Project Summary – Musical Positivity Over Time

Danish James

Project Design

The intent of this project was to analyze music over the past few decades. There was an observation that the type of music that people listen to varies with respect to their current emotional state. In addition, there has been articles about music as a form of expression for current events. Thus, the purpose of this project was to conduct an analysis on music over the past few decades and build connections from the polarity of popular music to current events.

In order to determine this, I utilized the Billboard top 100 chart as my primary source of popular music, and the Genius.com API to pull lyrics for analysis. I initially wanted to scrape weekly chart data, however the scraping ended up taking too long to complete in a single sitting due to inefficiencies in my code, and I ended up utilizing the yearly chart data instead. My final dataset was the top 100 songs from 1964 to 2015.

Modeling for this project was performed utilizing k-means clustering. I expected love songs to dominate the top 100 very strongly, and therefore decided it would be more prudent to separate the top 100 by category, then conduct sentiment analysis on each cluster and see how things were affected over time. I used Doc2Vec to convert my song lyrics into element vectors that the k-means clustering could utilize and the sentiment analysis was conducted utilizing two external models, nltk Vader and TextBlob.

Tools

- Python
    - Pandas, Numpy
    - Sklearn, Jupyter Notebook, matplotlib, lyricgenius
    - Nltk Vader & TextBlob
- Microsoft Word, Microsoft Powerpoint

Data

The dataset was obtained via scraping the Billboard top 100 songs chart. I utilized the yearly charts which only started in 1964, and scraped up to 2015. I then populated this with lyrics that were scraped via the Genius API using the lyricgenius wrapper. Some songs did not have lyrics available on Genius, and I chose to remove those from the dataset. It did lead to certain years being slightly unbalanced, however there were very few removals overall and the set did not need any sampling. The final dataset was approximately 6,700 songs.

What I Would Do Differently Next Time

I struggled a lot with how to approach this problem. I initially felt that topic modeling was the appropriate method, however after giving it a try I found that the topics I received were very nonsensical and difficult to distinguish. A large part of this was due to the strange format in which musical lyrics were delivered. In order to address this, I put a lot of focus into text processing and trying to clean up the data as much as possible. This resulted in my having a lack of time down the stretch to properly wrap up my model. In addition, I eventually chose to disregard topic modeling entirely and go with a clustering route, which effectively made all of my previous work useless.

In approaching the same problem a second time, I would definitely try to restrict how much time I spent with text processing and try to put more time in on the modeling. As well, I would want to discuss my methodology with the instructors a bit more in-depth as I feel my understanding of what techniques would be appropriate was inadequate and giving myself more knowledge would have been very helpful in making sure I stayed on track.