Spring 2021

What’s in a Song?

Natural Language Processing (NLP) Analysis of Billboard Top 100 Songs

Nour Mohamed, Advisor: Agustín Indaco

Introduction:

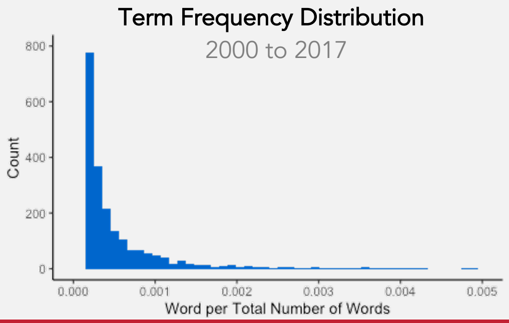
This project uses different NLP techniques to analyze music trends across the Billboard Top 100 songs. The methods used range from sentiment analysis, analyzing term frequency as well as topic modeling. The purpose of this study is to allow us to better understand changes in lyrics amongst the most popular songs in the music industry.

Dataset:

The dataset consists of 6,100 songs, from 1,770 artists, categorized into 6 different genres. The songs included in the dataset are ones which made it to the Billboard Top 100 Charts in the years 2000 to 2017. The 6 genres in the dataset are Pop, Rap, R&B, Rock, Country & EDM.

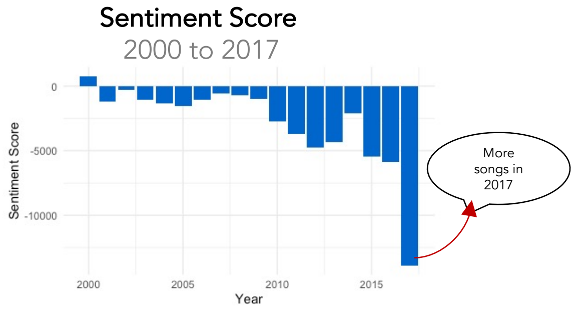
Method:

The dataset was converted into one token per row format; meaning each word of each song became a unique row. This project uses songs that have ranked at least once amongst the Top 100 in Billboard Charts from 2000 to 2017, and analyzes them based on their peak rank position each year. As an initial analysis, it can be seen from the chart below that within lyrics, many words occur rarely, and fewer words occur frequently.


Term Frequency Distribution

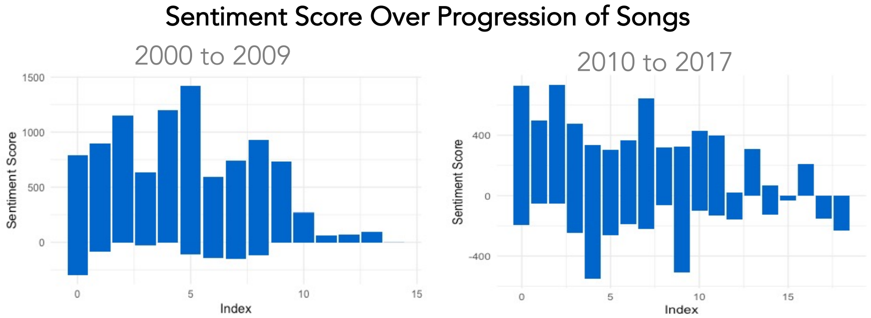
2000 to 2017

Results:


Sentiment Score

2000 to 2017

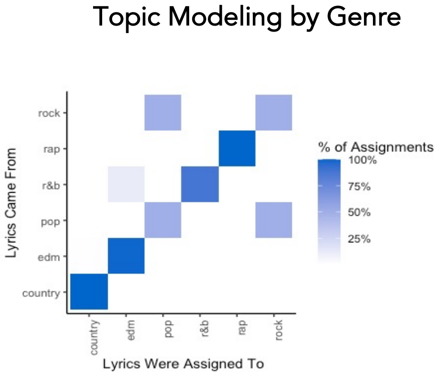
The chart above shows positive/negative sentiment scores for each individual song lyrics using AFINN lexicon. This was then grouped by year to find an overall sentiment score for each individual year. The year 2000 is the only year with a slightly positive overall sentiment score. All following years display increasingly negative scores, suggesting lyrics became more negative over time. Contributing to the exceptionally negative score in 2017 might be the total number of words in 2017 songs being around 90,000. In comparison, the average number of words in 2000 to 2016 is 45,000, and the median is 44,000.


Sentiment Score Over Progression of Songs

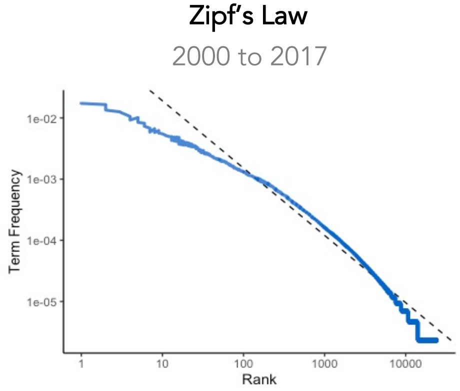
2000 to 2009

2010 to 2017

Songs released in the years 2000 to 2009 are most positive in the middle of the song, which may be attributed to the chorus. Songs released in years 2010 to 2017 instead begin with more positive words, receiving the highest sentiment score in the beginning. An interesting article published by The New York Times explained that the music industry began to change in 2010, with less focus placed only on the chorus and instead the goal being to “create music that will grab a listener’s attention from beginning to end”. The article also attributes this change to “hints of a future in which the chorus becomes subservient to the hook just so that our impatient, digitally addled brains don’t nudge us to hit skip” (Opinion I The Culture Warped Pop, for Good, 2021). This may help explain the more positive sentiment scores in the beginning of songs in the years 2010 to 2017.


Topic Modeling by Genre

The confusion matrix on the left suggests that Country & Rap have the most distinguishable lyrics, whereas all other genres had several words that were misassigned to another genre. This is interesting as lyrics from those two genres might be considered extreme opposites in terms of the type of language used.


Zipf’s Law

2000 to 2017

The blue line on the graph highlights the relationship between the frequency of a given word and its rank, showing a nonconstant negative slope. When fitting the graph with an exponent for Zipf’s law, we see that the largest deviation occurs at the higher ranks, meaning lyrics contain fewer rare words than predicted by a single power law.

Citation:

Nytimes.com. 2021. Opinion I The Culture Warped Pop, for Good. [online] Available at: <https://www.nytimes.com/interactive/2021/03/14/opinion/pop-music-songwriting.html>.

MEETING OF THE Minds Undergraduate Research Symposium

Carnegie Mellon University Qatar