Christian Schroeder (dbn5eu@virginia.edu)

DS 5001: Exploratory Text Analytics

15 December 2021

Exploratory Analysis of the Correlation of Lyricism to Awards Won in Rap Music

Introduction

For this project I decided to analyze the lyrics and sentiment of the most lyrical rappers. To define the 'most lyrical' rappers, I used the output data of a well-known project by Matt Daniels from 2019, "The Largest Vocabulary in Hip Hop." In his work, Daniels compiled a list of 160 rappers with the most diverse vocabularies throughout their discography. I used this list and lyrical rankings, along with data on Grammy and Billboard awards and nominations, to explore how diversification of lyrics could impact the success of a rapper. It should be noted the artists who appear on the list provided by Matt Daniels is ultimately subjective, as he could not analyze and include every rap artist, rather he relied on his estimates and the input of the rap community.

Data and Preprocessing

To download the lyrics for songs by select artists, I used the Genius API. Through this I could request the top k most popular songs by each artist in my list, where popularity is determined by the frequency of visits to that song's lyrics page on the Genius website. As Genius is the leading website for song lyrics, I felt confident that their popularity rankings for each artist would roughly match those that may be found from streaming services. I had to make sure I was pulling enough songs per artist to get a corpus that strongly defined each rapper's style and popular themes. I wanted as much data as possible but did not want to introduce possible bias by including songs from artists that were not popular. So, I decided to pull the top 25 most popular songs per artist. This gave me a substantial number of songs to work with, while hopefully not minimizing possible bias from unpopular songs. In total, my final dataset of songs included 160 artists and 3,975 songs.

Before I could analyze the lyrics of the songs, I needed to preprocess the data into the proper format. This involved creating several tables for storing the data, including LIB, DOC, TOKEN, VOCAB, and TFIDF. These tables were then used and as the input data for my exploratory text analysis.

Exploratory Data Analysis

Topic Modeling using Latent Dirichlet Allocation (LDA)

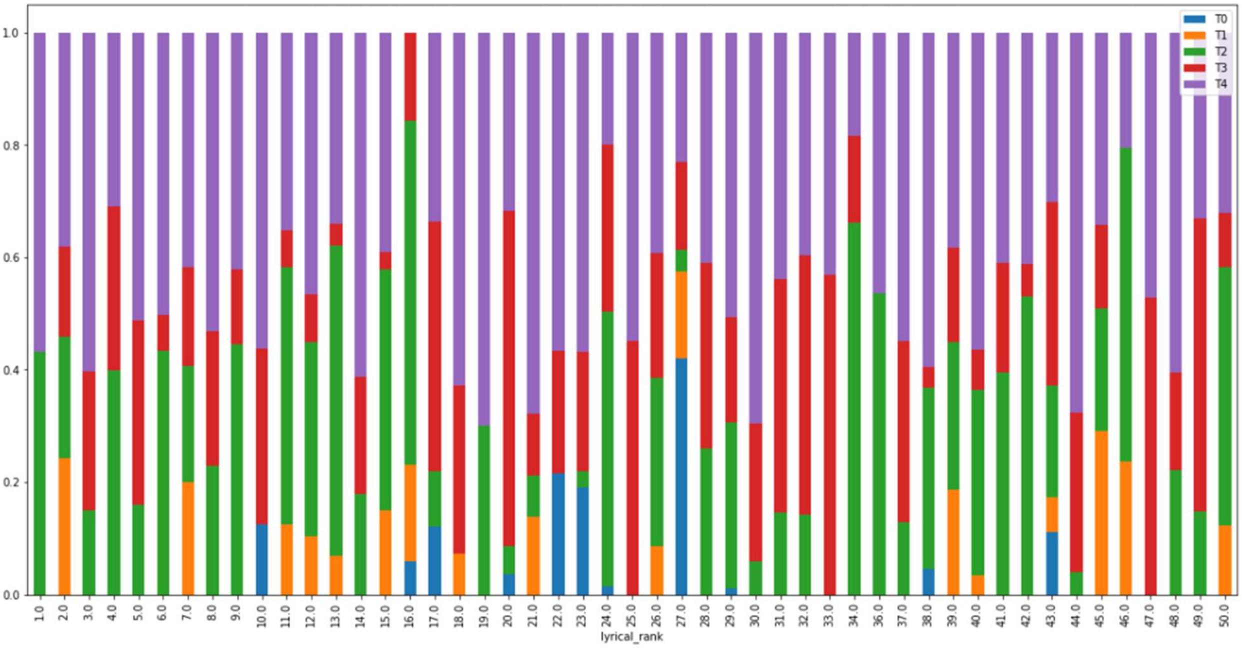
To perform topic modeling on the data, I used LDA in the form of collapsed Gibbs sampling. This allowed me to sample each possible document/word combination and determine the most probable topic to fall into. I chose to create 5 topics and iterate the sampling 500 times. I also used an alpha of 0.02 to and beta of 0.2 to allow more flexibility in the assignment of topics.

