Exploring Popularity in Music

Project Report and Data Analysis

Team Shrug: Amrish Selvam, Phan Trinh Ha, Timothy Jin, and Gary Yao

Special Thanks to Dr. Mehmet Koyutürk

Summary

How a particular song gains popularity is usually of great concern to artists who wish to replicate the success of other artists. As musicians ourselves, our group set out to investigate the popularity of music through statistically testing. However, predicting whether or not a song will be popular or not depends on a great variety of variables such as lyrics, genre, style, pitch, tempo, instruments used, and the time period in which the song came out in. Therefore, we have narrowed our view to one such variable, lyrics, and how the use of certain words and lyrics may be correlated to the popularity of a song. In this report, we discuss our methodology and our findings as well as our overall thoughts and conclusion.

Project Definition

Project Scope

Initially, our project attempted to consider numerous factors in order to create strong predictive model, including, but not limited to genre, era, time-frame, sentiment of lyrics, rhyme density, and influences of real world states such as economics and/or war periods. However, upon further evaluation, we quickly realized that the resources necessary to complete such a project was beyond what was reasonable for our team. We thus incorporated class feedback to create a much narrower scope which focuses simply on the lyrics themselves and the popularity of music. We retroactively included the genre of music later into our project after the information became available through our dataset. As a result, our project largely dealt with songs and four attributes of the songs, an identifier (song title), the lyrics of the song, the popularity of the song, and the genre of the song. Despite only having three attributes a plethora of possible features, for example the existence of a word in lyric, or the number of words in a song, could be extracted. Using the features extracted from the lyrics, we hoped to predict the song's popularity and/or genre.

Defining Popularity

For the sake of our project, we had to define what popularity meant. A few initial thoughts, included the number of weeks on billboard's top 100 list, the number of plays on a YouTube video, and the number of plays on SoundCloud. As we researched the topic, we realized that it would become difficult to find an attribute that could quantify a song's popularity from a song of any time frame. It quickly became obvious, that we would have to stick to more recent titles (within the past century) to have more data available. Sticking to this time frame meant that we could not use electronic play counts from YouTube or Spotify as they would highly bias more recent songs. Additionally, the population of the United States was 76,212,168 in 1900, which is less than a quarter of the United States population today, so the popularity attribute would have to be population independent. This meant that using the Billboard Top 100 charts would be the obvious choice. After successfully scarping the data from Billboard, we realized that all the songs that made it onto the list were already quite popular, so defining more popular songs from less popular songs within this list was in practical and did not represent all songs accurately. Additionally, after plotting distributions of the number weeks that a song would stay in the top 100, we saw a bimodal graph that loosely followed a power-log distribution (Figure 1). From this we concluded that this attribute would make for a poor definition of popularity.

After presenting our work and receiving additional feedback, we decided to narrow our scope to the past 15 years. Due to this shift in time frame, we could use Spotify, which could be scraped for play counts of more recent songs. This provided the primary advantage that unpopular songs (with far fewer play counts) could be compared against very popular songs with high play counts. Since the population difference is not too great within the past 15 years, the popularity attribute would not need to be population independent either. For these reasons, we defined a song's popularity as the number of plays it has on Spotify. We used this attribute in various ways within our data

Methodology

Data Collection - Scraping

We initially attempted to scrape data from MetroLyrics, but this proved to be difficult due to inconsistency in the url. We then scraped AZ Lyrics as the url was consistent, although it could still fail due to abnormal song titles. These arose when special characters were used or artists had multiple aliases. Although this method was initially successful, we ran into issues with the site blocking our IP address due to the nature of scraping. Because of the inconsistency of finding the right songs and the IP address blocks, we decided to switch to Genius lyrics, which has its own API for data miners to use. Fortunately, Genius could properly match songs with variations in song titles and names and since we used their approved API, we did not get blocked. By registering for an account, lyrical scraping was fairly accurate and consistent with _____ retrieved songs for every _____ attempts. On average it took _____ minutes per song.

