Math 381 Project 2 Sentiment Analysis of Songs

Eunji Lee
University of Washington
elee0025@uw.edu

Yuan Qu
University of Washington
quyuan0522@yahoo.com

Arnav Dubey
University of Washington
arnavd@uw.edu

ABSTRACT

In this paper we used markov chains to model and explore the transition of sentiments in the lyrics of songs of polular Artists.

1. INTRODUCTION

For the project, we modeled using Markov chains and the existing 'sentiment analysis' r package to capture the sentiment characteristics of each artist' lyrics in their songs. We were interested in exploring how the transition of sentiments in an artist's song might differ from those of another artist. We also wanted to investigate the similarities and distinctions in the sentiments of the lyrics from different pop stars.

2. BACKGROUND

Applications of Markov chains are found in many areas of sports, biology, chemistry, games to linguistics. The concept of Markov chains fascinated our group, especially its role in linguistics. Markov chains are used in linguistics to usually analyze a sequence of letters or words to capture author's writing style and to generate random texts with the most probable letter or word sequences.

However, finding a model that fits Markov property was not easy because there were several limitations to apply Markov chains. First, the model should change its state from a discrete set of states, in a discrete set of times (each second, each day, etc.). Also, the probability of a current state only depends on the previous state and is not affected by any other prior states. We considered generating random texts by analyzing poems of a writer or lyrics of an artist but concluded that we would probably get meaningless sentences because Markov chain uses random, memoryless process when changing its states. Creating valid (at least grammatically correct) sentences would require us to study linguistics and include a large number of grammar rules. Most importantly, it would also be hard to quantify a success of the project with created texts. Hence, instead of using lyrics to create random texts, we thought it would be interesting to do a sentiment analysis on the lyrics for some artists.

Some interesting researches were done to analyze sentiments of song lyrics. Oudenne and Chasins used Natural Language Processing (NLP) to analyze and identify the emotional polarity of song lyrics. According to the paper, an increasing number of researchers have started to focus on lyrics to classify music, whereas they used to only focus on the audio features to analyze music. Sentiment analysis is popular in NLP, and sentiment analysis on lyrics is often used in machine learning field to classify sentiments of the lyrics. It was also found that lyrics play more significant roles in sentiment or mood analysis than audio features itself. Yet, most other problems creating a model for music with Markov chains were usually about audio features of music (Markov chains with rhythms, tempos as states). In our project, we are going to explore the transitions of sentiments in a song lyrics and observe if similar patterns of sentiment transitions are observed for each artist. Since sentiment analysis is already widely developed and used in the field and its usage is out of the scope of this project, we are going to implement the existing codes to get the sentiment scores of the lyrics.

2.1 Model

We used 'beautifulsoup' package in Python to scrape the information for all songs of each artist from the website and transformed them into CSV files. The above image showed what the file looks like. It contained information about album name, artist name, song title, duration, URI, full URL and the lyrics. The websites that we had used to obtain information were under 'Full_URL' list. The file shown above was only for Beyoncé's songs. In total, we obtained five CSV files since we would analyze five artists (ED Sheeran, Beyoncé, Bruno Mars, Coldplay, and Maroon5). Additionally, for each file, there are about 80 songs, except for Beyoncé, there are more than 200 songs. Thus, our data contains about 570 songs.

2.2 Data

We used 'beautifulsoup' package in Python to scrape the information for all songs of each artist from the website and transformed them into csv files. The above image shows what the file looks like. It contains information about album name, artist name, song title, duration, URI, full URL and the lyrics. The websites we obtaining information from are under 'Full_URL' list.The file shown above is only for Beyoncé's songs. In total, we obtained five csv files since we will analyze five artists (ED Sheeran, Beyoncé, Bruno Mars, Coldplay and Maroon5). Additionally, for each file, there are about 80 songs, except for Beyoncé, there are more than 200 songs. Thus, our data contains about 570 songs.

