Trends in Lyrics of Popular songs, 1959-2017

Colin Allen*, Simon Caton†, School of Computing, National College of Ireland, Dublin
colin.allen@student.ncirl.ie, simon.caton@ncirl.ie

Abstract—**This paper presents the results of a content analysis of the words used in song lyrics, of the top ten songs of the year for every year based on the billboard top 100 of the week. The songs that appeared in the top ten of the billboard top 100 of the week most consistently over the course of the year are the songs that have their content analysed for that year.**

The challenge of this analysis was the gathering of data by filtering a dataset of the billboard 100 songs of the week for every year since 1959 up until 2017, through an API that held the lyric information. Each song was requested from the API individually and the lyrics of those songs were stored and fed back into the analysis system.

Once the data had been gathered I had a dataset of the most popular songs and the performers of the last 50 years and the lyrics to the years most popular songs. The results of this paper show that the words used in the most popular songs have remained consistent throughout the past 50 years, although the usage of certain words rise and fall throughout the years, their rank as one of the top ten used words remains. The results also show that songs remain among the top ten songs of the week for more weeks, and that the popularity of the performers of those songs lasts for longer now than it ever has before.

I. Introduction

There are many reasons as to why society enjoys and listens to music. Throughout history there has always been popularity around certain types of music, with different styles becoming more popular than others and trends growing and then dying out.

This report is an analysis of the word usage of the most popular songs of the last 50 years as an attempt to observe any change in the usage of words in lyrics over time. The initial objective is that over the last 50 years and with the passing of each generation that the usage of words song writers use for their lyrics will change, as different genres of music become more popular, certain words become popular and reduce in popularity as the next wave of music gradually takes over from the previous.

Another objective is that the popularity of individual songs will remain a constant throughout time. The assumption that the popularity of any given song can only remain in the public conscious for a given amount of time, before the public moves along to the next popular song and therefore we should not expect a significant change in how long a popular song remains in the top ten songs of the week, over the given time span.

Similarly, an objective of how popular an individual performer can become, and how long they can stay popular compared to their counterparts from other generations.

Note that popularity is not a sign of how long any given song stays at the ultimate No.1 spot in the charts, but how long it remains in the top 10 each week is representative of how long the song is still popular among the public.

II. Literature Review

For the literature review for this paper, papers were studied that have attempted similar analysis or have used the same data. It was found that generally they have a more specific agenda and a reduced scope.

The first paper that relates to mine is Innovation and Diversity in popular music industry, 1969 to 1990. [1] In this paper Paul D. Lopes conducts an analysis of the Popular music industry from the 70's and 80's to try and understand the United States Music industry and how it effects Innovation and Diversity in popular music. The proposed question of the paper is that an increased high market concentration should lead to homogeneity and standardization among the music that reaches the top of the billboard 100 charts.

The report finds that major record companies employ an open system of development and production that incorporates innovation and diversity as part of their strategy to remain in control of the market. As the music industry goes through different trends every generation it is essential that major record companies incorporate some form of innovation and diversity into their processes to remain in control of the industry as a whole. [1]

This open system of production had actually led to an increase in market concentration from 1969 to 1990. Compared to a closed system of production for ”top hits” during the early 1940s and 1950s, this open system of production had actually increased the oligopolization [1] of the music industry by major record labels. Meaning that the record labels now had more control over what music gets into the charts, although the open system meaning that more innovation and diversity was prioritized compared to before. [1]

This is relevant to this paper as it concludes that although major record labels control the market, that innovation and diversity still play a large role in the success of music in the industry, and therefore the result of the analysis should still find a variety in word usage for lyrics, and a variety in song and performer type.

Another paper that is relevant to this paper is Expressions of love, sex and hurt in popular songs: a content analysis of all-time greatest hits. This paper uses the same data set as this report but focuses on a different part of the data, as it explores the expression of love and sexual imagery within the lyrics and also examines the proportion of artists in the top 100 songs based on Gender and Race, and then breaks down those proportions to analyse which include the topic of love, sex and hurt within their lyrics over time, from 1958 to 1998. [2]

The results of the paper show that references to love by female artists had decreased over the given time frame although the proportion of female artists within the top 100 songs over had increased. Lyrics performed by women contained more love words than lyrics performed by men, and men made more references to sex than women. African-American artists in the top 100 songs had increased. [2]

This is relevant to this paper as it studies the same data, and therefore should be a similar analysis of lyrics and word usage. Although focused on the specific themes of Love, Hurt, and Sex. The results of this paper imply that the usage of words with those themes, specifically the word ”love” should be of high priority for my analysis. [2]

III. Data

This section of the paper will delineate the data that was chosen, why it was chosen and how that data was sources. The data for this report stems from an initial dataset, and the following dataset was obtained using data from the inital dataset.

