What can prevalence of personal pronouns in song lyrics tell us about how popular a song is?

Submitted by: Varsha Venkatesh | 1003646

Objective

It is claimed that songs that are personal are more relatable and listeners respond to them better. Through this research I hope find out if this is indeed true.

Background, purpose and significance

Analysing music to gain new insights about human behaviour, culture and lifestyles during the time in which they were produced is not a new concept. For example, "What has America been singing about? Trend in themes in the U.S. top-40 songs: 1960 – 2010" [1] is a study that explores the link between music lyrics and adolescent habits and welfare. The project, "Message in the Music: Do Lyrics Influence Well Being?" [2] explores the relationship between lyrics and its potential to influence positive psychology.

On the commercial side, "All You Need is Love?" [3] is a study that analysed songs and identified the top themes of each decade. It lead to the ability to predict if a song would be commercially successful. Such predictions would be of great interest to musicians, music companies, to create the next big hit, or advertising companies that might want to design advertising & marketing that resonate with the audience.

The results of lyrical analysis have many implications. For humanists, it is new knowledge about the human condition. It is possible to identify tropes, detect patterns, spot trends, and derive insights about an individual in a society and the society as a whole.

In the article "Decoding Songwriting With Data" [4] which is based on a talk that he gave along with Paul Jacobsen at the SXSW 2017, data artist and designer Eric Boam reports his findings from a study in which they analysed the lyrics of 25 albums (approximately 300 songs) using basic tools like readable.io, wordcounttools.com, and databasic.io. Among the many measurable dimensions they extracted from the lyrics, were pronouns. One of their findings was that the first person singular pronouns, "I" and "you" were the top two most used

pronouns in nearly every album. He suggests, “One way songwriters make songs more accessible and personal to listeners is through the use of pronouns” and “listeners respond to personal songs that they can inhabit.”

In this research project I propose to primarily test this claim – whether songs that are more personal also more popular among listeners? And if the answer is yes, I want to find if we can predict the popularity a song.

Brief review of relevant literature

I take inspiration from the Stanford Literary Lab Pamphlet “Loudness in the novel” [5]. The study measured the “loudness” of a novel by analysing all the voices in it. They were categorized according to the levels of loudness – loud, neutral and quiet, as determined by the dialogue’s speaking verb. Similarly, I would like to measure how “personal” a song is by analysing the pronouns in it. The method that they used was a most-distinctive-word test (MDW). The test “identifies the words which most distinguish one category from a second category”. Through their study they discovered that in addition to speaking verbs, there is also a correlation between loudness and grammatical mood. Based on this, they used different methods to do a structural analysis of novels. Additionally, to observe the trend of loudness over the decades they plotted the percentage of loud speaking verbs against neutral speaking verbs. We can use some these methods and insights and apply it to our corpus of lyrics.

After reading the pamphlet I felt that in some sections, for example the structural aspect, their choice of method to measure and depict the loudness was not explained. Amongst the methods they used some produce more significant results than others, since both their indicators of loudness greatly depend on genre, author and writing style. Their conclusion was that the loudest points in the novel are not effective at conveying information, and the reason must be because “loud structures affect the reader in a way that details, no matter how interesting, cannot.” I felt that there was no strong distinction between the loud structures and other characteristics of the novel. There was also no indication of how we might confirm that the loudness affects the readers. Moreover, loudness is not a structural aspect of all novels. Their study is largely experimental and many aspects remain speculative.

There are many parallels between study and mine, as I am also focusing a key feature of the text to measure something more abstract. In my study, however, I will attempt to focus more on this single variable and keep all other factors such as the quantity, nature, standard of the corpus as uniform and consistent as possible. I will be applying the same method of analysis on all data, and will have a control group to minimize the effects of other variables to produce a conclusive result.

Research questions

- Are songs that are more personal also more popular among listeners?
- What combination of pronouns best predict the popularity of a song?