2.3 Assumptions

In order to create the model, we made following assumptions. First, the sentiment score for lyrics transition from sentence to sentence follows the nature of Markov chain. This means we assumed that the transition process changes discretely among a set of states (negative, neutral and positive). And the next state (sentence) only depends on the current state (sentence) and is unaffected by the state (sentences) of the system at earlier times. We would present how we obtained the sentiment data and converted data to the states that we would use later in the paper.

The second assumption was that the sentiment scores obtained from the 'sentiment analysis' r package are a good representation (estimate) of the sentiments of the lyrics. This was the basis of our analysis.

Our third assumption was that the lyrics does reflect the sentiments of the songs. This assumption was important because we could analyze the sentiment of songs by analyzing lyrics. In addition, this assumption was also based on the researches mentioned before that sentiment of lyrics has a significant role in identifying sentiments of a song as a whole.

2.4 Getting and processing the data

We got the sentiment data for the songs by using the 'sentiment analysis' r package. It is a very popular open source package for text sentiment analysis. The package entails three different dictionaries: Harvard-IV dictionary, Loughran-McDonald Financial dictionary, and QDAP dictionary. Each dictionary categorizes the sentiment of words slightly differently by various linguistics. More importantly, each dictionary uses different algorithms to return various form of results. Harvard-IV dictionary only returns positive and negative scores. Loughran-McDonald Financial dictionary returns positive, negative scores and uncertain words. QDAP dictionary returns positive, negative and 0 scores. Since we wanted positive, negative and neutral scores as our states in transition matrix, QDAP dictionary gave us the most desired result. Additionally, the 'sentiment analysis' package can take in a word as an input and returns a score for a single word. It can also take in the sentence as an input, and returns a sentiment score for the whole sentence. We chose to analyze sentence by sentence for a reason that we wanted a relatively large set of data points. If we chose to analyze word by word, it would have had very high variabilities. Therefore, we don't think such transition can capture the sentiment transition feature of a whole song more accurately than sentence by sentence. On the other hand, we could have inputted the whole lyrics of a song and got a single sentiment score. In this case, however, we would not be able to know any sentiment transition within the lyrics, which meant not enough variations and data points. This would result in difficulty and inaccurate to approximate the sentiment transition feature of an artist. Considering the appropriate level of variation, the proper amount of data points, and the accuracy of approximation, we choose to analyze sentence by sentence.

Moreover, one of the essential steps of constructing our model is to transform the sentiment scores to the states in the transition matrix. The data from the 'sentiment analysis' r package are continuous values within the range of -1 to 1. To classify the data into negative, positive and neutral states, we rescale the data by raising them to ??? power to magnify all the decimal values. For instance, the value of 0.125 will become 0.5 after the rescaling. Then we classify the values between -0.125 and 0.125 to be neutral (0 state), other negative values to be negative (-1 state) and other positive values to be positive (+1 state). We implement such method because the sentiment score for sentence tends to be neutral (A lot of words like 'a' and 'the' are neutral, and the algorithm from the package will average the score for each word to get the score for a sentence), but we want to increase the threshold of the data values being classified as neutral for better representing and emphasizing the unique features of an artist' lyrics. More specifically, we want the probability of being classified as neutral to be significantly less than the probability of being positive and negative. We calculated that for the interval [-0.125,0.125] as neutral, the probability of being neutral is 13% and the probability of being positive equals the probability of being negative, which is 43.5% (P(neutral) = 13% < P(negative) = P(positive) = 43.5%). This interval of classifying neutrality is also one of our assumptions which can be changed for adjustments and extensions of the model.

{EXAMPLE OF A TRANSITION MATRIX}

To construct the transition matrix, we now have the three states and sentiment scores being classified into the three states. Then we start to calculate the probabilities of each transition and put the data into the matrix. Indeed, We use the 'markovchain' r package to get the transition matrix. Specifically, for each song, we enter its lyrics as a vector into the function 'get_markov_fits', and each sentence as an element in the vector. Then it counts the number of times of each possible transition, for example, -1 to -1 and -1 to 1. And it divides the number by the total number of transitions to obtain the probability of certain transition. This is how we construct the transition matrix for each song.

