Predicting Song Popularity

Team - Name: NA

Shailesh Sridhar
Department of Computer Science
PES University
Bangalore, India
shailesh.sridhar@gmail.com

Shashank Prabhakar
Department of Computer Science
PES University
Bangalore, India
shashank.prabhakar@yahoo.com

Shrey Tiwari
Department of Computer Science
PES University
Bangalore, India
shreymt@gmail.com

***Abstract*—Exploration and analysis of songs is a task that would interest and excite any music lover. In this research we wish to study how different parameters affect the popularity of songs among different user groups. The aim is to be able to extract information from multiple sources and use it to be able gain insights on the relationships between the song popularity, user groups and the numerous other factors, according to us, that could play an important role in determining the performance of the song in the market.**

I. Introduction

Since the advent of digital music and online availability of songs, there has been a lot of change in the way we listen to and consume audio. There are new artists coming up with every passing day, adding to the ever growing collection of audio tracks. As users, it is becoming harder and harder for us to find songs of our liking. Due to the large variety of songs and various new age features, present day recommendation systems are suffering a loss in their accuracies.

Gaining insights into what makes a song popular and what attributes please what kinds of target audiences would help the music industry. In Spite of the diversity in the field of Music Information Retrieval, there has not be much research on the task of predicting a song's popularity by taking into account the diversity of the user groups, the song's features and the different markets whose needs the audio track might cater to. We wish to build models (maybe neural nets) that can account for these factors and use the rich data available to be able to accurately predict the popularity of various songs.

II. Datasets

[a] Million Song Dataset

The huge Million Song Dataset is provided for free by LabROSA, Columbia University. The Million Song Dataset is a cluster of complementary datasets contributed by the community for academic purposes and research in certain aspects of music data. It contains information about songs from 1922 to 2011.

[b] Taste Profile Subset

The dataset contains real user-play counts from undisclosed partners, all songs already matched to the Million Song Dataset

[c] musiXmatch Dataset

The musiXmatch dataset contains lyrics to many songs in the Million Song Dataset. The lyrics in this dataset are directly associated with the MSD tracks and are in the form of bags-of-words.

III. Problem Statement

The aim of this project is to predict song popularity.

In the process, we will analyse how the features of popular songs have evolved over the years and vary across locations, aiming to determine the features that are the most important in determining the popularity of a song.

By the end of this project, we wish to use this information to come up with a robust method or model to predict how popular a song is.

IV. Past Research

A lot of work has been done on predicting the popularity of songs. Most of these publications, for example, [5] classify a song as a hit or not a hit based on whether or not it has appeared on the BIllboard Top 100 charts or any other regional music hits chart, or utilise the song's ranking on these charts as a measure of its popularity.

In 2011, Borg and Hokkanen investigated if they could predict the popularity of a song based on its audio features and Youtube view counts. The features for audio tracks were obtained from The Echo Nest. For this task, they used a number of Support Vector Machines were, and the achieved results were very modest. The Support Vector Machines, regardless of feature choice and parameters, never achieved more than 53% accuracy. They draw the conclusion that audio features alone do not seem to be good predictors of what makes a song popular.For instance, for two given songs, even though song features were similar (examined from the data extracted from audio), there seemed to be a huge difference in the view counts. They suggest that popularity is likely driven by social forces.**[6]**

A study from 2016 conducted by Pham et al., at Stanford University, evaluated different machine learning algorithms and their ability to predict popularity of music tracks using the Million songs Dataset.Using Support Vector Machines, Neural Networks, and Logistic Regression etc. In their research, they used a subset of 2717 tracks and a hit was defined as a song with a high hotttness value(A field in the MSD dataset). The results of the research showed that all models performed with similar accuracy, with a values ranging from 0.70 to 0.85 **[8]**

As recently as April 2018, J.Berger and G.Packard attempted to analyse the role that lyrics play in determining the popularity of songs. Using natural language processing of thousands of songs, they examined the relationship between lyrical differentiation (i.e., atypicality) and song popularity. The 'normal' lyrics of a particular genre were determined and compared with the lyrics of hit songs belonging to that genre. The result was that hit songs seemed to have lyrics that were very atypical for their corresponding genres. **[9]**

There has been very little research on utilising both Lyrics and audio features to predict song success.

IV. Data Processing and Visualisation

A. Data

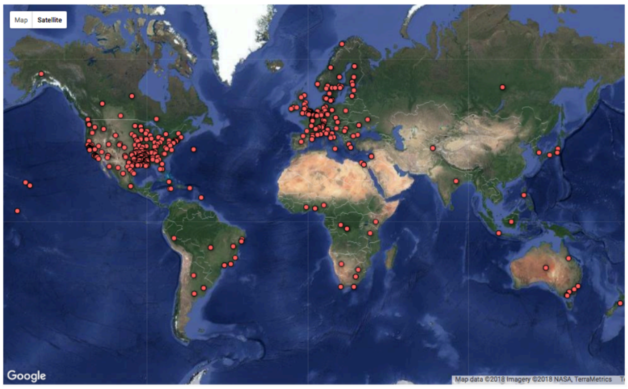
Data in the Million Song Dataset is in HDF5 format. For Visualisation, we use the already available 10000 song subset in CSV format. For other analyses, we convert the HDF5 files to the appropriate format, like data frames in Python arrays. The large dataset in HDF5 format does contain duplicates. However, the CSV subsets used for visualisation and lyrics analysis do not.