A. *Billboard Hot 100 1958-2017*

The first dataset chosen was Billboard Hot 100 1958-2017. This dataset was downloaded as a .CSV file from the website data.world. [3]

The Billboard charts are a music industry standard ranking issued weekly by Billboard magazine. The chart ranking is based on sales of both physical and digital media, radio play and online streaming. There are other sources of chart information, however Billboard are considered the most accurate and that is why I chose this data set for my analysis. [3]

The initial dataset that was downloaded had Ten columns, (Billboard Chart URL, WeekID, Week position, Song name, Performer name, SongID - Concatenation of song and performer, Current week on chart, Instance, Previous week position, Peak Position , Weeks on Chart) [3] But ultimately I cleansed it down to the essential data that was needed.

WeekID	Week Position	Song	Performer
DATE	INTEGER	STRING	STRING
1990-05-05	84	Lonely Boy	Paul Anka

TABLE I: Billboard Hot 100 Dataset (Cleansed)

After removing the columns that weren’t needed, I cleaned the remaining columns. I checked for null values and didn’t find any. I cleaned up the WeekID column as I didnt need the week information only the year.

From there I cleansed the Performer column to remove anything extra that was alongside the name of the performer, such as ”And His Orchestra”. I removed ”Featuring” and anything after it, as I needed a clean value of song and the performer to enter my API. Afterwards, I split the data set into sections, one for each year. Each row in these sections contain the Year, the song name, the position and the performer.

These sections were then run through a sequence of Map Reduce Patterns. [4] The first pattern Isolated the songs that were at position 1-10 throughout the year. If a song was in any position between 1 and 10 for the year, it would be selected. After this selection process, those songs would be placed into another Map Reduce Pattern. [4] An alteration of the Word Count Map Reduce Pattern, ”Song count” would count the song and how often it came up in the section. The output of these Map Reduce Patterns resulted in my second dataset, named popularity.

B. *Popularity*

The Dataset named Popularity has four columns, Year, Song and Performer are all self explanatory. Popularity is the result of the Map Reduce Pattern Song Count, this is how many times the Song was found within the section. The example below, means that Y.M.C.A by Village People was in the top 10 songs of the week, 10 times in 1979. [4]

Year	Song	Performer	Popularity
DATE	STRING	STRING	INTEGER
1979	Y.M.C.A	Village People	10

TABLE II: Popularity Dataset

C. *Lyrics*

The Lyrics dataset is a result of the popularity Dataset being passed through to the Genius API [5] [6] to store the Lyrics of each song. It is essentially Identical to the Popularity dataset, although including another column called Lyrics.

Year	Song	Performer	Lyrics
DATE	STRING	STRING	String
1960	I’m Sorry	Brenda Lee	I’m sorry..

TABLE III: Lyrics Dataset

These datasets were then combined to make my Second dataset, which is named PopLyrics.

D. *Lyricswc*

The final dataset for my analysis is the Lyricswc Dataset. This is the result of the Lyrics from the PopLyrics dataset being put through a final sequence of Map Reduce Patterns. [4]

First I take the Lyrics column from the PopLyrics dataset and split it by year, each year is then written to an output file with the title being the year. It contains the lyrics of the top 10 songs for that year in a single file. 1959, 1960, 1961..etc.

These files are then ran through a preprocessing script that eliminates any noise from the lyrics. This means removing any Non ASCII characters such as punctuation and commas, removing numbers, setting everything to lowercase, and removing any stop words that would not be valuable to the analysis, such as ”The”, ”And”, ”But”, etc. I also ensure that certain words are filtered out that appear often in the Lyrics column but aren’t relevant. When the lyrics are retrieved from the Genius API [6], they are sectioned. So things like [Chorus], [Verse 1], [Verse 2] are found heavily within the lyrics and definitely need to be removed.

From there the output of the preprocessing script is run through the Word Count Map Reduce Pattern, and the output of that is run through the Top 10 Map Reduce Pattern, which results in the top 10 words used per year.

