The Billboard Top 100: An Analysis of Timelessness and Lyrical Content

Katie Cao

1. Research Question

In today's music industry, pop hits are a dime a dozen. But what makes a song special? What makes a song timeless? Given the wealth of data available via song-tracking music charts and online music streaming platforms, it is possible to evaluate the persistence of a song's popularity over the years and the factors that determine it. Hearing frequently hear from the older generations that today's top hits are repetitive and formulaic-- lacking in the lyrical genius of older gems-- I was motivated to tackle the challenge of quantifying the quality of lyrics and examining their changes over time. And, while a song's musical features have been well studied for their impact on song popularity[1], I found an investigation of the relationship because the quality of lyrical content and a song's timelessness to be a worthwhile and novel pursuit.

My hypothesis is that song lyrics have become more basic, repetitive and negative over time, but that all these qualities are negatively correlated with a song's timelessness, as complex lyrics stand the text of time.

2. Methodology and Results

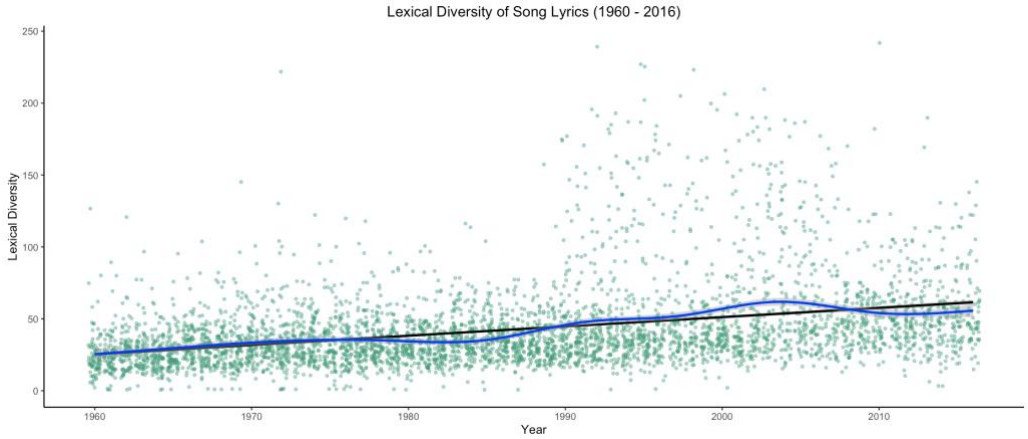
I compiled a variety of data sources in order to complete my analysis. I needed to be able to analyze song popularity over time, audio features of songs (as control variables), and the lyrics. I used three datasets. The first dataset contains every weekly Hot 100 singles chart from Billboard.com between 8/2/1958 and 6/22/2019. There are 317,175 entries with details on the song and artist name, rank, date, and total number of weeks charting at that point.[2] The second dataset contains values for each of the 28,000+ tracks pulled from the Spotify Web API. A complete and thorough data dictionary can be referenced on Spotify.[3] The last dataset contains lyrics of 5701 Billboard Top 100 Songs from 1960 to 2016.[4] It's important to note that while combining this dataset with the prior two, I lost quite a few observations because of varying song titles and missing values. Ultimately, the regression is performed on a dataset of 4607 observations. The lyrical analysis however, breaks out the words for each song, and looks at 39,000+ observations.

The lyrics data required the most cleaning. I removed entries with no lyrics, expanded contractions, removed special characters, made all lyrics lower case, added a decades column, and standardized the song ID variable to the most matchable format between datasets (Song by Artist). Many lyrics, have phrases like "Repeat Chorus", or labels such as "Bridge" and "Verse". After some preliminary exploratory analysis, I created a list of undesirable words to be removed.

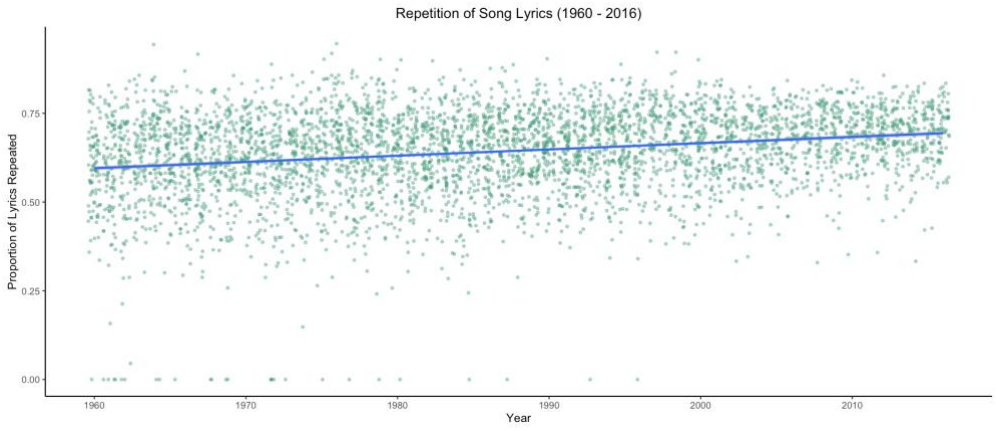
I measured a song's timelessness using its popularity over time. Figure A shows a song's Billboard Weekly Top 100 rank over time. Timelessness is calculated using the area under the rank over time curve, weighted by the total number of weeks spent on the top 100 chart. The timelessness score is out of 100, with 100 being the most timeless and 0 being the least timeless. This score accounts for each song's weekly rank and the number of weeks spent at those ranks, heavily favoring songs that chart for many weeks.

