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 lyrical styles and semantics 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 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, styles, and semantics could impact the success of a rapper.

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.

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

1. 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 0.05 to and beta of 0.4 allow more flexibility in the assignment of topics. Of the five topics created, two were focused on explicit or vulgar words. The other three generally lacked that language.

Topic 0 seemed to be a middle ground of observation and introspection, with terms hinting towards experiences like drugs or friends, as well as possibly more observational terms like everything, nothing, and dreams.

Topic 1 lent itself more towards the observational or story-telling side with the most common terms being people, world, thing, and place. It also included terms for family members.

Topic 2 was heavy towards lyrics about desires, sexual or physical, that included terms like tonight, dirty, clothes, pound, dough, watch, and chips.

Topics 3 and 4 were very similar and geared themselves towards explicit or vulgar language like n\*\*\*a, bitch, and pussy, as well as material items like paper and diamonds.

1. 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, 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.

A picture containing Word

Description automatically generated

Topics 3 and 4 show to be very similar, as expected, but I did not expect such spread for topic 0.

I also checked if the components may explain any difference between artists that had ever been nominated for awards or not. I also found an overlap between the two, but PC0 did show some difference from those who won or not.

Chart, scatter chart

Description automatically generated

1. Word Embedding
   1. TSNE
2. Sentiment Analysis
   1. VADER