Proposed Methodology

I propose a two phase research process. In the first phase I hope to answer the easier question of whether the popularity of a song correlates with the prevalence of personal pronouns in the album. This is the central claim made by Eric Boam that I want to test if true. If it is true, the second phase of the research will delve deeper into the analysis of the types of pronouns, their combination, and proportion in the lyrics to arrive at an "algorithm" that best predicts the popularity a song. The algorithm will then be tested on a new corpus to validate its efficacy.

Corpus

I propose to compare the most popular songs ('popularity' as defined by most listened to, liked, or bought by the population) with the most critically acclaimed songs ('critical acclaim' as defined as the best songs by music critics, experts, musicians and music industry personalities).

For the first set of songs, I will use the "Billboard Hot 100 Chart" [6] , which is a chart produced weekly with rankings based on sales, radio play and online streaming as one set of data. They have archives from 1958 onwards, and the songs in the charts best reflect what people enjoy listening to.

The second set will consist of songs that are highly rated by critiques, music experts and industry figures from reputed lists such as "Rolling Stone's 500 Greatest Songs of All Time" [7],

New Musical Express's "The 500 Greatest Songs of All Time" [8] and Pitchfork's "The 200 Best Songs of the 2010s" [9]. This will ensure the songs selected are all almost equally good, diverse, and from the same period.

Of course, there will be overlaps between the two sets so I will take only the mutually exclusive songs in each set into consideration. I will also remove the duplicates within each set. The first set of data will be scaled down to match the second. The lyrics of these songs will be extracted for analysis using the Genius API[10]. The diagram (fig.1) below illustrates the two set of songs for comparison study.

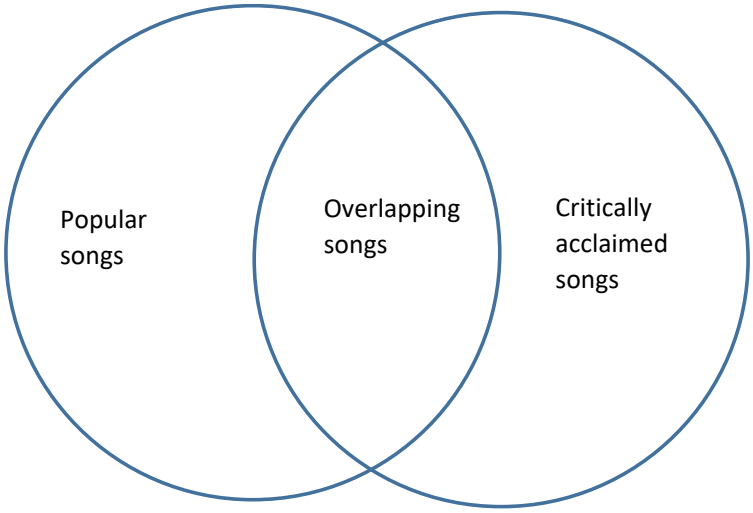

Figure 1: Song sets for the comparison study

One of the assumptions that I am making is that while the songs that made it to both the sets are liked not merely for their lyrics (but also for the overall quality of the music such as vocals, instruments, tune etc.), it is more likely that the critically acclaimed songs are assessed more holistically on a broad range of criteria than the popular songs; and that the popular songs are more likely to be popular because people personally relate to them better.

Phase 1

In this phase of the study, I will use the web-based digital text analysis environment, Voyant[11], to perform a simple frequency analysis of the personal pronouns I, you, he, she, it, we, they, me, him, her, us and them in each song in each set of songs. If the frequency of

these personal pronouns is statistically significantly more prevalent in the first set (popular songs) than in the second set (critically acclaimed songs), I can then reasonably conclude that songs with more personal pronouns are more popular.

Phase 2

Moving beyond the frequency of pronouns, this phase of the study will analyse the workings of the pronouns (table 1).

