CPE 695 Final Project Proposal - Zheyu Xiao

Title: Chart performance predictions of *Billboard “The Hot 100”* songs

The *Billboard “The Hot 100”* chart, initiated on August 4th, 1958, is a weekly published standard and widely- acknowledged record chart of the music industry in the United States. It is the most prominent and straightforward indication to the popularity and public-acceptance of songs. Chart rankings are based on sales (physical and digital), radio play, and online streaming in the United States.

The currently existing projects online are mostly simple data rankings, visualizations, and categorizations. Examples are rankings of total weeks on chart and artist-based filtering. This project aims at standardizing the chart performance of songs from the weekly data, and looking for performance predictions based on several parameters concluded from data cleaning. The project will endeavor to formulate a predictive model of the chart performances; as a lot of entries (mostly “short-lived” songs) are likely to be unhelpful, the data feed for the machine learning algorithms are subjected to continuous adjustments to finally present a sound prediction.

Main Sources (and not limited to):

1. Billboard "The Hot 100" Songs: A Collection of "The Hot 100" Charts on Billboard, https://www.kaggle.com/dhruvildave/billboard-the-hot-100-songs
2. The “All Time Top 100 Songs”, Official rankings from *Billboard*, published in 2018 (60th anniversary)

The project will be divided into four major parts:

1. Cleaning of raw data
   The raw data set does not provide any columns of chart performances, but rather the weekly ranks of songs (current week and previous week) that repeatedly appear until they drop out of the chart. This step aims at reorganizing the data into intuitive and (relatively) concise entries and categories based on individual songs. It is a necessary process for all tasks afterwards

2. First-round regression
   One important aspect of *Billboard Hot 100*, and the world we are living in, is that the way we obtain and enjoy music records have been changing over time along with technological advancements. As music videos, MTV, online purchases, and online streaming emerges, the chart performance of songs changes significantly. One obvious change is that songs in our age (2010-20s) tend to stay on chart significantly longer. *Billboard* has an undisclosed method to weight these

parameters to make a fair comparison for hits of different ages. We will run a regression based on the disclosed “All Time Top 100” to calculate the weight of songs with respect to the decades they are in.

3. EDA
   This step focuses on seeking for proper classifiers and categories that identify songs and their chart performances. Examples are, and are not limited to, month/season of entry, entry with or without albums, entry position on chart, and total weeks in top 10/20/50. New (and revised) data sets will be generated from the cleaned intermediate data
   This step is not limited to fixing and normalization of the data. It is necessary to combine Step 3 and Step 4 (i.e. running through algorithms) to determine the classifiers.

4. Model training with various machine learning algorithms
   Several algorithms should be applied to the post-EDA data set. Decision trees, ANNs, genetic algorithms, etc. will be used to classify the datasets and train the models. The activations, inner structures and model evolutions should be carefully identified and taken as the essential steps of this research project. The reasons and rationality of parameter selections (layers, learning rates, etc.) will also be important parts to be addressed in the final report.
   Details of training models are, of course, subjected to updates throughout the span of the project.