Music discovery through song lyrics

Project LYRICAL - Mateus Surrage Reis, Indrek Ardel

Introduction

Music is an area of wide human interest, and as such, generates a proportionally large amount of raw data in today's connected world. The abundance of music often makes it challenging to discover or suggest new songs that a person might enjoy. In this project, we sought to find how songs are connected with each other through their lyrics and how these methods could be used to aid finding similar music.

Data

We used various sources for obtaining lyrics. Datasets were discovered through Kaggle. All data and scripts used to process data is available at our project repository[1].

1. Billboard Year-End Hot 100 charts from 1959-2018
   This is an enhanced version of the dataset provided by a Github user Kaylin Pavlik[2], which combines data from Wikipedia, lyricsmode.com, genius.com, songlyrics.com and metrolyrics.com.
2. 500 000 songs from lyrics.com
   We re-scraped the website using the sources provided on its corresponding Kaggle page in hopes of getting more and most recent songs, however out of 1.3 million records we still found only around 500 000 to have unique lyrics.
3. 380 000+ lyrics from MetroLyrics
   This dataset was sourced from Kaggle, it is worth mentioning that this dataset includes the year of release and genre in addition to the typical artist, title and lyrics of other datasets.
4. 55 000+ songs from LyricsFreak
   While this dataset was considered originally, we quickly discovered that it did not contain many new songs compared to what we already had in datasets 2 and 3, which is why we decided to discard this dataset.

Challenges

Our datasets were drawn from a variety of online sources, and those sources in turn drew from mostly-unfiltered, minimally curated crowdsourced input. This, along with other factors inherent to using online sources, song lyrics, and text in general as data brought with it a unique set of challenges when it came to data preparation.

First and foremost is simple inaccuracy, not in the lyrics, but in attached information such as year of release and genre, after all, just lyrics by themselves are typically not that interesting. To this issue there's no good solution but to find better databases, or scrape and compose some yourself. Due to the lacking quality of year information, we had to abandon any plans of providing analysis that would have compared songs by year.

We also had to take into consideration that songs come in many languages, and this was in fact the most computationally expensive section of pre-processing, although we relied on external libraries for the task of removing non-english songs.

Finally, there are minor, common data-cleaning trials such as: duplicate entries (many artists are attributed to same song even if they aren't the original author), encoding issues, and unwanted artifacts in data. As an example, lyrics text often contains indicators for chorus, verses, or how many times a part of the song is sung. These were filtered as they were noticed the normal way.

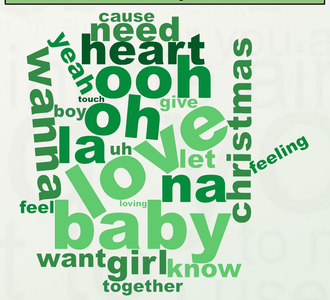
Country

Old	Lonesome	Cowboy	Tennessee	Country
Man	Road	Boots	Memphis	Boy
School	Whistle	Cocaine	Nashville	Song
Years	Sound	Ride	Waltz	Little
Friend	Town	Hat	Hills	Road
Time	Day	Songs	Rocky	Side

Jazz

Thrill	Dear	Lovely	Romance	Blue
Gone	God	Day	Bad	Skies
Hope	Hold	Weather	True	Deep
Worth	Mother	Lady	Want	Eyes
Hold	One	Face	Find	White
Never	Life	Sweet	New	Moon
Found	Friend	Night	Sweet	Feeling

Pop



Baby	Heart	Oh	Girl	Feeling
Please	Break	God	Good	Inside
Tell	Take	Please	Pretty	Right
Let's	Inside	Tonight	Every	Lonely
Tonight	Soul	Never	Met	Good

How did we reach these words?

The words in the word clouds, as well as in the highlighted rows were the ones with highest *log likelihood* of appearing in their respective genres, calculated with the following formula:

$$LL_{word} = (O_g * \ln \frac{O_g}{E_g}) + (O_o * \ln \frac{O_o}{E_o})$$

Where:

- O_g is the amount of occurences of that word in that genre's lyrics,
- O_o the amount of occurrences in all other genres added up,
- E_g the expected amount of occurrences of that word in that genre, if it had been uniformly distributed across the lyrics, and
- E_o the expected occurrences in all other genres, as above.

Finally, the log likelihood was negated if E_g was greater than O_g.

The remaining words in the tables were discovered by going through the lyrics again and extracting all words within two words of the highest-log-likelihood one, then ranking them by absolute number of appearances. The contents of the tables were then manually selected from amongst the top-ranked words. That is to say, in the columns are words that appeared frequently together with the entry at the top of the column.

References

1. Project LYRICAL repository
   https://github.com/Ingramz/ds-2018-project-lyrical
2. Kaylin Pavlik - 50 Years of Pop Music Lyrics, Billboard Year-End Hot 100 charts dataset
   https://github.com/walkerkq/musiclyrics
3. Simon White - How to Strike a Match
   http://www.catalysoft.com/articles/StrikeAMatch.html