Word	Amount	Year
STRING	INTEGER	DATE
Love	25	1969
Baby	20	1969

TABLE IV: Lyricswc Dataset

IV. Methodology

This section describes the use of KDD in the analysis found in the report. [7]

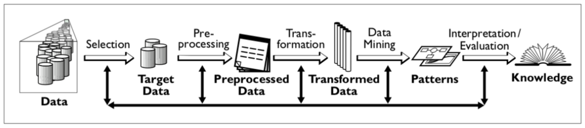


Fig. 1: Steps of KDD

A. Selection and Goals

My Initial goal for this analysis was to provide enlightening information about the music industry and the popular music that is common today, and to compare it with the popular music of years previous to see if there is any difference in what type of music is popular in regards to word usage in lyrics. This data would be valuable to those who have an interest in the Music Industry, specifically the types of music that appear on the top of the charts and maybe some understanding as to why it is popular, or any standardization to the popular music and its lyrics.

B. Targeting data

The Dataset that was selected for my target data was the cleaned subsection of the initial Billboard Hot 100 dataset, mentioned in part III. These columns were selected as they were important to the analysis and would allow me to extract the lyrics needed from the Genius API. The dataset was then subset by selecting only the songs that appeared in the top 10 each week, instead of performing an analysis of the whole 100. This was done because I wanted to focus on only the most popular songs of each year, and because there was such a volume of data I needed to cut it down to a sample. This Processes was done by using a the top 10 Map Reduce pattern to select only the top 10 songs from each week.

C. Cleansing and Preprocessing

The Cleaning and Preprocessing strategies that were implemented for the analysis. Checking for any null values within the inital dataset was essential. Afterwards, setting each column to the correct type was important. Most of the columns were set to factor when they were originally obtained and those had to be changed to Date and Char where appropriate.

A significant amount of preprocessing was done to prepare for the use of Map Reduce Patterns, Preprocessing the Initial dataset by removing anything extra in the Performer column, because the Genius API would not accept anything but exact names. [5]

Lyrics would also have to be preprocessed before they entered the Word Count Map Reduce pattern by removing punctuation and such, detailed in chapter III above.

D. Transformation

Once I had all of my datasets collected, they could then be extracted from in order to create my graphs and visual representation of the data.

First I analysed the lyricswc dataset to see if there were any common occurrences for words. At a glance I could tell that there were certain words that appeared many times over, and I wanted to check how common those words were. I extracted a subset from my lyricswv dataset, called wordByYear. This was the most popular word of each year, ignoring the other 9 words. I also did a similar extraction where I produced performer per year, this was the performer who was the most popular for each year.

E. Patterns

I took my subsets from my datasets and used those to allow me to graph the data, which allowed me to get a stronger understand of my data as it was presented visually. I used a libraries plotly and ggplot to aid in this.

V. Implementation and Architecture

The workflow for the analysis works as a singular pipeline, all ran through a single .sh script. It starts with Rstudio and switches back and forth between Rstudo and Python files when needed. This is mainly because of the heavy use of Map Reduce for the purposes of gathering data and the analysis itself.

Technologies used are R, Rstudio, Python, Sublime text editor, and Microsoft Excel for CSV files.

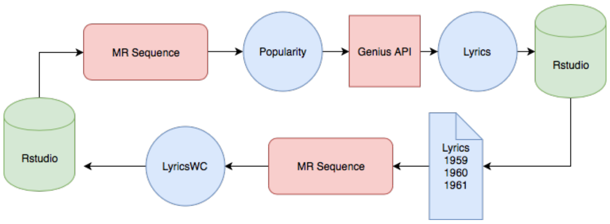


Fig. 2: Overall Architecture

The initial dataset is downloaded and loaded into R studio, where it is checked for null values, unwanted columns are removed, and columns are cleansed. Then it goes into the first series of Map Reduce Patterns, the result of these being the popularity dataset. Then this dataset is sent to a Python script that will retrieve the lyrics for each of the songs from the

Genius API [5] in the popularity dataset, based on song name and performer. This results in a new dataset called Lyrics, which is Popularity but now with an extra column storing the lyrics for each column. Both these datasets are then entered into R where they are combined to form the PopLyrics dataset. From the PopLyrics dataset, the lyrics column is cut by year and then produces 59 individual txt files, each named after the year they represent. They are then individually run through a second sequence of Map Reduce Patterns that ultimately produce a lyricswc dataset. This is the result of the lyrics being ran through the Word Count Map Reduce Pattern.