Furthermore, we will compare the lyrics sentiment features from artist to artist by comparing a set of matrices to another set of matrices. We calculate the mean of the set of matrix and obtain the mean matrix. We also calculate the distance from one matrix to its mean matrix using the 2-norm function and get vectors of such distance for the set. The mean can represent the average transition of an artist' lyrics, but it loses plenty of important properties for each individual matrix. Thus, we also calculate the distance between each matrix to the mean, which helps us observe the variation of the lyrics sentiment. Both the mean matrix and the norm vector help us capture certain properties of an artist's lyrics as a whole and compare the similarities and differences between artists.

When processing the data, we notice that the number of songs for each artist is not the same. For example, Beyoncé has more than 200 songs while Bruno Mars only has 40 songs. But from a statistical perspective, 40 is not a small sample number. Since all other samples are more than 40, and we are trying to capture the sentiment characteristics of an artist lyrics as a whole, we conclude that unequal sample size for each artist does not affect much of the result.

Another noticeable fact is that after processing the original data, some songs are getting lost because they contained non-UTF-8 characters like ’*‘and’ˆ’. Based on the time constraints, it is not feasible to figure out why certain data points weren’t compatible with the package we were using and how to fix it. Besides, the number of songs being lost is about 5 for each artist, which is not too much and we still have a good amount of data to analyze. Thus, we choose to discard certain songs directly.

Based on the assumptions we make and our method to obtain and process the data, our model will capture sentiment transition from sentence to sentence to analyze the sentiment properties of the lyrics for different artists. On the other hand, we left out the audio part of the song since we only analyze the lyrics. As mentioned earlier, when taking the music into account. the sentiment of a song might vary. Moreover, we do not consider the sentiment transition from verse to verse or the transition from three sentences to the following three sentences (or four and five). Processing the data differently might change the result, too.

3. CODE

Nullam semper imperdiet orci, at lacinia est aliquet et. Sed justo nibh, aliquet et velit at, pharetra consequat velit. Nullam nec ligula sagittis, adipiscing nisl sed, varius massa. Mauris quam ante, aliquet a nunc et, faucibus imperdiet libero. Suspendisse odio tortor, bibendum vel semper sit amet, euismod ac ante. Nunc nec dignissim turpis, ac blandit massa. Donec auctor massa ac vestibulum aliquam. Fusce auctor dictum lobortis. Vivamus tortor augue, convallis quis augue sit amet, laoreet tristique quam. Donec id volutpat orci. Suspendisse at mi vel elit accumsan porta ac ut diam. Nulla ut dapibus quam.

4. RESULTS

From the set of transition matrices from each artist, we calculated the mean matrices by calculating the mean of each element of matrices from same state transitions. For example, we calculated the mean of all transition probabilities from -1 to -1 and placed the mean into the same position in a new mean matrix. Then, in order to analyze how far every transition matrix is from its mean transition matrix (i.e. the spread) of its artist, we calculated the “distances” between each matrix and the mean matrix. The distances are calculated such that we find the norm 2 of difference between each transition matrices and the calculated mean matrix. The same process was repeated to calculate mean matrices and the list of distances for each artist. The following is the mean transition matrices and the list of distances for each artist

{Show mean matrices and list of distances for each artist. Or at least mention in the previous sentence that it’s in the appendix.}

First, we will compare mean transition matrices for all artists. Probabilities of transitions from state to state (i.e. sentiment score to sentiment score) show similar patterns for all artists overall. The transition of any sentiment scores (negative, positive, and neutral) to neutral score is most probable, followed by the transition to positive score, then transition to negative score. We found the same pattern observed for all artists interesting. In addition, generally, the transition from neutral state to neutral state is most probable for most artists, except Coldplay and Ed Sheeran. In their cases, the transition from negative to neutral was most common. Although our interval for a neutral score when we convert continuous sentiment scores to discrete sentiment scores was pretty narrow (i.e. (-0.125, 0.125), the probability of transitioning from neutral state or transitioning to neutral state was still high. Even though the mean matrix is a good mathematical tool to observe the overall set of transition matrices for each artist, we could not observe how each transition matrices are related to its mean matrix.

