Homework 6 (save as PDF in the end)

Final Project Topic:

<https://developer.musixmatch.com/documentation/music-meta-data>

<https://developer.musixmatch.com/documentation/api-methods>

For the final project topic, I decided to explore and look into the music field. By calling the API to Musixmatch, I was able to retrieve the datasets that I will use for my project. Musixmatch provides access to song lyrics as well as other attributes. With further analysis using some of the datamining tools shown in class, I’m interested in how, if any, of these attributes correlate to song lyrics.

Spotify provides data on

Number of user followers related to genre of music user listens to?

Popularity of genre? Popularity of album?

Lastfm provides data

Attributes/Description: Nature of the attributes

Text/string

Timestamps

Numbers

(represent? Taking specific values?)

Frequency of a word that occurs in a song/track

Frequency of a word that occurs in an album of an artist

Frequency of a word that occurs in a genre (e.g. top hits)

Does the frequency of word correlate to song ratings?

Properties/Basic Statistics:

range of values, their distribution, their density (are there any missing values? if yes how do you intend to cope with them?) and any irregularities that you may have noticed.

Some irregularities that are seen in values are songs that don’t have lyrics. These instrumental tracks will be handled with binary coding where if the track has lyrics the value will be assigned as 1 or true. Tracks without lyrics will be replaced with a 0 value or empty for false (e.g. fillna(0)). It is important to filter tracks without lyrics in separate genres.

Hypotheses:

Frequency of a word that occurs in a song/track

Frequency of a word that occurs in an album of an artist

Frequency of a word that occurs in a genre (e.g. top hits)

Does the frequency of a word correlate to song ratings and popularity? How does genre factor?

After obtaining top tracks from the same genre, are there any trends that cluster words to the same group or centrality?

Can song lyrics predict music genre?

We don't really need to see your code (the data gathering and the plotting). The only thing we want to see is a report with what I mention in the homework description and the corresponding figures. You should create a new directory for this homework under your submissions directory and upload your report there. **Make sure that the report is in pdf format.**

In this homework, you are going to work on the first part of your project assignment, i.e., you are supposed to collect and do some primary analysis on the data that you will work on.

To be more specific: you need to collect (srape, call APIs or download) the dataset or datasets that you will use for your project.

Once you have acquired your dataset, you are expected to provide a description and some basic statistics for it. In the description, you will talk about the nature of the attributes that you have (is it text? timestamps? numbers? what do they represent? are they taking specific values?) The basic statistics that we expect to see are some insight about the range of values, their distribution, their density (are there any missing values? if yes how do you intend to cope with them?) and any irregularities that you may have noticed. The goal of this step is to allow both you and us to better understand what you are working with.

After you show different properties of your dataset, you should list and briefly describe the different hypotheses that you plan to test/verify in your project and highlight how you intend to use your data for this purpose.

Submission deadline: Friday, 3/17 midnight