Project Proposal: CSE 519 Data Science Fundamentals
Do Popular Songs Endure?

Objective

The objective of this project is to predict the popularity of english songs with time i.e. how popular or not-so-popular songs when originally released, are doing today? We will design a model to predict the current popularity of an old song by analyzing factors affecting it. While such a model will help us in analyzing the song's hits trend over time, it would also be able to predict how a newly released song will perform in the future.

Background Research

Over the last few decades, the music industry has witnessed exponential growth in terms of volume of business, manpower employed, number of albums produced each year and also the global reach. The music industry in the United States alone generated a revenue of $8.72 billion in 2017. With cheaper mass storage, better compression algorithms, and growing internet connectivity, many companies like Spotify, Apple music, Pandora, etc have started generating huge revenues in the songs industry. Thus it will be of immense interest to know what makes a song a popular. Is it the recent technical advancements or the singer's popularity? Does it depend on songs acoustic or do the lyrics generate revenues? There have been numerous studies on analyzing what makes a song popular and how their popularity has changed over time. Studies like "Top Songs in the Sixties" by Richard R Cole. have done a content analysis of popular lyrics. Similar studies have used lyrics to predict the popularity of a song. 'Automatic Prediction of Hit Songs' by Ruth Dhanaraj, Beth Logan dealt with developing classifiers to find likely hit songs. Similar previous studies have analyzed different features like time of release, major product labels, etc to predict a song's long-term success.

Dataset

We have taken data from various sources to get the most informative training sample for building an accurate model. Those data sources are:

a. **Million Song Dataset** - This dataset is provided by Columbia University and is a collection of features and metadata for a million of popular music tracks. This dataset contains data from different communities like Last.fm dataset, thisismyjam-to-MSD mapping musiXmatch dataset etc.

b. **FMA : A dataset for Music Analysis** - This dataset contains per track metadata such as ID, title, artists, genres, tags, play counts etc.

c. **Track Popularity Dataset** - This dataset provides different sources of popularity definition ranging from 2004 to 2014. It contains 23,385 tracks of which 9.193 are designated as popular by appearing in any of the popularity charts, while remaining are not designated as popular by any popularity scoring sources.

d. **Billboard** - The HOT 100 Songs (1958-2017) : By scraping the billboard's website, we obtained all the songs that appeared on charts from 1958 to 2017. The data for each year is scraped, for e.g. below is the data for year 2007 obtained from Billboard.com

	year	title	name
0	2007	Irreplaceable	Beyonce
1	2007	Umbrella	Rihanna Featuring Jay-Z
2	2007	The Sweet Escape	Gwen Stefani Featuring Akon
3	2007	Big Girls Don't Cry	Fergie
4	2007	Buy U A Drank (Shawty Snappin')	T-Pain Featuring Yung Joc
5	2007	Before He Cheats	Carrie Underwood
6	2007	Hey There Delilah	Plain White T's
7	2007	I Wanna Love You	Akon Featuring Snoop Dogg
8	2007	Say It Right	Nelly Furtado
9	2007	Glamorous	Fergie Featuring Ludacris

e. **Youtube API** - To get the view count of the songs to measure as one of the popularity factors.

f. **Grammy.com** - To get the list of artists winning grammy in particular year which will affect the popularity of the song.

Approach

Goal

We intend to create a model **to predict the current popularity of a song which appeared at position p on the charts w weeks ago?**

Data Collection

We have identified scraping, API's and manual data download as 3 ways to collect the datasets mentioned above. We plan to use Python for all our scripting and analysis tasks. We will collect the song's data from the 1950's until 2018.

Data Cleaning

Our group went through several of the datasets mentioned above. In particular, we identified that Million Songs Dataset is a large dataset containing many helpful features for our analysis. However, the format and size of the data given by this dataset are not very helpful. We plan to convert the h5 format file for each song into CSV format. Out of 280 GB of data, we plan to pick 10,000 songs which will have all the features. We will clean all the null values and keep a varied distribution of songs over time. Further, many songs get recorded by multiple groups, and many recordings appear repeated on different albums after this release. Thus we need to make sure to link the right recording of a song to its Billboard rank.

Feature Selection

In order to achieve the above goal, we grouped the features into 2 categories.

1. Dynamic/ Time-Varying Features

These features are observed in last w weeks and given a relative value for the current week.

