our knowledge (and Google searching) there isn't a large visualization of clusters of taste in the world based on a large dataset. There are a lot of blog posts on how to visualize individual taste or on how to find the minimal distance between artists, but we would like to see what's out there other than our own taste.
- **Expected Result of the Project:** We would like to get an analysis (maybe a type of classification) on people's taste of music, a clear visualization on the "tribes/clusters" of music listeners, and possibly a recommendation algorithm based on the way we have classified people's taste. We would like to find out the prominent features that make the decision between one music taste to the other and what it takes for a music piece to be more popular than the other. Since we might be using datasets dating from previous years, the analysis will be very task-dependent and might not be reflective of the current state of music taste.

Data:

- **Where & How:** the following are examples of sites/APIs we would like to collect data from:
    - https://webscope.sandbox.yahoo.com/catalog.php?datatype=r
      We would like to collect the music ratings datasets dating from 2004. We can get this dataset since we are students. We will probably use this instantly available dataset to experiment, hypothesize and analyze while we scrape from the Spotify API for the final dataset (**size:** millions of ratings, depends on which year)
    - https://developer.spotify.com/documentation/web-api/reference/
      Spotify's API is very extensive and allows to get information on artists and users. The endpoints are referenced clearly. We can use these for ratings.

- **Cleaning:** We plan on cleaning the data by dropping features that don't seem like impacting variables in the problem we are tackling. We also think of grouping certain genres with IDs and replace other types of "string" data points with numbers if it seems suitable for the analysis.
- **Storing the Data:** Although it is hard to pin down at this point of the project, we would like to use a database with SQL.

Methodology:

- **Analysis:** there are several types of techniques we would like to use to analyze the data
    - Clustering Models: since our goal is to visualize tribes/collections of music taste we would like to use clustering models like kNN or k-Means
    - Predictive Modeling: for recommendations we would like to use a predictive modeling
- **Visualization:** We would like to visualize our results in a way that shows the classification we have have found.

Plan of Action:

- First TA check-in:
    - Collect the data: scrape Spotify.
    - Clean the data: both of the databases.
    - Store the data in an SQL database.
    - Jot down certain hypothesis around what type of functions/algorithms to use and which features seem to matter.
- Midterm Report:
    - Start the analysis listed out above.
    - Visualise at least a small part of our dataset (if it takes too much processing).