Proposal on Trends and Indicators of Popular Music on Spotify

Project members

Jason, Wang Cheuk Yeung 1155127081

Hans, Nathanael Junoes 1155147304

Felix, Tsui Fan Yau 1155143241

Problem Statement

One of the greatest concerns of music producers is the lack of ideas or directions to create a hit song. Many music producers try to create a popular song by catering to the public's flavor, often affected by following the contemporary public music trend.

This project aims to alleviate the difficulties of music producers who are faced with the dilemma of selecting a trend or style. The concept of our design is to use many popular songs that were released in the past as a reference, and therefore to suggest a popular song and categorize them according to themes that the music producer wishes.

By analyzing varying attributes in our dataset, we expect to find certain factors that correlate more positively in comparison. Utilizing various data science techniques, we desire to identify the trends and indicators of popular music on Spotify and reasonably predict the likelihood a song is to rise in the charts compared to others.

Motivation

In consequence to the global, disastrous impact the pandemic has brought, Spotify has seen a significant rise in users over the last few years as many found themselves spending more time in the comfort of their homes. As much as a 27 percent increase in monthly active users was recorded in early 2021, totalling to 345 million. Subscribers of the premium kind, which account for the majority of the streaming service giant’s revenue, were up 24 percent in the late fourth quarter of 2020 [1].

However, seemingly defying natural sense, Spotify’s streaming service was discovered to be inversely correlated with the increase in active users. On average, audio music consumption decreased by 12.5 percent after the World Health Organization (WHO) declared a pandemic in March 2020 [2]. Owing to this, Spotify suffered a major loss of $838 million in revenue in the first three quarters of 2020. In a time when live concerts and tours are put to an end, it has been brought to our attention that the current situation is detrimental to artists, especially smaller ones, aiming to make a name for themselves.

Thus, this project is aimed to provide one with a deeper understanding into the different metrics that are correlated with the popularity of a song. We believe that looking at the history of hit tracks in Spotify will lend us information of trends in music and how it changes over time. It is in our best hopes that our findings may prove useful to artists who seek recognition, and, if anything else, the reader may gain insight into the inner workings of Spotify’s recommendation system.

Data Source

There is a wealth of music-related data available on Kaggle - a free, online community for data science and machine learning datasets. There already exist datasets collected using the Spotify Web API, listing the top music on the app . For the purpose of this project, we shall use the dataset 'Top 100 tracks of Spotify from 2001-2019' from Kaggle [3].

From 2001-2010, the number of subscribers in Spotify was relatively low. Spotify reached 1 million subscribers in March 2011 [4], so in order to have a reasonably large sample size to determine the popularity of a track, only data from 2011 to 2019 are used.

Data Processing Approaches

The core idea of this project is to investigate the factors that correlate to a track's popularity on Spotify. The aforementioned dataset contains several attributes that describe the property of a song - for example, the danceability, the energy and the loudness of a song. Spotify classifies genres using these attributes [5], so they should prove useful for our purpose.

We propose to first analyze the data year by year with a regression analysis using the attributes as variables. We aim for a best-fitted regression with appropriate weightings for each variable. Thus, we can then visualize the weightings year by year as percentages, and view the trend in significance for different variables across the years. We then do the same for an aggregate of all 9 years and visualize the significance and percentage of each attribute on the popularity of songs based on the ranking of the tracks. Essentially, we can hope to predict the popularity of a new song with this system.

With some of the weightings in mind, we also propose a clustering system to group different songs together via unsupervised learning. Using the different variables, we expect the system to do an automatic grouping of the songs based on their similarities. We expect such a system to provide users with a basic recommendation of other songs based on songs they have listened to.

Implementation / Experimentation

For the regression analysis, we intend to use the sklearn library in Python to do linear regression. We will divide the songs into training and testing data sets - the former to do the regression and the latter to test the accuracy of the regression model. We will split the dataset depending on the outcome of the accuracy itself. As the genre (which is categorical in nature) of a song depends on the other attributes, it will not be included as part of the regression. Hence, all the attributes are quantitative in nature, and there is no need to assign any custom numerical values to any categories. We will visualize the regression with matplotlib in Python and Tableau.

For the clustering system, we will use a simple k-means clustering system for unsupervised learning to calculate the Euclidean distance between each song - calculated using the variables. Similar to the regression model, we will divide the data into training and testing sets to verify the model.

Expected Conclusions

In visualizing and doing regression analyses for songs from 2011-2019, we expect to find a general trend for which variables hold the highest significance on the popularity of a song. For instance, perhaps the energy of a song is shown to be a better indicator of popularity than the key of the track.

There may not be a significant difference in almost a decade, though we expect at least gradual shifts in music taste on the main traits of a song that draw the most audience. In general, we hope that with the ever-changing entertainment landscape, one can extrapolate an overall regression and even predict the music trend in the near future, and perhaps prove itself useful as a general guide for music artists.

In doing clustering of the songs, we aim to classify or group different songs into subgenres, based on their similarities. For instance, the system may group together songs in minor scales with those that are lower in decibels. With this in operation, we may recommend similar songs to any existing user based on their own preferences.

References

[1] Mukherjee, S. (2021, February). *Spotify outlook weakens as pandemic uncertainty persists*. Reuters. Retrieved April 5, 2022, from https://www.reuters.com/article/us-spotify-tech-results-idUSKBN2A31JB

[2] Philiptrapp. (2022, February). *Why streaming on Spotify actually declined during the pandemic*. Loudwire. Retrieved April 5, 2022, from https://loudwire.com/spotify-streaming-declined-during-pandemic-study/

[3] Delaney. (2021, April). *Spotify Top 100 Tracks (2001-2019)*, Version 1. Retrieved April 3, 2022 from https://www.kaggle.com/datasets/delaneyisabella/spotify-top-100-tracks-20012019.

[4] BBC News. (2011, March). *Spotify hits milestone with 1 million subscribers*. Retrieved April 3, 2022, from https://www.bbc.com/news/business-12676327.

[5] Patch, N. (2016, January). *Meet the man classifying every genre of music on Spotify - all 1,387 of them*. thestar.com. Retrieved April 3, 2022, from https://www.thestar.com/entertainment/2016/01/14/meet-the-man-classifying-every-genre-of-music-on-spotify-all-1387-of-them.