As for scraping population metrics, we ran into fewer issues overall. Scraping Billboard did not take very long in itself, although setting up the scraping algorithm took some time. There were hardly any inconsistencies with this algorithm. When we switched to Spotify, we initially attempted to use Spotify's of API for retrieving metrics. However, their API did not provide the play count attribute, which was essential to our project scope. As a result, we created our own scraping algorithm by adapting public code found on GitHub. We formulated the algorithm to search for pseudorandom songs within a genre so that we could find songs with varying popularity. Finding 500 songs per genre took approximately _____ hours to run.

Dataset Description

We initially intended to MetroLyrics (with an already scraped dataset on kaggle), Billboard Top 100, and the iWeb Corpus to conduct our study. However, as our project scope changed, we shifted to using Spotify and Genius Lyrics. Our attributes within our final dataset included Title, Artist, Count, and Lyrics. We used the Title and Artist as the primary key for both tables and used it to join our data. Count was an attribute of the Spotify table and Lyrics was an attribute of the Genius Lyrics table. We also had a hidden 5th attribute, Genre, as we had multiple tables, each corresponding to a genre. Technically, a song could fall under two genres, meaning that the true primary key is Title, Artist, and Genre, but count and lyrics would stay consistent amongst repeats. The song would simply be repeated amongst both titles. So each row or instance in our table has a title, artist, count, and lyrics, but each song belongs to a table with a specified genre, meaning that a song could have multiple genres if it appears on multiple tables.

Title, artist, and genre were all short strings, while lyrics was a much longer string, and count is an integer. Title describes a song's title. Artist describes a song's artists, which may have a few primary artists and featured artists. Count describes the number of times a song was played on Spotify by the date it was scraped on. Lyrics are all the lyrics in the song, with profanity uncensored.

One major consideration included how lyrics would interpreted and used. Often times, lyrics would have erroneous characters that do not have any meaning, such as a new line character, tabs, and so on. Our data was modified as a result to remove these characters and clean up the data. Additionally, songs in a foreign language may have a completely different subset of words used, which would make later analysis and the drawing of comparisons difficult. Many popular songs may also contain no lyrics at all. Therefore, we restricted ourselves to songs that are popular in the United States, which have mostly lyrics in English and genre's which typically have lyrics. A final consideration are words that have the same spelling but have distinctively different meanings and pronunciations such as "read". Since we could not distinguish the difference by the data given, nor account for all such words, there was no fix to this solution and we did not account for this in our model.

Overall, we had five tables, corresponding to five genres: rock, pop, hip-hop, folk, and country. There was quite a few overlap between pop and hip-hop, but very little or no overlap between the other genres. Each table has 500 instances or songs, totaling to about 2500 instances, and _____ distinct songs.

Statistical Analysis

As a baseline, we first visualized our data via histograms and boxplots to gain a better understanding of what how are data was represented and analyze for any anomalies. We created histograms of the variation in the words (distinct words in a set of lyrics), number of select popular words in a song (ie. Love), number of common words in a song (ie and, the, or a), total number of words in a song, and number of appearances of certain words in a song. These histograms served as both a sanity check and a way we could better understand the data we are dealing with. We also attempted to fit statistical models to these distributions. We applied range normalization where appropriate.

Discretization; Confidence Intervals; T- test; Correlation/Covariance; Multiple Hypothesis Testing; Jaccard Index; Entropy; Mutual Information; Clustering;

Predictive Model

We attempted to create two predictive models. One of them used the lyrics, and features derived from lyrics as well as the genre to predict the popularity of a song, and the other used the lyrical features to predict the genre of the song.

Simple Linear Regression; Multiple Linear Regression; Naïve Bayes Model; Confusion Matrix; Specificity vs. Sensitivity; ROC and AUC; Training-Validation-Testing (Cross Validation); Feature Selection; Dimensionality Reduction (Words – Topics)

Results

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Conclusion

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.