Feature	Observations
Song's/Album's Popularity	We plan to analyze 3 sources for this feature. 1. Popularity on Google Trend 2. Wikipedia Page Visits/ Edits 3. Youtube View Counts 4. Spotify downloads 5. Radio Airplay
Singer's/Band's Popularity	First 2 sources as above
Composer's Popularity	First 2 sources as above
Lyricist's Popularity	First 2 sources as above
Any Awards Won	Billboard Music Award, Grammy Award, Glenn Gould Prize, Pulitzer Prize for Music, etc
Any Cover Released	Ex. Elvis Presley's version of Carl Perkins' original "Blue Suede Shoes
Related Incidents	Extract Google News to look for incidents like Singer's Death or Band separation

2. Static Features

Audio Analysis Features
Tempo, duration, mode, loudness, key, time signature, section start, length

Song Related Features

Feature	Observations
Billboard Ranking p weeks ago	
Sales in the first few weeks of release	
Release Year	This gives how old the song is
Any Festive Theme	Christmas/Halloween/New year
Genre	We will rank the genres based on their current popularity and give the corresponding rank to this song's genre

Lyrics	Meaningful/Repetitive
Record Label	Individual/ Major Record Label (like Universal Music Group, Sony Music Group, Warner Bros Music)
Is Cover	True/False

Brand Related Variables

Feature	Observations
Distributor's/Label's Market Share	High/Medium/Low
Publishing Company's Popularity	High/Medium/Low

Miscellaneous Variables
Artist's Location

Model

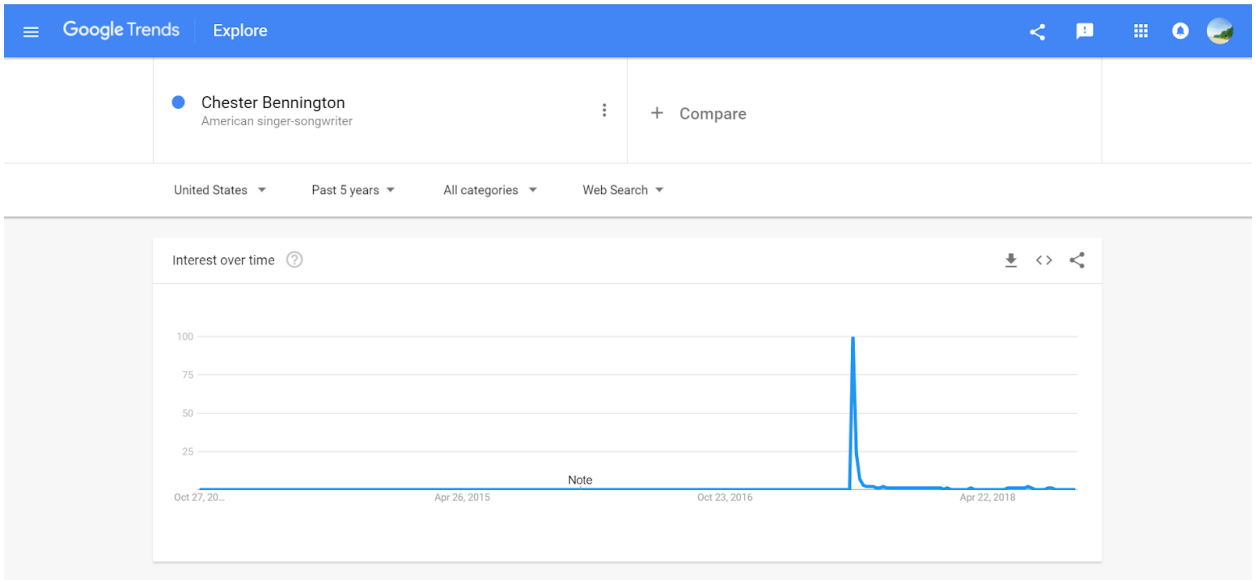
We plan to create a regression model as our baseline model, which will predict the current chart's rank of a song which appeared at position p on the charts w weeks ago.

Depending on whether the dataset is balanced or imbalanced, we will be testing different machine learning algorithms like Simple Linear Regression, SVM, Neural Networks, Random Forest Trees, Gradient Boost Models, etc. We will also have to tune the hyperparameters using K-Fold cross-validation.