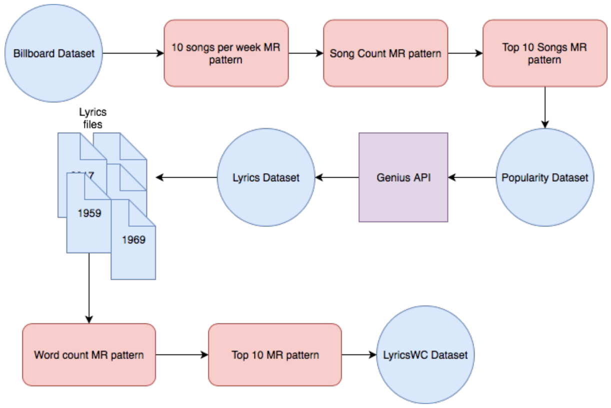


Fig. 3: Map Reduce Patterns in Detail

The above diagram shows are more detailed look at the process, it shows exactly how the cleansed Billboard dataset is transformed and how it allows me to extract the lyrics for all of the songs needed in my second dataset. It then shows the files from that second dataset being put through word count and top 10 map reduce patterns to finally have a dataset on the lyrics of popular songs.

VI. Results

The main aim of this paper is to determine if there is any significant change in popular music over time, from word usage in lyrics to individual song and performer popularity. To meet this aim, an initial dataset was cleansed, preprocessed and processed to extract the data needed for this analysis. The data extracted from an API was then combined with the initial data to answer the objectives outlined in the Introduction of this report.

Putting my data through the various forms of Map Reduce multiple times meant that I spent a lot of time with the data, the results of this analysis went against the initial assumptions I had made when starting this paper.

The first time I received the results of the second sequence of Map Reduce, One glance at the result and I could see that there was a tonne of repetition. The same few words appeared again and again throughout the years, even though my application of the Map Reduce Patterns were definitely correct.

A simple barplot on the most frequently found words in the lyricswc dataset shows this, that the word ”Love” shows up

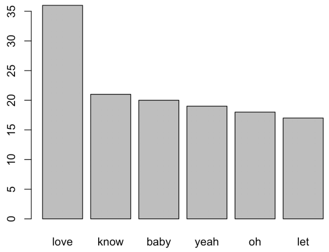
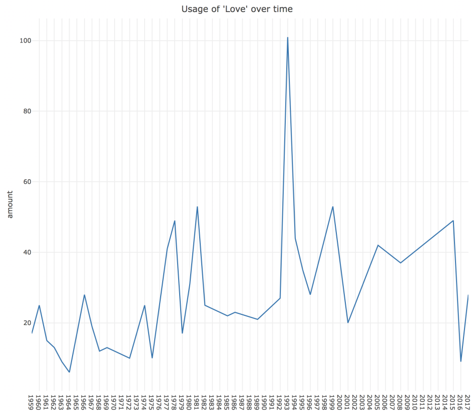


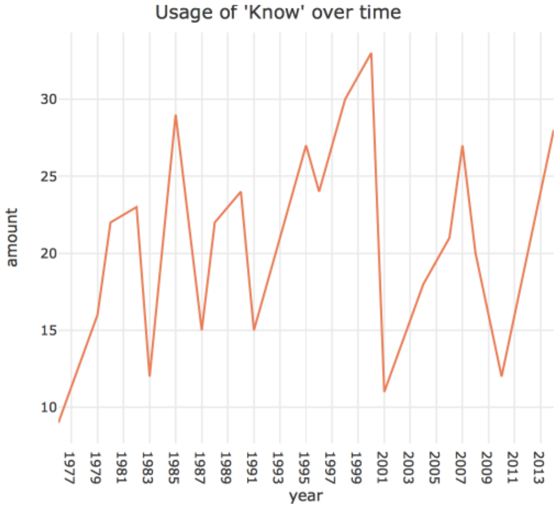
Fig. 4: Top Word Usage from all Lyrics

35 times. Followed by ”Know”, ”Baby”, ”Yeah”, ”Oh” and ”Let”. This meant that ”Love” shows up somewhere in the top 10 most used words per year, for 35 years. Thats more than half of the total years in the dataset.