Pronoun as subject	Pronoun as object	Possessive pronoun	Reflexive pronoun
I	me	mine	myself
you	you	yours	yourself
he	him	his	himself
she	her	hers	herself
it	it	its	itself
we	us	ours	ourselves
you	you	yours	yourselves
they	them	theirs	themselves

Table 1: List of personal pronouns

Using natural language processing (NLP) techniques[12] through TextBlob (Python library for processing textual data) [13] we should be able to perform part-of-speech tagging, pronoun phrase extraction, and classification. Using algorithms such as the RAKE (Rapid Automatic Keyword Extraction)[14] we can determine key phrases in a body of text by analysing the frequency of word appearance and its co-occurrence with other words in the text.

I want to see how the following measures correlate with the popularity of the songs.

- Frequency and types of personal pronouns
- Ratios of types of personal pronouns as a proportion of the total
- Co-occurrence of pronouns with other pronouns

And if there emerges a consistent pattern of occurrence of these measure across songs that are popular, we should be able to codify the pattern into an algorithm. This algorithm will then be tested on a new corpus of songs to predict their popularity.

Research Limitations

The research attempts to answer two questions. However only the first is based on a valid claim. The second question is personally interesting to me, but is purely based on a speculative claim. Furthermore it is not clear how extensively the analysis of pronouns have to be carried out to arrive at the algorithm that has the power to predict the popularity of the songs.

References

1. Christenson, Peter G., Silvia De Haan-Rietdijk, Donald F. Roberts, and Tom F.m. Ter Bogt. “What Has America Been Singing about? Trends in Themes in the U.S. Top-40 Songs: 1960–2010.” *Psychology of Music* 47, no. 2 (2018): 194–212. https://doi.org/10.1177/0305735617748205
2. Fox, Tricia. 2015. "Message In The Music: Do Lyrics Influence Well-Being?". *Repository.Upenn.Edu*. http://repository.upenn.edu/cgi/viewcontent.cgi?article=1094&context=mapp_capstone.
3. Chang, Bettina. “Can a Song's Lyrics Predict Its Commercial Success?” Pacific Standard, March 19, 2014. https://psmag.com/economics/can-songs-lyrics-predict-commercial-success-76936
4. Boam, Eric. “Decoding Songwriting With Data.” Medium. Medium, April 18, 2017. https://medium.com/@ericboam/decoding-songwriting-with-data-9d06117eed93
5. Katsma, Holst. “Loudness in the Novel.” *Stanford Literary Lab Pamphlets*, September 2014. https://litlab.stanford.edu/pamphlets/
6. “Billboard Hot 100.” Wikipedia. Wikimedia Foundation, October 28, 2019. https://en.wikipedia.org/wiki/Billboard_Hot_100
7. “Rolling Stone's 500 Greatest Songs of All Time.” Wikipedia. Wikimedia Foundation, October 29, 2019. https://en.wikipedia.org/wiki/Rolling_Stone's_500_Greatest_Songs_of_All_Time.
8. Barker, Emily. “The 500 Greatest Songs Of All Time: 100-1.” NME. NME, February 9, 2014. https://www.nme.com/photos/the-500-greatest-songs-of-all-time-100-1-1421753
9. Pitchfork. “The 200 Best Songs of the 2010s.” Pitchfork. Pitchfork, October 21, 2019. https://pitchfork.com/features/lists-and-guides/the-200-best-songs-of-the-2010s/

10. Chrishyland. “Scraping Lyrics from Genius API.” Chris Hyland, February 18, 2018. https://chrishyland.github.io/scraping-from-genius/
11. “Voyant Tools.” Voyant Tools. Accessed November 1, 2019. https://voyant-tools.org/
12. “Natural Language Processing (NLP) Techniques for Extracting Information.” Search Technologies. Accessed November 1, 2019. https://www.searchtechnologies.com/blog/natural-language-processing-techniques
13. “Simplified Text Processing.” TextBlob. Accessed November 1, 2019. https://textblob.readthedocs.io/en/dev/index.html.
14. “Rake-Nltk.” PyPI. Accessed November 1, 2019. https://pypi.org/project/rake-nltk/