Chart size

Similarity range

Show most similar song

N-gram matches

Ranking

- Similarity %
- N-grams matched
- Average n-gram ranking

#	Title	Artist	N-grams matched
1	The Easy Way	Westlife	8990
2	Need Me	Mashd N Kutcher	8987
3	If I	Jesse Powell	8976
4	Dangerously in Love	Beyoncé	8962
5	Do You Wanna Rock	N-Trance	8956
6	Loverboy	Twenty II	8954
7	Baby	She & Him	8950
8	Shiver	Coldplay	8949
9	Blue Jeans	Yasmeen	8945
10	Am I Losing You	Chanté Moore	8931
11	La La La La La	Marsha Ambrosius	8930
12	Don't Wanna Lose You	Lionel Richie	8922
13	Slow and Easy [All Night Mix]	Zapp	8922
14	Shivers	Rick Astley	8921
15	Say You Will	Brandy	8920
16	Come to Me	Tommy James	8917
17	Don't Let Me Be the Last to Know [Hex Hector Radio Mix]	Britney Spears	8894
18	Don't Leave Me This Way	Jimmy Page	8889
19	I Try	Will Downing	8882
20	April Showers	Dru Hill	8878

Finding similar songs based on word sequences

In addition to *log likelihood*, another method of discovering songs with similar themes is through matching word sequences found in lyrics. If songs share a significant amount of word sequences, they can be thought to be similar.

The word sequences used in this project are simple word n-grams of size 3 in order to strike a balance between the amount of tokens generated while still having the ability to convey some meaning for songs written in English.

Songs in dataset 1 (Billboard Year-End Hot 100 charts from 1959-2018) were tokenized and paired up with the rank of the song on the year the song made it to charts. Then songs from dataset 2 were tokenized in a similar fashion, but only information about tokens matched with dataset 1 was retained. This helped us to create 3 different ways of creating a ranking of songs of our own:

- Highest similarity % when compared to to single song based on Sørensen–Dice coefficient[3]

$$similarity(s_1, s_2) = \frac{2 \cdot |triples(s_1) \cap triples(s_2)|}{|triples(s_1)| + |triples(s_2)|}$$

- Number of 3-grams matched (each 3-gram from a single song adds 1 to count)
- Average of rankings gathered for each n-gram

Metal

Eternal	Evil	Flesh	Shall	Darkness
Life	Woman	Bone	Never	Light
Night	Good	Human	Come	Comes
Flame	Man	Made	Rise	See
Fire	See	Like	See	Falls
Light	One	Burning	Find	Let
Rest	Speak	Weak	Fall	Night

Hip-hop

N####a	Wit'	F##k	B####es	Ass
Young	Rock	Around	Bad	Shake
Real	Lean	Shut	Bottles	Big
Lil	Hit	Police	Bow	Fat
Rich	Ride	World	Aint	Punk
Dope	Blow	Cant	Fake	Kick

Make your own playlist

In order to explore the large number of songs and in which way they are similar to the Billboard Year-End Hot 100 chart songs, we crafted a web application which you are encouraged to try out by scanning the QR code found on the poster.

By adjusting the sliders and switching between ranking metric it is possible to create a wide variety of different playlists. The default settings should provide a diverse enough playlist that contains songs not found in Billboard's charts.

As an unintended feature this application can serve as a plagiarism detector. Can you guess how?

Clicking on a table row will take you to the heat map of that song.

The database contains about 500 000 songs which our data cleaning process deemed to be in English and also unique.

Hot in Here

by Danny Jacobs

hot in so hot in here so hot in hot oh with a little bit of uh uh and a little bit of uh uh just a little bit of just a little bit of just a little bit of just a little bit of i was like good gracious ass is bodacious flirtatious trying to show faces im waiting for the right time to shoot my steez you know waiting for the right time to flash them kis then im leaving please believing oh me and the rest of my heathens check it got it locked at the top of the fo seasons penthouse roof top birds i feeding no deceiving nothing up my sleeve and no teasing i need you to get up up on the dance floor give that man what he asking for cause i feel like busting loose and i feel like touching you uh uh and cant nobody stop the juice so baby tell me whats the use i said its getting hot in here so hot so take off all your clothes i am getting so hot i wanna take my clothes off its getting hot in here so hot so take off all your clothes i am getting so hot i wanna take my clothes off uh uh uh let it hang all out why you at the bar if you aint popping the bottles cmon what good is all the fame if you aint fucking the models i see you driving sports cars aint hitting the throttle and ill be down to do a hundred top down and goggles get off the freeway exit 106 and park ed it ash tray flip gate time to spark it gucci collar for dollar got out and walked it i spit game cause baby i cant talk it warm sweating its hot up in this joint vokal tank top on at this point you with a winner so baby you cant loose i got secrets cant leave cancun so take it off like your home alone you know dance in front your mirror while youre on the phone checking your reflection and telling your best friend like girl i think my butt getting big let it hang all out mix a little bit of with a little bit of let it just fall out give a little bit of with a little bit of let it hang all out with a little bit of and a sprinkle of that let it just fall out i like it when ya girl baby make it stop pacing time wasting i gotta a friend with a pole in the basement what im just kidding like jason oh unless you gon do it extra extra eh spread the news check it nelly took a trip from the luna to neptunes came back with something thick and it fitting in sasoons say she got a thing about cutting in restrooms oh let it hang all out mix a little bit of with a little bit of let it just fall out give a little bit of with a little bit of let it hang all out with a little bit of and a sprinkle of that let it just fall out i like it when ya girl baby make it

Getting hot in here?

To visualize how songs make their way to our playlists, we also provide a heat map of 3-grams for all of the songs in our database.

In the example provided, hotter areas indicate that the 3-gram is present in a greater amount of Billboard chart songs.

It should be also pointed out that the original author of the song is rapper Nelly, which shows the weakness in our deduplication strategy, which had no way of knowing the original Author.