**Madison Yonash**

**Susan Nunez**

**Jaylyn Leatherwood**

**Nicolas Garay**

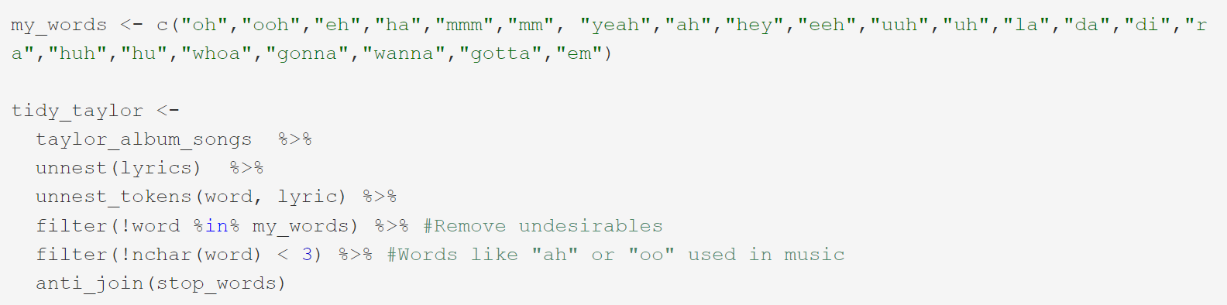
**Data Mining and Text Mining Final Project: Taylor Swift Songs**

**Methodology**

To process the data, principal component analysis and sentiment analysis were performed. This was done to both examine the way that a song could be defined based on features of the music, as well as to discover more insight into the nature of the lyrics. R programming was utilized for the analysis of the dataset, with various open-source packages available on CRAN providing useful support.

*Sentiment Analysis*

The sentiment analysis for this data set focuses on the lyrics column of the taylor\_album\_songs dataset. To get the data ready for text analysis, we must take some preprocessing steps. Each observation in the original data set represents one song from Swift’s discography. In the lyrics column, there are nested tibbles that contain the lines of each song. We must first unnest this and remove stop words as well as unnecessary/small words and sounds that are included. This allows us to perform much more meaningful analysis. For example, if this is not done, the song “I Did Something Bad” produces the most frequent word as “di”, which isn’t actually a word, but she stutters the word “did” during the chorus. The following code snippet shows the steps to prepare the dataset for text analysis:



We focus on two lexicons for this sentiment analysis: NRC Emotion and Bing. NRC Emotion Lexicon is a list of English words, and these words are associated with eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. It also contains two sentiments: negative and positive. Bing Sentiment manually classifies text by looking at the difference between the positivity and negativity of the tokens. These differ in some of the words that are included as well as the categorization. By utilizing different lexicons, we can view the lyrics from different perspectives.

To perform the sentiment analysis, we can utilize different manners of aggregating the lyrics and analysis. This is useful as different insights can be derived by viewing the different layers of the analysis. While looking at the emotions of her discography, we can get an idea of aggregate sentiments, but this can be further broken down into other subcategories such as album, track number, or song. Utilizing these different levels of comparison and different visual techniques can give a better understanding of the sentiment analysis that is being performed.

Additionally, packages such as the tidyverse, tidytext, radarchart, and wordcloud provide useful tools to create better visualizations and make more effective analysis. This allows us to use premade functions and commands that avoid the need to manually label text sentiment or display data with inefficient displays.

The exploratory data analysis (EDA) for this dataset was focused on four main columns, energy, danceability, acousticness, and tempo of the songs for the taylor\_album\_songs dataset. To get the histogram charts had to use packages like tidyverse, color, and ggplot2. Using ggplot2 allowed us to create the histograms while the color package allowed us to put the outlines in color depending on the album name.

**Results and Discussion**

*Sentiment Analysis*

After the preprocessing steps, we are able to perform sentiment analysis with our now tidy data set. The first piece of analysis is simply looking at the most common words for each song. This is done by counting the number of words in each song, selecting the most common word of each song, and then, for sake of demonstration, showing the songs with the highest counts.

A screenshot of a computer

Description automatically generated

Some of the most common words of songs appear in the title. If you’ve ever listened to her pop songs like “Red (Taylor’s Version)” or “Shake It Off”, you’d be familiar with the catchy choruses that contain repetition of the title.

To create a visualization of word usage of her discography, a word cloud can be used to compare frequency of words. Doing so with her 150 most used words in all her songs, the following output is produced:

A circle of words

Description automatically generated

The most commonly used word is “love,” with “time” and “baby” also following. The topics of Swift’s songs range from heartbreak, to friendship, to the occasional murder-mystery, but she is well-known for her songs about relationships and love. This word cloud shows that she does in fact say love quite a bit.