- Topic 0 geared itself towards explicit or vulgar language, as well as material items like house, chain, paper, and diamonds.

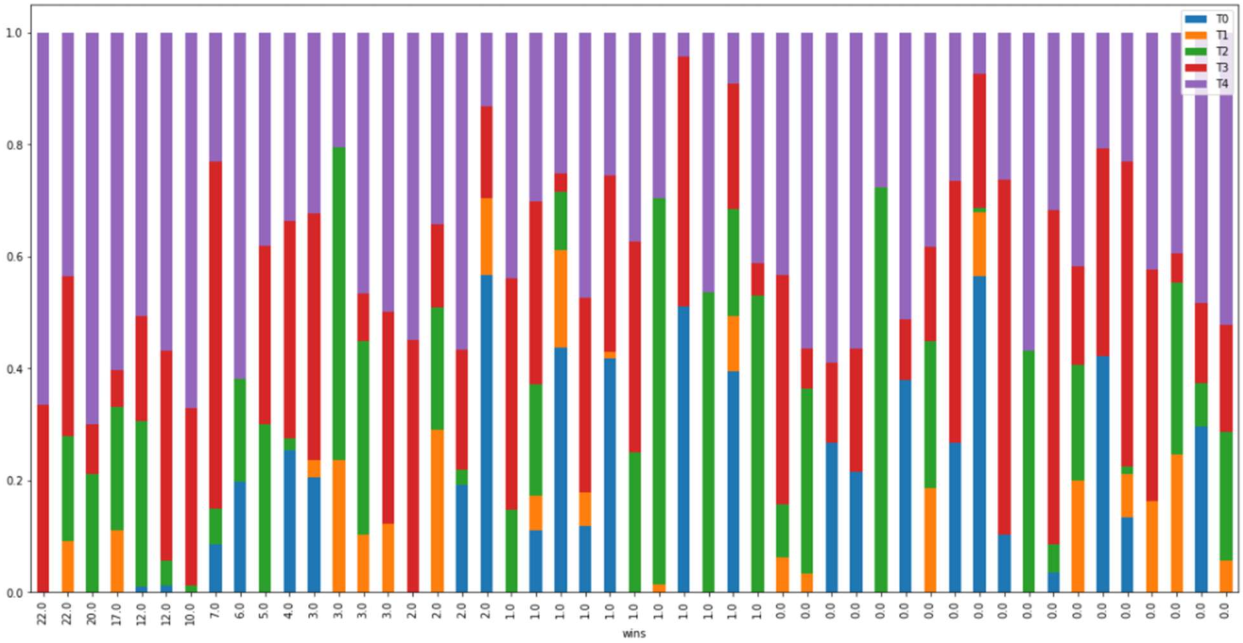
- Topic 1 was more towards lyrics about desires or actions, with terms like moment, tongue, and motion, but also mentioned material items like diamonds and watch.
- Topic 2 lent itself more towards the observational or story-telling side with the most common terms being people, rappers, school, everything, and friend. It also included terms for family members.
- Topic 3 seemed to be more associated with songs related to life in impoverished areas with terms like streets, ghetto, and drugs.
- Topic 4 appeared to also be more observational and lent itself to possible storytelling with terms like world, night, people, times, and years.

Common explicit language expected in rap music appeared regularly in all topics except for topic 2. Which may imply artists most related to that topic are less likely to use that language in their music.

When looking at the distribution of topics for the top 50 artists by unique vocabulary, I did not see a noticeable pattern as the ranks decreased. Although, I did notice that an increased presence of topic 4 usually came with a decreased presence of topic 2, which may have come from both topics being more observational lyrics and similar lines being classified as either or. Aesop Rock (1) is an interesting example where he was only classified as those two topics, which given his style of music may be expected. We also have A Tribe Called Quest (33) who only had two topics, 3 and 4, showing their music to be observational storytelling through the lens of impoverished living.



