Akeem Wells (ajw3rg)

Analysis of Music Genres of the Billboard Top 100 from 2009 to 2020

When listening to music, is it possible to assess the sentiment of various genres across time and across genres through their lyrics? This is the question that is posed to be answered through an evaluation of the lyrics of specific genres using topic models and sentiment analysis of the Billboard Top 100 from 2009 to 2020. Through this pipeline, determination will be made if Rap, Country, Pop and R&B/Hip-Hop displays strong direction to the emotions of anger, anticipation, disgust, fear, joy, sadness, surprise and trust while also measuring the polarity in those songs across years. It's worth a note that sentiment analysis might not work well with standard lexicons on Rap sounds due to some curse words. In order to mediate this issue, we added a few words to the standard lexicons list but not every single word. We do this because we still need to consider that those that are not native to the vernacular can still listen to the song within a genre if they understand the language of the song and; unless we have a measure of lexicons at the time of song release, we need to make use of a static lexicon so words that may have changed meaning across years in Rap have been omitted.

The data in this analysis consists of Top 100 billboard.com year-end hot 100 songs for the years 2009-2020. (www.billboard.com/charts/year-end/2020/hot-100-songs) Each chart consists of 100 songs x 11 years for a maximum library of 1100 songs. Artists are divided into 4 categories(genres): Rap, Country, R&B / Hip-Hop and Pop. This data is retrieved from the year end top artists of each category (www.billboard.com/charts/year-end/2020/pop-songs-artists)

We use this to create a list of records in the format: Song-Title, Artist, Genre. Each record is fed into a lyricsgenius api (python module: lyricsgenius) which then returns a string of song

lyrics for that song saved in a .txt file in the format /year/genre/<title>---<artist>.txt with each line in the .txt representing 1 line of song text (a lyric) from Genius.com.

To remove the duplicate genres and songs spanning multiple years and genres, we use a top down approach to import the data into the corpus maintaining the artists first recorded song and genre for that song is used as their genre for the analysis (i.e. artists that start in country and move to pop remain in country for this analysis). We do this under the assumption that the artists reaching the top billboard in early years did so appealing to a specific audience and that audience will remain for subsequent years. Artists that fall under other categories (Rock, Alternative, Jazz, etc) have been excluded from this song aggregation leaving us with a LIB (index: song_id, columns: title, artist, year, song_file, genre) of 881 songs:

- Pop Songs: 277
- Country Songs: 145
- R&B / Hip-Hip Songs: 228
- Number of Rap Songs: 231

After preprocessing, tokenizing, annotating and stemming. We have a

- VOCAB
  - index: term_id
  - columns: term_rank, term_str, n, num, stop, p_stem, pos_max)
- TOKEN
  - index: genre, artist, title, sent_num, token_num
  - columns: pos_tuple, pos, token_str, term_str, term_id
- DOC
  - index: song_id, sent_num
  - Columns: line_str, artist

For this project, we decided to perform a sentiment analysis to explore sentiment and emotion values as features in VOCAB. We can look into the sentiment polarity and emotions for each song that is part of our dataset. In order to do this analysis, we take our TOKEN and LIB tables to apply the lexicon. Prior to the sentiment analysis, we conducted a Latent Dirichlet

Allocation (LDA) topic model with respect to each song within the data subset by each genre.

Exploring the topics, we can see the top terms are allocated as:

- Country: **im youre hey way cause baby dont time girl yeah**, 0.143131
- Pop: **baby yeah dont youre cause love time im night ooh**, 0.115541
- Rap: **im yeah bitch money ayy nigga shit man niggas fuck**, 0.195870
- R&B / Hip-Hop: **im youre hey way cause baby dont time girl yeah**, 0.105652

We can see that the genres of Country and R&B / Hip-Hop share the same topic model. Looking

at the clustering level the topics, we can see the dendrogram linking various topics together

shows that there are distinctions between various genres , however; genres such as R&B/Hip-hop

and Country can be seen as close to each other, which can be an indication that their topics do

not differ greatly. With this topic model execution, we can now look at the sentiment analysis

between the various genres. First we can compare the genres using Jockers's Discrete Cosine

Transform (DCT) method:

country (DCT)