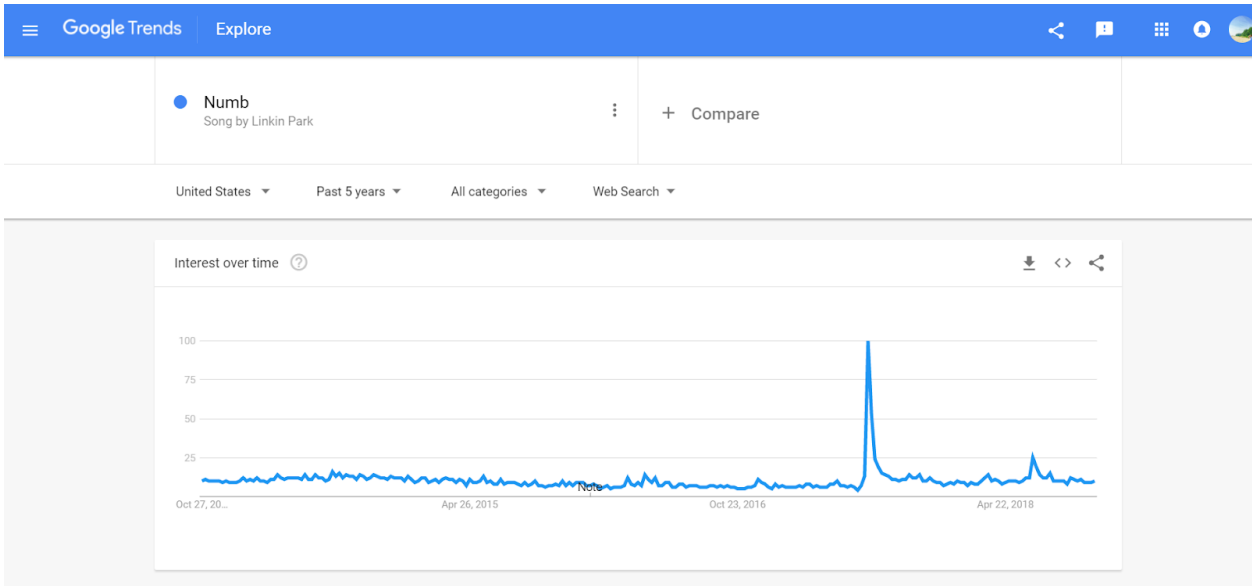
Analysis
Out of all the features mentioned above, we wish to find a subset of features which significantly affect a songs hit over time. Such a set of features would be quite helpful for music companies as this would help them in producing music which lasts long in the industry. We will be ranking major music groups like Beatles, Grand Funk, the Beach Boys, the Bee Gees, Kiss, Sonny and Cher, Michael Jackson, Britney Spears across different time periods. We will also use clustering techniques to analyze songs which our model predicts as outliers. We are seeking to prepare illustrative charts and visualizations to answer the following questions.

1. How a song is considered hit/popular?
2. Do the rank of songs improve over a period of time?

3. Is there a trend in music chart rankings?
4. Do early/late songs of bands tend to over/underperform? Does a popular group's lesser hits retain more/less popular than their top hits?
5. How have different genres performed? Do movie themes or dance songs behave differently?
6. Does the change in artist's fame over time impacts songs endurance?



The trend for singer popularity - which has suddenly increased in the mid-year 2017



Song Numb by Linkin Park (Vocalist - Chester) Trend

From the above two trend analysis, we can see that the singer's popularity affects the song's popularity.

7. How is over/under-performance a function of the number of charted songs they have? Do one-hit wonders retain more/less popularity than expected by their chart positions? Does this differ depending on how high their charted songs were?

Evaluation

The current actual ranking of the song on Billboard, when compared with the predicted rank, will give us the accuracy of our model. Further, we will perform sniff tests on our computational results. The test will make sure that our model has learned over different time periods and performs well for most of them.

What's Next?

We will continue doing a literature review on this topic. Next, we plan to create a unified database for all our features. We want to create a baseline model as quickly as possible and then improvise on it by further data cleaning, feature selection, and other dataset explorations.

References

1. Million Song Dataset: https://labrosa.ee.columbia.edu/millionsong/
2. Track Popularity Dataset: http://mir.ilsp.gr/track_popularity.html , https://link.springer.com/chapter/10.1007/978-3-319-44944-9_50
3. Youtube API: https://developers.google.com/apis-explorer/#p/youtube/v3/
4. Google Trends: https://trends.google.com/trends/?geo=US
5. Wikipedia: https://en.wikipedia.org/wiki/American_popular_music
6. Top Songs in the Sixties: http://journals.sagepub.com/doi/abs/10.1177/000276427101400311?journalCode=absb
7. Automatic Prediction of Hit Songs https://pdfs.semanticscholar.org/2de2/34f32c268879e0aa331f286f50c3426837ad.pdf
8. Song Popularity Predictor https://towardsdatascience.com/song-popularity-predictor-1ef69735e380