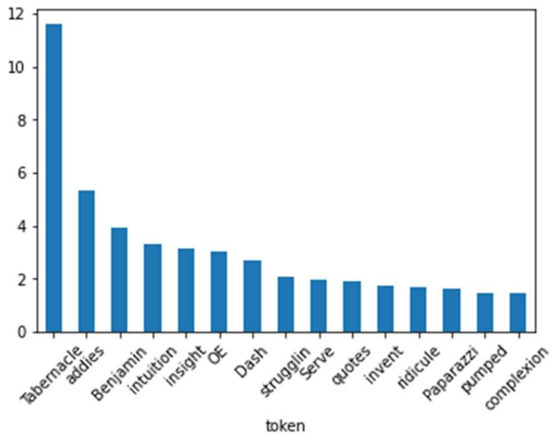
I also looked at the same distribution for the 50 artists with the most awards won. When comparing this distribution to the lyricist distribution, I noticed a much higher presence of topic 0 and slight increase of topic 1. This may mean that the use of explicit language or lyrics about desires could be beneficial to an artist's success.



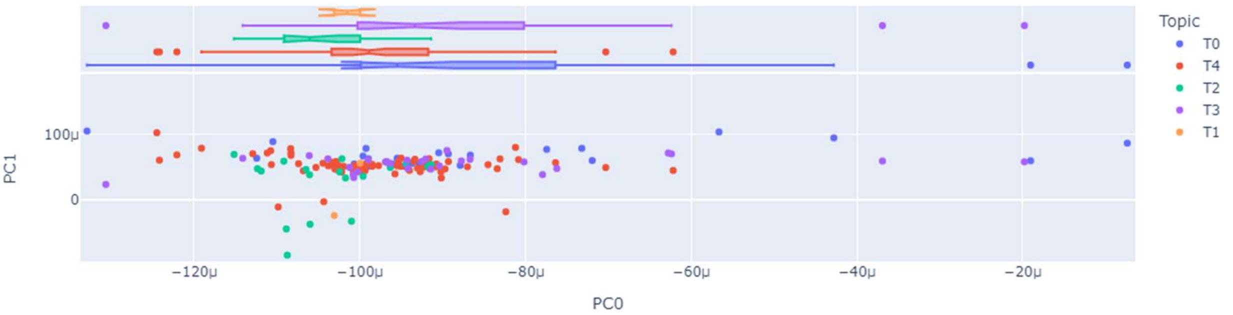
After defining the topics, I was also interested in seeing how different clustering methods would handle grouping together artists by their assigned topics. I tested Manhattan, Euclidean, Euclidean 2, cosine, and several others. I found the first three were the most effective and produced very similar results by grouping topics. After doing a spot check of how the artists themselves were further grouped, I believed the Euclidean method to be the most effective. This was based on the proximity of certain artists to others based on my perception of their styles. This method grouped artists like Aesop Rock and MF DOOM as a pair and included Del The Funky Homosapien and A Tribe Called Quest in the same area of the dendrogram.

Principle Component Analysis (PCA)

I also performed PCA to determine which terms are responsible for the most variance in the data. I was also interested in seeing if the principal components would show any variance between the modeled topics. After getting the principal components and loadings per each term, I found the terms with the highest explained variance were the ones below.

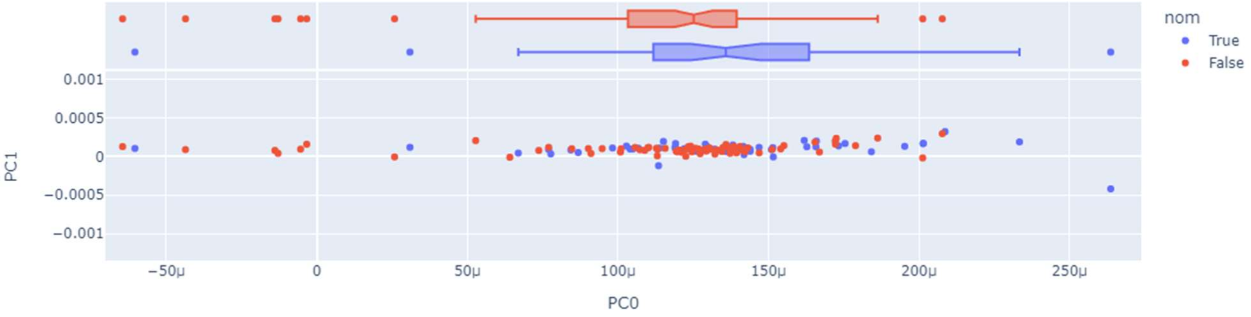


I found that the first component, PC0, explained a lot more variance between each artist than PC1. I also found that there may be a difference between the topics, but not a strong one, as they all overlap in the plot below. This plot and the others are zoomed in to display any spread seen on PC1.