The lyrics come in bag-of-words format: each track is described as the word-counts for a dictionary of the top 5,000 words across the set. The dataset comes in two text files, describing training and test sets.

We make use of Python's built-in library, *h5py*, for reading and processing HDF5 files. Columbia University has provided getter functions which makes it easy to extract the desired attributes from .h5 files. The data extracted is stored in arrays, and finally combined into dataframes.

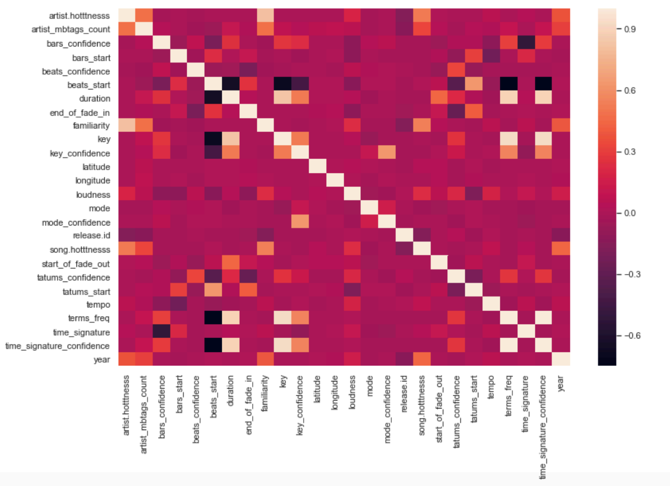
B. Plots

- Song locations across the globe



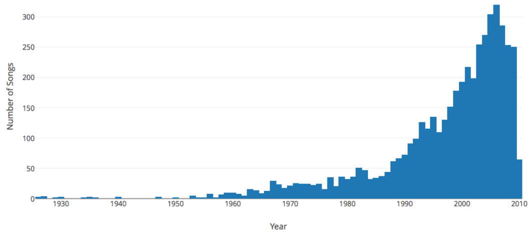
Most songs come from Europe and North America.

- Correlation Plot



High correlation between *familiarity* (LabROSA's measure of popularity of a song) and *artist_hotttness* (measure of popularity of artist), *key* and *terms_freq.*

- Number of Songs vs Year



Number of songs released every year sees a general increase over the year, with the most being in 2006.

V. Approach

As there is an apparent lack of in-depth research with respect to this approach, we will attempt to use both lyrics and audio features to predict song popularity.

While most papers use billboard ranks to predict song success, we will also factor in the listening history of users as provided in dataset [b], with the assumption that a greater total number of times a song has been listened to implies that is more popular.
This will make the popularity of a song a value which lies within a continuous range and allow us to approach the problem as a regression problem as well and not just a classification problem(Hit/Non-Hit) .

We hope these unique facets to our approach will help improve the accuracy of our predictions.

VI. Citations

[1] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. The million song dataset.
In Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR), pages 591–596,
2011.
[2] J Serrà, Á Corral, M Boguñá, M Haro, JL Arcos.
Measuring the evolution of contemporary western popular music.
Scientific Reports 2, 521, 2012
[3] Mauch M, MacCallum RM, Levy M, Leroi AM. 2015. The evolution of popular music: USA
1960–2010. R. Soc. open. sci. 2, 150081
[4] Interiano M, Kazemi K, Wang L, Yang J, Yu Z, Komarova NL.
Musical trends and predictability of success in contemporary songs in and out of the top
Charts. R Soc Open Sci. 2018 May
[5] Mohamed Nasreldin, Stephen Ma, Eric Dailey, Phuc Dang. "Song Popularity Predictor",
Towards Data Science(blog), May 9 2018, https://towardsdatascience.com/song-popularity-predictor-1ef69735e380
[6] Borg, N. & Hokkanen, G. (2011), 'What makes for a hit pop song? What makes for a pop song?', Unpublished thesis, Stanford University, California, USA .
[7] Ni, Y., Santos-Rodriguez, R., Mcvicar, M. & De Bie, T. (2011), 'Hit song science once again a science?'.
[8] Pham, J., Kyauk, E. & Park, E. (2016), 'Predicting song popularity', nd): n. pag. Web 26.
[9] J. Berger, G. Packard(2018) Are Atypical Things More Popular?https://doi.org/10.1177/0956797618759465