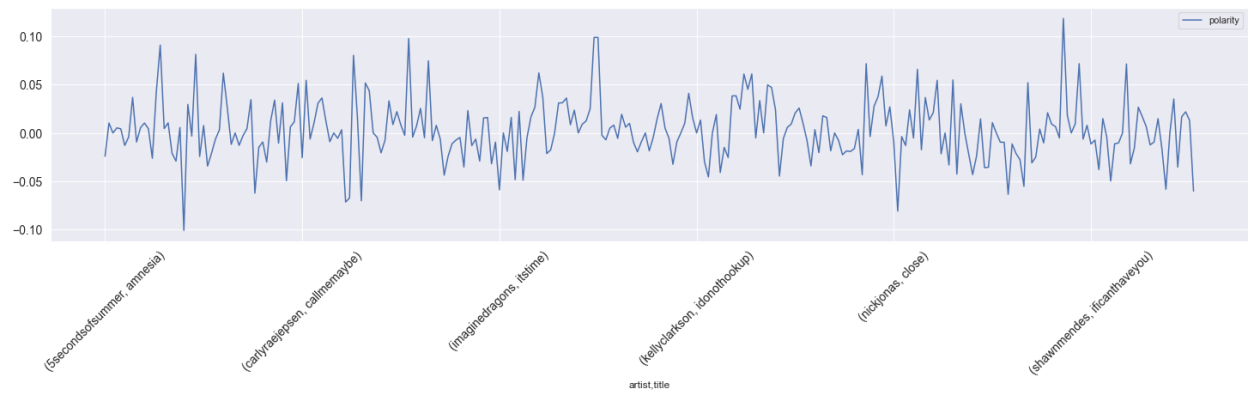pop (DCT)

rap (DCT)