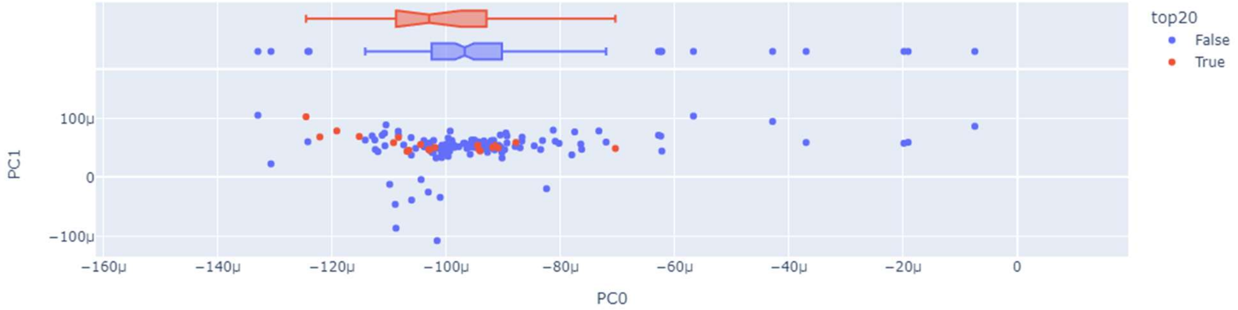


Topics 3 and 0 show to be very similar as they overlap completely. There is a noticeable difference between topics 2 and 4, which may disprove my earlier idea that they were very similar, rather they both may be observational topics but there is something causing a strong distinction.

I checked if the components may explain any difference between artists that had ever been nominated for awards or not. The groups do overlap a good amount, but PC0 did show some difference from those who won or not.



A similar distribution can be seen across PC0 when looking at the top 20 artists by unique lyrics. The median was more to the left, but the quartiles weren't as far spread.

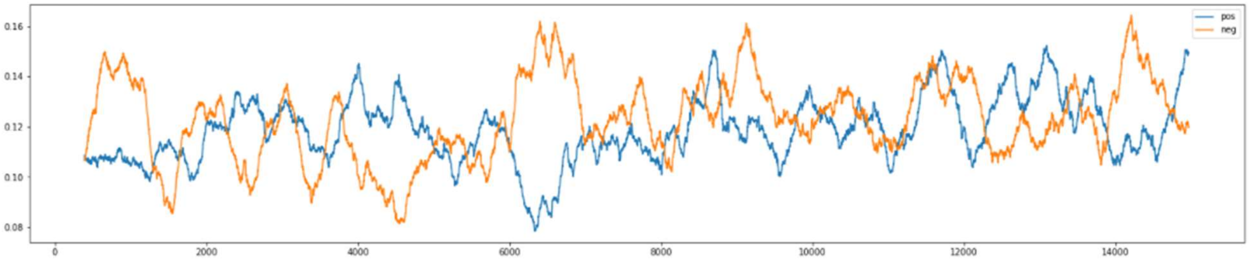


The general conclusion I came to on the principal components was that PC0 explained almost all the variance between each artist.

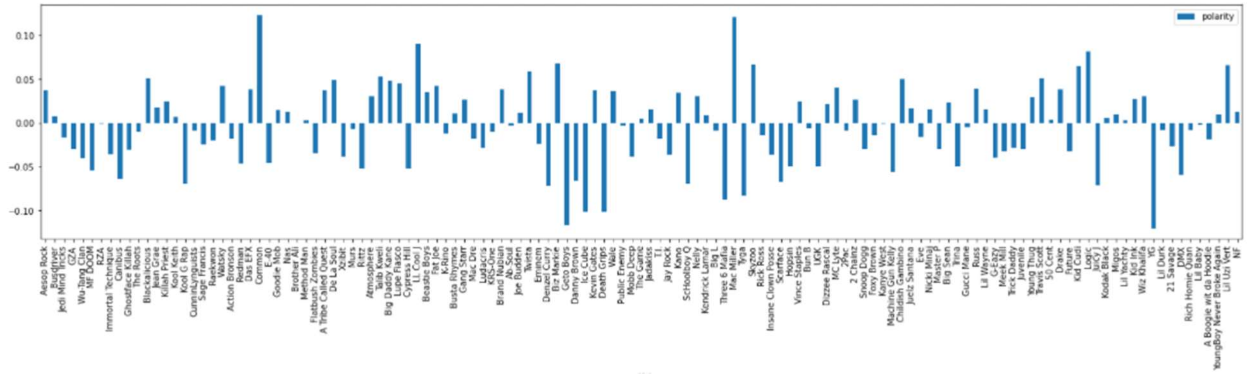
Sentiment Analysis

For sentiment analysis I used the VADER package to get the positive, negative, and polarity values of each verse on the songs. I went by verse to retain any ideas that artists display across multiple lines in a song.