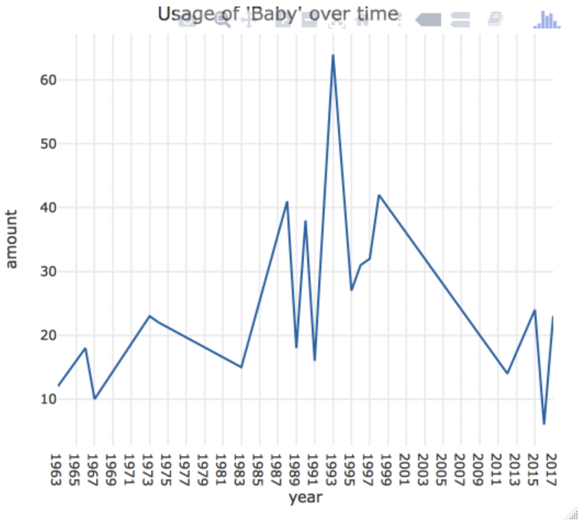
From there I wondered, If the same words are reoccuring all of the time, can I analyse their usage over time?



From the graph above you can see that the usage of ’Love’ in the top ten songs of the years lyrics dips in 82’ and doesn’t come back until the start of the 90’s where there’s a huge increase in its popularity.



In the graph above you can see that ’Know’ has strange graph, where every two years or so it drops in popularity only for its usage to increase again the following year.



In this final graph you can see the usage of the word ’Baby’ over time, it looks a though the usage of the word is minimal up until the end of the 80s where it sees a surge in popularity, with a peak in 1995, and by the end of the 90s into the 2000s its popularity gradually reduces, with a short spike in 2014.

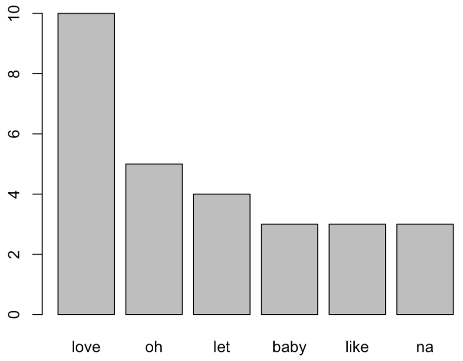


Fig. 5: Top Word Usage from 1st position

From these graphs you can see that the words used over time are actually quite consistent, even if we filter by the most used word every year, the same set of words appear multiple times. Again with ’Love’ being the most popular, being the most used word of the year for 10 years.

This completely changes my outlook on song lyrics used in popular music, the initial objective was to observe a change in word usage over time when it came to popular music, instead the data suggests that word usage has largely stayed the same over time, with ’Love’ being the most popular word used for most years.

Another analysis that can be performed with this dataset is an analysis of an individual performer and their popularity. In my objective I was under the assumption that the popularity of a performer would be consistent over time, that the legends who wrote music before would be just as popular as a performer from today. That performer popularity wouldn’t change over time.

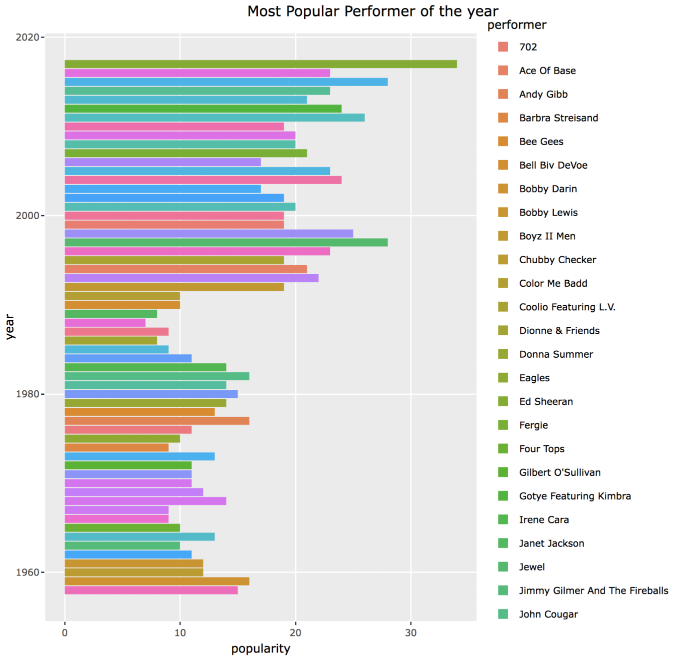


Fig. 6: Most popular performer per year

From the graph we can see that the assumption is wrong, that the popularity of the most popular performer of the year is higher the closer you get to today, popularity as a measurement of how long their songs stay in the charts. There is a relatively clear distinction halfway between 1980 and 2000, where the popularity of the top half are all consistently more than the popularity of the bottom half.