rnbhiphop (DCT)
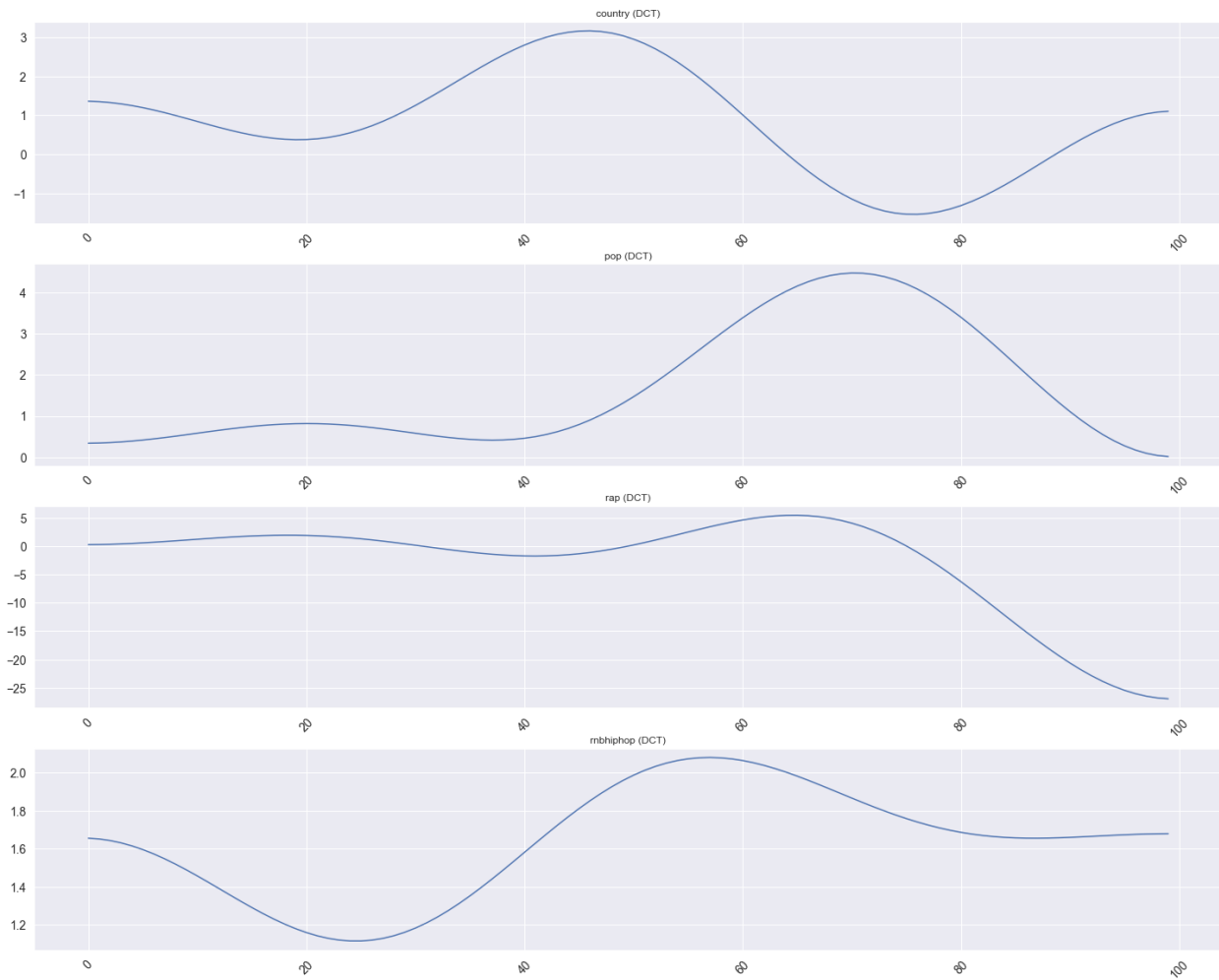
We can see from the above plot, that Country and R&B / Hip-Hop relatively follow the same trend that was presented from the topic model. We can also see that Pop is not too far away from the others. Looking at the polarity plots for
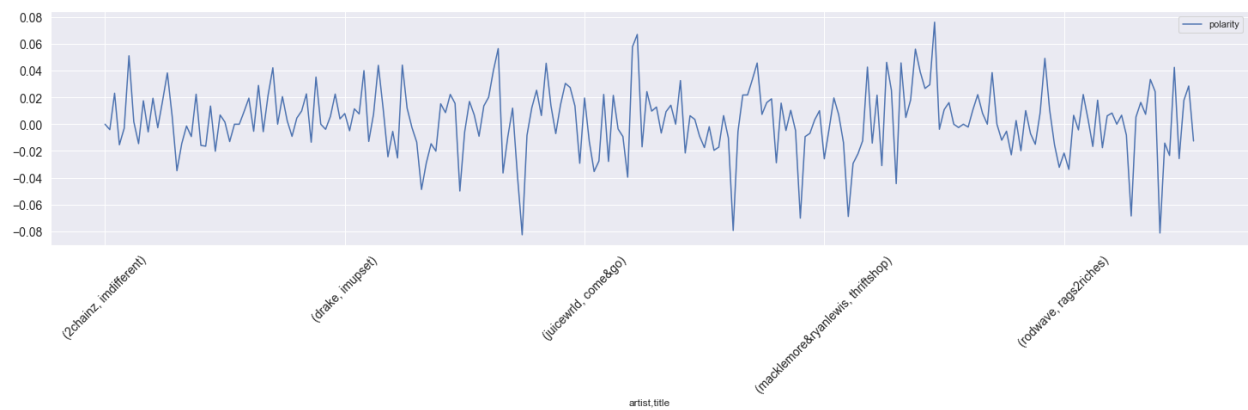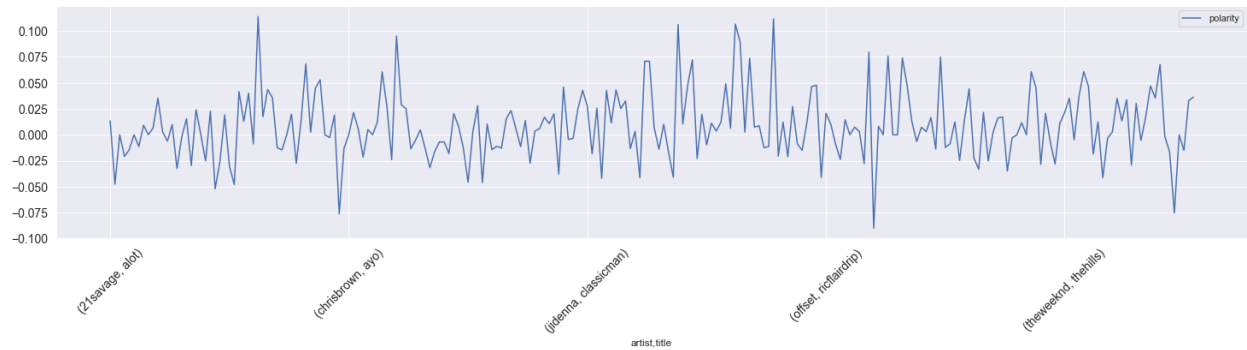
- Country