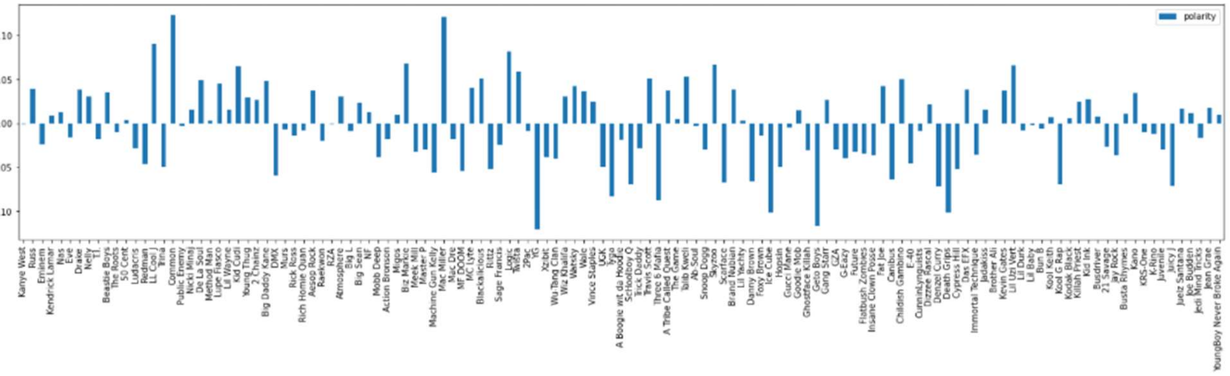
The positive and negative values plotted across lyrical rankings showed an average distribution of sentiment as the rank decreases. This showed little change across the ranks, although there is a slight increase in negative sentiment as the plot goes along.



Graphing the mean polarity of each artist showed the same results, where no real pattern was recognizable when following lyrical rank. Though, it is interesting to see the lack of artists that are overall positive in polarity.



However, when comparing artists by most awards won, there was perhaps an increase in positivity seen in artists that won awards more than others. Of the top 10 artists by awards won, there were several that were positive, with the others hovering close to neutrality. This was different from the top 10 most lyrical artists that were relatively heavily negative.



Although there was no definite pattern between the sentiments of artists that won awards and those with the most unique vocabularies, there could possibly be an underlying pattern of positivity that could improve an artist's chances of becoming more popular.

Conclusion

This analysis of the most awarded and most lyrical rap artists did not conclude exactly how I expected when I started. I did expect to see a noticeable pattern or distinction between the artists, but I ended up showing they are not that much different lyrically, at least from the scope of this analysis. Although, I did find the results of the LDA topic modeling to be insightful, and I believe the topics created do properly classify the style of lyrics of each artist to an extent, shown by the ability of the Euclidean method to successfully group them together.

I think the principal component analysis could have been more insightful if I had treated each song as each "document" rather than each artist. I would like to further investigate that if I have time.

I did not have the results I expected from this analysis. If further investigated, one may uncover more specific topic distinctions or perhaps semantic differences between the artists that I investigated. A possible cause for the lack of distinction between the artists with the most unique lyrics and those without is the fact that the most lyrical artists use specific words less. Without repeated use of certain words or topics, it may have made it harder to classify the artists into specific groupings, opposite to the artists with less unique lyrics that ended up having more robust vocabularies that were easier to classify. Perhaps that same principle takes effect in fans when listening to artists. New fans may end up steering away from songs with an excess of unique lyrics because they may be harder to remember.

Overall, I believe any investigation of music that strictly focuses on lyrics will not be completely accurate. The lyrics of a song are only a piece of what creates music. Even if we may classify the lyrics of MF DOOM as being closer related to Aesop Rock or K.A.A.N., when listening to the production behind the beats and style of delivery, that same relationship isn't as direct. The speed, rhyme scheme, and style of delivery of lines all play a very important role in the success of a song, and that can't be captured strictly from the lyrics. If someone is to accurately identify and classify rap music, I believe those other factors would need to be considered.