CSCI 1951A: Data Science - Spring 2019

Final Project Pre-Proposal:

Title: The Sound of Censorship

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/Sound-Of-Censorship

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 user data and data science techniques. Although data for musical analysis is mostly open-source, not a lot of user data can be given to the public for privacy reasons. However, one dataset collected from Last FM called LFM-1b dataset gives us important information like the country/age/listening habits of the user (http://www.cp.jku.at/datasets/LFM-1b/). There is also explicit discussion on internet censorship and other types of freedom of expression restrictions analyzed and indexed per country from international organizations like Freedom House or the World Bank.
- **Expected Result of the Project:** We would like to get an analysis on how censorship affects the listening habits of a user and how this is controlled by predictors like user information and country indexes/rankings (ex: are younger people more willing to use the internet as a resource to go around censorship when the content involves music?).

Data:

- **Where & How:** the following are examples of papers/sites/APIs we would like to collect data from:
    - **User + Listening Events Data:** http://www.cp.jku.at/datasets/LFM-1b/ This very large dataset consists of tracks/albums/users/listening events. Although it's pretty much well organized and clean we would like to make certain modifications and drop certain features that hypothetically don't seem too prominent to us.
    - **Press + Internet Freedom:** The World Bank and Freedom House datasets (https://tcdata360.worldbank.org/indicators/h3f86901f?country=BRA&indicator=32416&viz=line_chart&years=2001,2015 and https://freedomhouse.org/report/freedom-net/freedom-net-2018) consist of indexes and rankings related to the freedom of expression in a country.

    - (potentially: https://developer.spotify.com/documentation/web-api/reference/ Spotify's API is very extensive and allows to get information on musical features. The endpoints are referenced clearly. We can use these to learn what type of features of songs are prominent when the user is located in a country with heavy censorship.)

- **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 replacing 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. However, NoSQL could also be a choice because of the size of the dataset. We're also thinking of Google Cloud and Colab.

Methodology:

- **Analysis:** there are several types of techniques we would like to use to analyze the data. However the most important one would be *predictive modeling*: we would like to predict how censorship can affect listening events.
- **Visualization:** We would like to definitely visualize the model we have have found. Other visualizations could include facts/charts about the specific train/test population.

Plan of Action:

- **First TA check-in:**
    - Collect + 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).