- Pop

- Rap
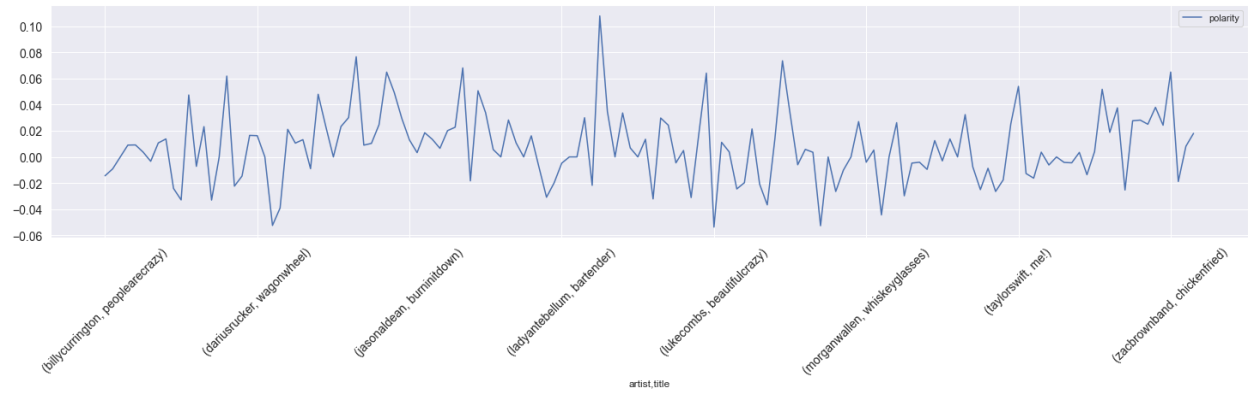
- R&B / Hip-Hop

We can see that the overall polarity of the songs per genre appear to be distributed across the 0 axis indicating the songs are both positive and negative. This does not display that one genre is exclusively associated as positive or negative. This is expected since various artists express different styles and convey different emotions so lyrics will be conveyed as both positive and negative. In addition, looking at the sentiment plots for
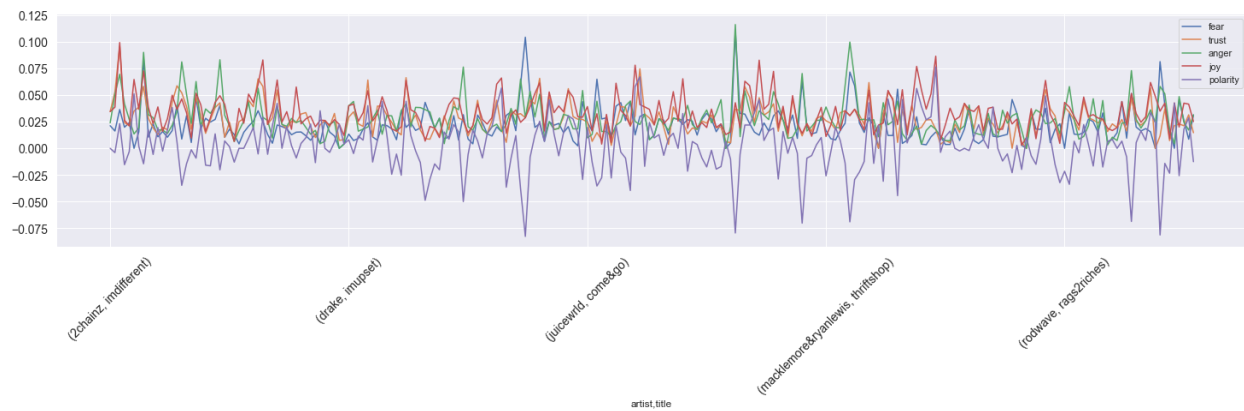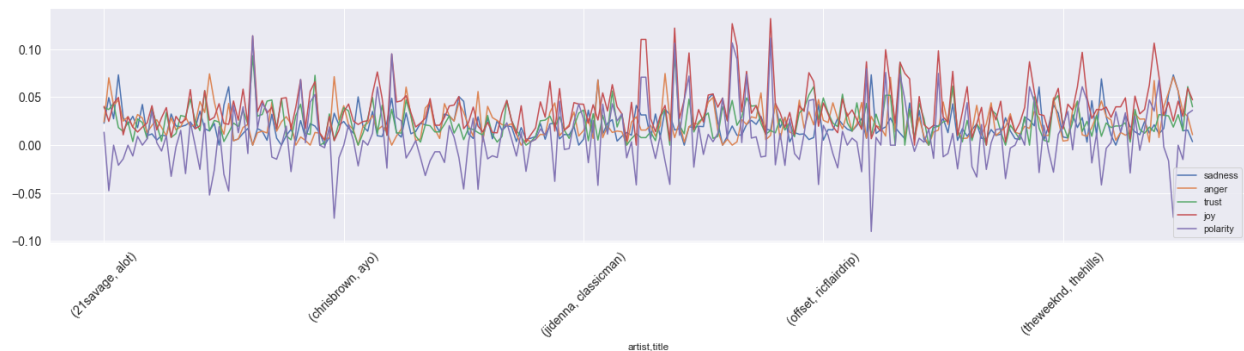
- Country