I examined four components of lyrical content and their changes over time: 1) Timeless Words 2) Lexical Diversity 3) Repetitiveness 4) Sentiment Analysis. I found the most frequently used lyrics in songs for each decade and found that the most popular lyrics, ie. “love,” “girl,” and “time” are persistent throughout time.

The more varied a vocabulary a text possesses, the higher its lexical diversity. Song vocabulary is a representation of how many unique words are used in a song. This can be shown with a simple graph of the average unique words per song over the years. After fitting both a linear model and generalized additive model, the positive correlation between the year of release and lexical diversity is evident.



Lexical density is the number of unique words divided by the total number of words, in a range of 0 to 1. This is an indicator of word repetition, in that the lower the lexical density, the higher the repetitiveness of the song lyrics. Thus, repetition = 1 - the lexical density. A linear model fitted to the graph shows that song lyrics have become more repetitive over time.



Sentiment analysis is a type of text mining which aims to determine the opinion and subjectivity of its content. When applied to lyrics, the results can be representative of not only the artist's attitudes, but can also reveal pervasive, cultural influences. I used a lexicon-based approach to categorize song lyrics as

either positive or negative, and as one of eight categories of emotion, finding that Lyric sentiments has remained relatively the same over time.

Lastly, I performed a multivariate model predicting the timelessness score of a song using the Repetitiveness, Vocabulary Diversity, and Sentiment of the lyrics. The model uses controls for various song features including the Tempo, Energy, and Valence.

Results

	Timelessness	
Predictors	*Estimates*	*Conf. Int (95%)*
Intercept	-0.51	-3.16 – 2.14
Repetitiveness	14.22 ***	11.22 – 17.23
Vocabularty Diversity	0.02 ***	0.01 – 0.04
Positive Sentiment	-1.39 *	-2.65 – -0.14
Tempo	0.01	-0.01 – 0.02
Energy	4.70 ***	2.73 – 6.66
Valence	-7.07 ***	-8.51 – -5.63
Observations	2545	
R^2 / R^2 adjusted	0.082 / 0.080	

** p<0.05 ** p<0.01 *** p<0.001*

According to the results of the regression, the lyrical repetitiveness and vocabulary diversity are both significant predictors of the timeslessness of a song. Though the effect of a higher proportion of positive sentiment words in lyrics is not significant at the 0.01 level, it is still negatively associated, meaning on average, more timeless songs have more negative sentiments.

3. Discussion

One limitation I noticed immediately is that data is inconsistently available throughout time. I removed all songs not within the 1960-2016 range because of date limitations between the three datasets. The datasets also did not cover the same songs-- the billboard data reported on weekly charts while the lyrics data reported for the top 100 of each year. The lyrics data was significantly limiting the number of observations used for the regression, and it would be beneficial to perform a manual webscrape to get lyrics for all the songs in the billboard and spotify datasets. The last limitation is inherent to the sentiment analysis package. This lexicon-based approach assumes that the emotional and positive/negative classifications of a set dictionary of words is correct. However, a manual sanity-check of some of the assigned sentiments shows that many words colloquially used positively were marked as negative.

A few additional avenues of research would be to investigate timelessness using metrics outside of Billboard charting data, because timelessness doesn't necessarily mean topping the charts.[5] In additional, it would be interesting to utilize natural language processing and machine learning to create and test a more comprehensive model for predicting a song's popularity using its lyrics.

References:

1. https://medium.com/@albert.w.berger/what-makes-a-song-popular-in-a-certain-country-feat-p%CC%B6i%CC%B6t%CC%B6b%CC%B6u%CC%B6l%CC%B6l%CC%B6-spotify-cd705abc59
2. https://data.world/kcmillersean/billboard-hot-100-1958-2017
3. https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/
4. https://cran.r-project.org/web/packages/billboard/billboard.pdf
5. https://pudding.cool/2017/05/song-repetition/