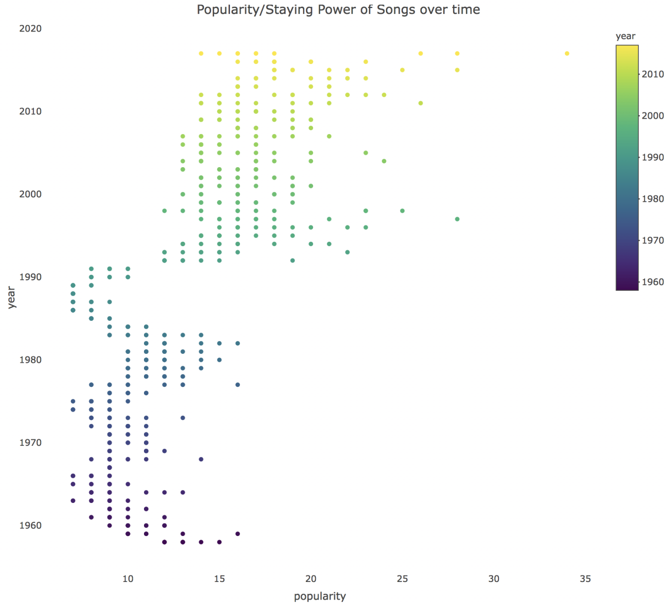


Fig. 7: Song ’staying power’ over time

The last analysis that I performed with my dataset is an analysis of a songs 'staying power' over time. What is meant by this is how long can a song stay in the top 10 songs of the week for, and for which years was this achieved? and is there any difference between older songs or newer songs? The graph below shows the difference in popularity based on year very clearly. The new songs that have released have stronger 'staying power' than the songs that released long ago. It looks as though newer songs stay in the top 10 of each week for more weeks than their older counter-parts did. This means that the popular songs of today are better at staying around and not getting knocked off by new songs coming onto the charts.

The results from these graphs show that in someways songs and their popularity have changed over time, but in other ways they have relatively stayed the same. When it comes to word usage in the lyrics of popular music, it seems that the same words are consistently used over and over, as likely that the stories these words create, words like love, are relatively universal and therefore can have mass market appeal

VII. Conclusions and Future work

In general, I learned that most of the assumptions that I had made coming into this report were wrong, that the data actually pointed to some more interesting results than the contrived assumptions I had made at the beginning. In regards to lyric usage over time, it seems that as humans we write songs about the same topics fairly consistently over time. The story of love and loss is as old as shakespeare, and this story continues to be told and retold throughout popular music regardless of the decade. However there were changes that time had made in relation to the popularity of performers and the staying power of their songs. It seems that songs are sticking around in the top 10 of the charts for a lot longer than they used to, and the performer themselves seem to be more popular than they ever have before. I can only guess as to what the reasons for this might be, but my assumptions are that the use of technologies and invention of social media and streaming platforms have made performers more famous than they've ever been before - and in turn, their songs are more popular than before. Not to mention years of iteration on what makes a good pop song, and how to consistently create songs that resonate with the public.

If I had more time I would've liked to spend more time analysing the dataset, as although I am happy with the results section of this report, I would have preferred more deeper results, and the ability to explain why the results I have exist, rather than just mention their existence and only guess at the possible reasons.

The future of this analysis could go deeper into the different aspects of each data point and why the data shows what it shows. I'd also bring in more data and perform an analysis using genre alongside the data I have here. An understanding of genre could lead to interesting visualizations where the genre could be related to color.

References

[1] Paul D. Lopes. Innovation and diversity in the popular music industry, 1969 to 1990. *American Sociological Review, Vol 57, No.1, 56-71*, 1992.

[2] Tara M. Bisel Richard L. Dukes and Others. Expressions of love, sex, and hurt in popular songs: a content analysis of all-time greatest hits. *The Social Science Journal 40 (2003) 643650*, 2003.

[3] kcmillersean. Billboard hot 100 1968-2017, 2018.

[4] michael noll. Writing a hadoop mapreduce program in python, 2018.

[5] Genius. Genius api, 2018.

[6] johnwmillr. Lyricsgenius, 2018.

[7] Gregory Piatetsky-Shapiro Usama Fayyad and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data, 1996.