- Pop

- Rap
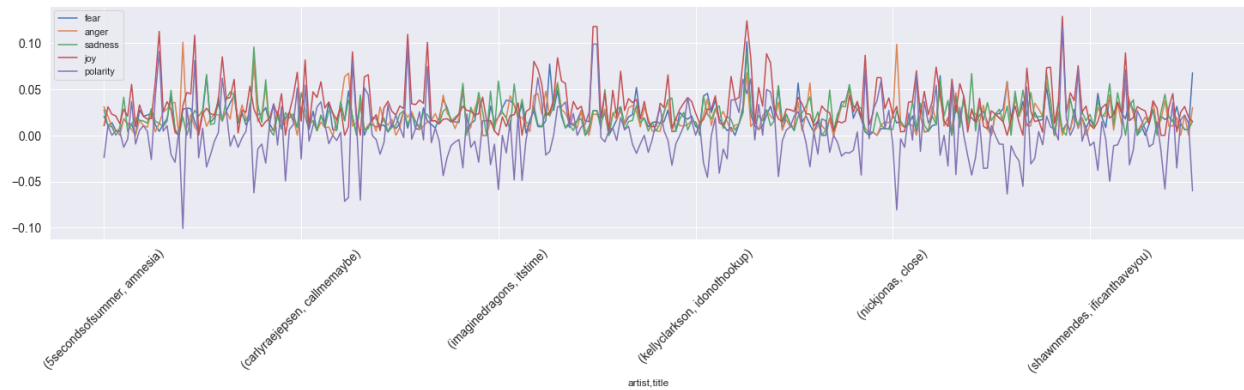
- R&B / Hip-Hop

we can see with the updated lexicons, the songs per genre are a range of various emotions. With large spikes across different emotions, no genre particularly stands out as 'joyous', 'angry', etc. The analysis has shown that there are similarities across genres. This can be attributed to that

lyrically, songs are written following similar patterns talking about life experiences, love, money and people with a limited vocabulary.

For future work, we would like to add various lexicons for songs within each genre for each year in the analysis in order to accurately perform a sentiment analysis. Music overall, seems to be also dependent on external factors such as the audio structure in addition to the lyrics. Although this analysis has shown that songs across genres share topics and similar sentiment, it does not show representation of the society and culture at time of release. This information will need to be addressed to further extend this project.

Resources:

https://towardsdatascience.com/song-lyrics-genius-api-dcc2819c29

https://genius.com/

https://www.billboard.com/