Project 2 Description

COSC 311

Kalyn Howes & Chloe VanCory

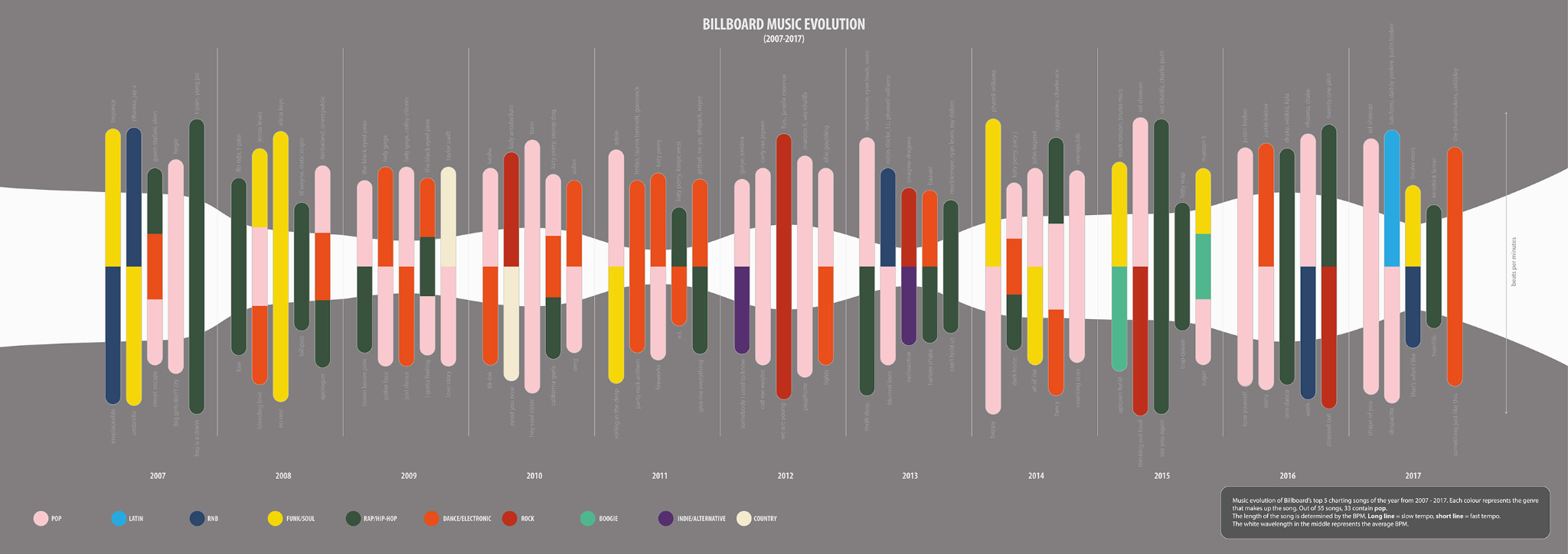
11.23.21

In this project we will use the Spotify Top 200 Charts (2020-2021) dataset from Kaggle, found here, [Spotify Top 200](https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021). We will implement several machine-learning algorithms and show their effectiveness on the chosen data. This data includes all of the songs that have been on the Top 200 Weekly (Global) charts of Spotify in 2020 & 2021. Each song has a number of features including: the highest position of the song, the number of times that the song has charted, the week when the song had the highest position, the name of the song, the song ID (provided by Spotify), the approximate number of streams the song has, the artist, the number of followers the artist has on Spotify, the genre, the release date, the weeks charted, and many more. This dataset was mostly gathered from spotifycharts.com, with some features being calculated based on numerous other features. The most obvious “classes” in this data are the different genres of each song. However, the songs could also be classified by artist, tempo, energy, chord, etc.

The first algorithm we plan to implement is for decision trees. Decision trees are used to go from observations about an item (the branches) to make conclusions about an item (the leaves). More specifically, we would create a classification tree. In these, the leaves represent class labels and the branches represent the conjunctions of features that may lead to those class labels. In our data for example, the decision tree can be used to go through a song’s features such as genre, tempo, danceability, etc to determine how many times it is charted. A decision tree would be successful with the Spotify charting data since there are so many qualities of a song that could cause it to be popular. Once the tree is created, we could see the “perfect formula” for getting a song to chart on Spotify.

We also plan to implement the nearest neighbors algorithm. The nearest neighbor method takes in all available data in order to train the algorithm. From here new data can be classified based on how its neighbors are classified. This data includes many features that have been quantified, which is helpful in determining a “distance” between two songs. An example use of this algorithm could be to determine the number of streams a song has based on when it was released, the number of weeks the song charted, the genre of the song, or many more.

Visualizing this data could be very intriguing. We could use an animated visual of which genres were most popular based on when they charted, and move through month by month. Another idea is a graph where the x-axis includes the date released and the y-axis includes number of streams while each point is a song, however, the size of the point is determined by the number of followers the artist has and the color is determined by the song’s genre. Perhaps we could even include a graph that looks like a soundwave (example below) or is on sheet music with each note representing how long a song was on the charts for (a sixteenth note representing a shorter period of time and a whole note representing the longest).



Using the Spotify data we could draw many interesting conclusions. Some obvious ones could include the biggest current artists, the top genres, etc, however we can go even further. Since this data was from 2020 and 2021, COVID is obviously a big factor in this time frame. We could draw conclusions about the types of music people needed to listen to during this time due to the saddening events. It may also be good to consider the impact TikTok has had on the top charts of music these past few years and how TikTok use skyrocketed during quarantine.