Thus, we drew a box plot of the collection of distances for each artist in order to analyze how each transition matrices are distributed from the mean for each artist.

{Box Plot of distances}

```
From the box plot, we observed that the distribution of dis
```

We also constructed the density plot for distances of each artist to observe if similar distributions are found. Here is the density plot of 5 artists.

{Density Plot overlap}

In order to test if the distributions of distances for all artists are similar, we conducted an analysis of variance (ANOVA) test. As a side note, we are not comparing mean transition matrices among artists. we are going to compare mean distances for all artists. We will construct the typical hypothesis testing of the common alpha value of 0.05. Then, hypothesis testing is set as follows:

Let μ_1 be the mean distances for Ed Sheeran, Let μ_2 be the mean distances for Beyonce Let μ_3 be the mean distances for Coldplay Let μ_4 be the mean distances for Maroon 5 Let μ_5 be the mean distances for Bruno Mars

H_{0}: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ H_{a}: At least two of the means are different.

We constructed the test in R code. From the analysis, we found out that the p-value is 0.0236, where p-value indicates the probability of observing the outcome given the null hypothesis is true. In this case, the probability of mistakenly rejecting our claim that at least two of the mean distances are different is 0.0236. Since we set 95% hypothesis testing (with an alpha value of 0.05), it means we are willing to reject the null hypothesis (H_{0}), when the p-value is less than 0.05. Hence, we can reject the null hypothesis and conclude that at least two of the mean distances of artists are different. If we had constructed the hypothesis with an alpha value of 0.01, another typical value of alpha, we would have failed to reject the null hypothesis (since p-value of 0.0236 > 0.01). Thus, we could have concluded that we failed to conclude that at least two of the mean distances are different. We are not very confident that if transition matrices for all five artists (Ed Sheeran, Beyoncé, Coldplay, Maroon 5 and Bruno Mars) are similarly distributed from its mean transition matrix because we were only able to compare mean distances between transition matrices and mean matrix for each artist. Also because the answer varies depending on the alpha value. We are 95% that at least two of the mean distances are different, but we are not 99% confident of this statement that at least two of the mean distances are different. Furthermore, it's hard to figure out which of the mean distances are different when we simply conduct a hypothesis test with an alpha of 0.05. However, we are pretty confident that there are lots of similarities in their distribution based on hypothesis testings and distribution graphs of distances.

5. ADJUSTMENTS AND EXTENSIONS

Two adjustments: (change of assumptions or choices, discuss what happens when you changes these assumptions and choices.)

1. Changing the interval of neutral.
2. Instead of using the norm, calculate the standard deviation of each matrix.
3. Changing the way to calculate the norm.
4. Process the data paragraph by paragraph or three sentences by three sentences?

There are several methods to explore transition matrices of sentiment scores for all song lyrics for five artists. When we calculate "distances" between transition matrices to mean matrix, we used 2- norm to calculate them. But instead, we will use a different algorithm for getting distances, which is to use infinity norm instead. Infinity norm of matrices will give maximum row sum norm.

6. CONCLUSION

The most time-consuming part of our model was to process the data. We needed to make specific choice about using the dictionary, and think of reasonable methods to convert continuous sentiment scores to discrete states and find feasible approach to compare a set of matrix to another. These insights, thinking process and the final approaches were essential for developing the complete model.

To conclude, the sentiment transition within a song might not help us detect the unique style of an artist's lyrics because by comparing the sentiments for one artist to another, their transition distributions were not significantly different. But at the same time, by comparison, we could observe their similarities and some worth-noticing distinctions about certain transition states for an artist. Through further exploring their background, we can hypothesize and conduct further exploration about the factors affecting such differences.

7. REFERENCES