Music Words Slap*

Do Words Make Hits?

Cole Gaito†
Applied Comp Sci Student
University of Colorado Boulder
Littleton, Colorado, USA
cole.gaito@colorado.edu

Problem Statement

When it comes to the business of Music, we see a method of communicating with one another. Whether it's the latest cringe Tic-Tok video theme or yet another one of Taylor Swift's break-up songs… music has the ability to pull people in and capture the audience's attention, heart, emotions, and ultimately wallet. Understanding this continuously growing 28.6 billion dollar[1] industry and what captures the attention of this consumer group is the focus of this research project. I am specifically interested in utilizing the Billboard Hot 100 list as the basis for measure. While this list reflects the interests of the American consumer, this group is often the one driving interest and trends worldwide. Most of major music stars arise from the United States, and the typically the US spends the most on music.[2] It also doesn't hurt that a majority of the record labels and technology platforms also reside in or were developed within the borders of the United States.

In this research project in addition to the Billboard Hot 100, would like to do analysis based on Spotify (given the company's pre-eminence within the streaming market and availability to API calls), and lyrical word analysis. Based of the categories available I seek to derive additional data points and see if there is a correlation between the songs as a whole within their genre, album, time and list on the billboard chart.

While ambiguous, I am also interested in seeing if there is a correlation that can be derived to world events and the Billboard's top 100 chart. While I personally highly double the availability of this to measure, I'm interested to see if the songs, by virtue of their words point to influential events in the world. Again due to the highly subjective nature of lyric interpretation, this is unlikely.

As a point of acknowledgement, this project will not survey the rhythmic form of songs, nor its melody. Only the words of the songs themselves and their associated metadata will be analyzed.

Literature Survey

Looking around the research landscape, there are a couple of previous works that I have come across. In *Analysis of Billboard's Top 100 Songs and Lyrics (1964-2015)* by *Elaine Hsu & Hattie Xu* there is a similar analysis of the ranging and the word counts to study the decrease in one-word, two-word, and three-word phrase complexity. There is a generalized perspective that with each passing decade there is an increase in pessimism that grows out of the lyrics. Ultimately some of their hypothesis' held true – more profanity, less complexity, though their idea of a more pessimistic outlook was incorrect. See the reference for additional details.[3]

There is also a Machine Learning Algorithm that was created by GitHub user *yamnihcg.* Here data was used across Spotify and Billboard data pipelines to model if a song would be a future hit.[4]

A Medium author Josh Viner, also used data from Billboard Hot 100 and Spotify's labels identify artists, songs and popularity.[5]

Ultimately though, I could not find anyone who is specifically looking at the text to see if there is a correlation.

Proposed Work

After having hinted at the work proposed, I'll look at outlining what is necessary for understanding and working through the questions stated above. The following numerical values are based on what was theorized in my previous submission.

1. Data Cleaning – Almost None

Having worked with the data so far, this statement remains mostly accurate. The data coming from highly used datasets provides me with little "cleaning". I have found though that there is roughly a ~2%, totaling to 850 items, with an issue between all the unique items. Whether lyrics are not found, names incorrectly identified, or with odd labels for the set there doesn't seem to be a large issue.

2. Data Incorporation – Significant

I knew that this would be the main bulk, capturing the data. As I am pulling the lyrics from a website, I spent ~2 weeks with my machine constantly pinging Genuis' API for Lyrics data. This was also after I managed to get a script written to scrape the data, and then periodically save it into an excel file. The complexity is not necessarily the issue, rather the amount of time to ping, then compile the file was more then I anticipated.

3. Data Preprocessing – Some

Here I recommended having to perform some data manipulate to extract and understand the lyrical text.

Overall, I expect that the interesting part is still to go. I have generally completed the pull of data from the Genuis API scraping the lyrical data from the internet. I want to incorporate additional data point from Spotify and the associated Metadata around the artist, album and song, this is work yet to go.

In addition to the data itself, deriving the most common word in a song or subset and other statical metrics whether by decade, year, month etc. is also outstanding work. I'm not overly concerned about this part though. There are a variety of tools that have and will simplify the process.

Data Set

The original sources for the data can be located at Kaggle[5] and within the self-maintaining GitHub[7]. The dataset can be found in my public facing GitHub account in an excel file called gaitocole/ Final_Project/ Music_Trends/ *charts.xlsx*[8]. This format is accessible to Pandas, MatplotLib and other coding tools and provides an easy-to-read format that is universally available to both coders and non-coders alike.

Currently there is 3.4 million pieces of data, will additional data to be imported. There are roughly 31.5 thousand unique songs across the dataset starting from August 4th, 1958. As previously mentioned, importing the additional metadata and extracting the metrics is the work to go.

Evaluation Methods

The evaluation methods will be the derivation of the words based in the lyrics, the correlation to the Billboard Hot 100 chart. I assume that this will be along the range of having similarity to the idea of a correlation co-efficient. I would also like to see the demonstration based off a word cloud. I often find myself to be a visual person, so having the visualization is a huge point for me personally.

Tools

The current list of tools are as follows:

GitHub – Data Configuration Management

Git-LFS – Open-Source Large Dataset Git Extension

MS Excel – Basic Data Storage File Handler

Python – programming language v3.13.0

VScode – Integrated Development Environment

Genuis API – Website with API call interface for Lyrics gathering

Spotipy API – Spotify website call interface

Various Python Libraries

Milestones

I assume that this project will continue to be something that evolves through the rest of the semester. I would like to see this project completed in

the next month (by Thanksgiving ’24) as I am going on a scheduled holiday.

REFERENCES

[1] **Statista.** 2023. Global revenue of the music industry from 2002 to 2023. *Statista.* Available: https://www.statista.com/statistics/272305/global-revenue-of-the-music-industry/#:~:text=In%202023%2C%20the%20total%20revenue,compared%20to%20the%20previous%20year.

[2] **Wikipedia.** 2023. List of largest recorded music markets. *Wikipedia, The Free Encyclopedia*. Available: https://en.wikipedia.org/wiki/List_of_largest_recorded_music_markets.

[3] **Brown University Department of Computer Science.** 2011. CS100 Student Projects - Project 11. *Brown University*. Available: https://cs.brown.edu/courses/cs100/students/project11/.

[4] **Yamnihcg.** 2023.Billboard100. *GitHub Repository*. Available: https://github.com/yamnihcg/Billboard100.

[5] **J. D. Viner.** 2023. What makes a hit song? Analyzing data from the Billboard Hot 100 chart. *Medium*. Available: https://joshdviner.medium.com/what-makes-a-hit-song-analyzing-data-from-the-billboard-hot-100-chart-74c1d5ad3fa3.

[6] **D. Dave.** 2023. Billboard - The Hot 100 Songs. *Kaggle Datasets*. Available: https://www.kaggle.com/datasets/dhruvildave/billboard-the-hot-100-songs.

[7] **UT Data.** 2023. rwd-billboard-data. *GitHub Repository*. Available: https://github.com/utdata/rwd-billboard-data.

[8] **C. Gaito.** 2023. Music Trends - Final Project. *GitHub Repository*. Available: https://github.com/gaitocole/Final_Project/tree/main/Music_Trends.