Moving on to the actual sentiment analysis, there are a range of different manners to perform sentiment analysis. We can use different lexicons to examine the same data in different manners. We will start with the NRC Emotion Lexicon. Taking her discography, we perform NRC Sentiment Analysis on her lyrics to produce the following output:

A graph showing a number of different shades of brown

Description automatically generated

Here we see that this lexicon categorizes more words positively than negatively, and joy and anticipation being the most common sentiments. She shows the least amount of surprise and disgust in her songs according to the NRC Emotion Lexicon.

Utilizing different methods of sentiment analysis may produce different results in the analysis of Swift’s discography. Another lexicon that is used is Bing Sentiment. Utilizing the same process as before, we can use this lexicon to see how her discography is categorized using Bing Sentiment:

A graph showing a number of words

Description automatically generated with medium confidence

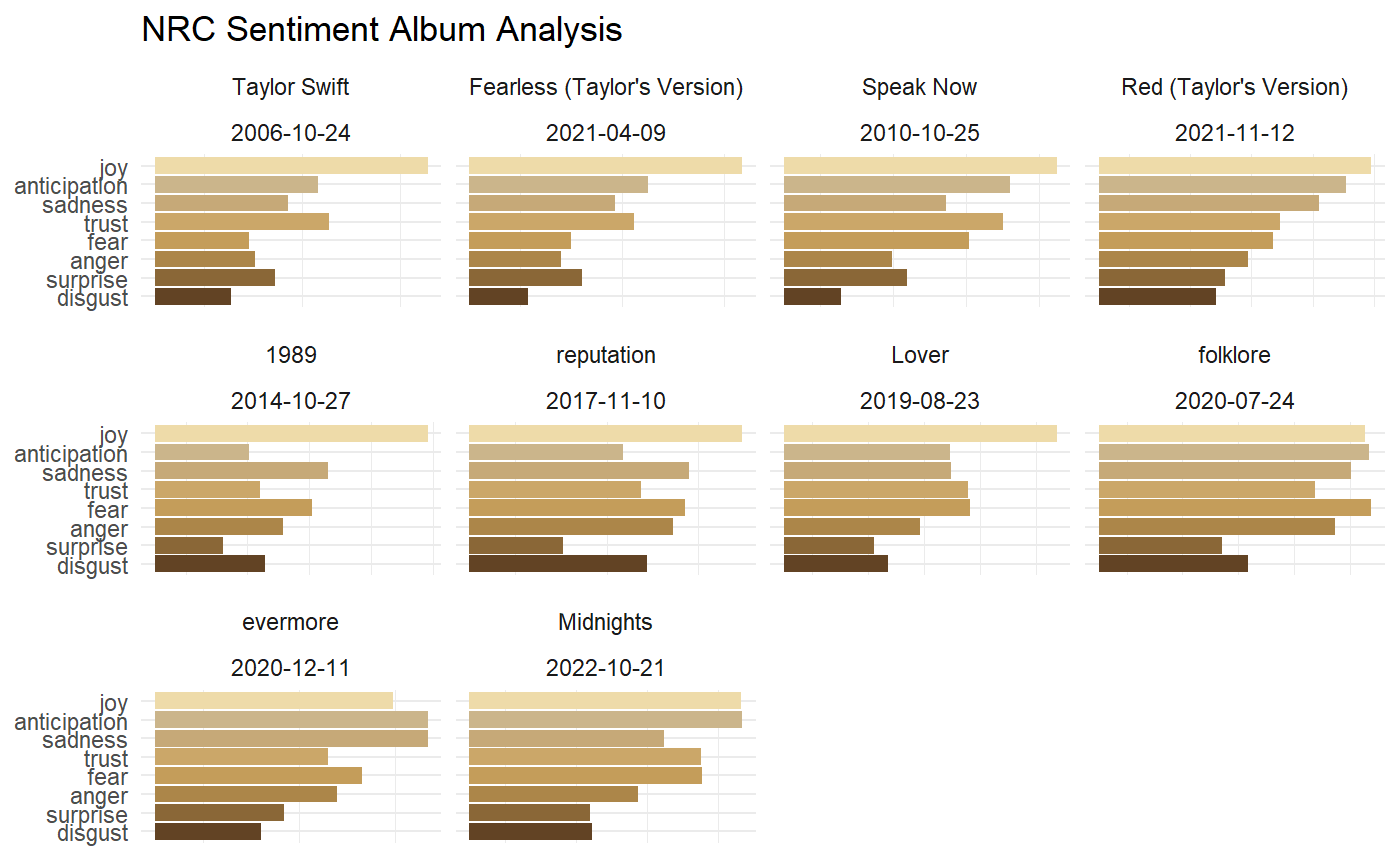
Unlike that of the NRC Emotion Lexicon, we see that Bing categorizes most of Swift’s discography as negative.

Continuing with the use of the Bing Sentiment, we will see what percentage of each album is positive. Taylor Swift’s discography ranges from her first self-titled album to her recently released Midnights, and we can compare the overall sentiment of each album:

A graph of different shades of brown

Description automatically generated

Her most positive albums are her Lover and Fearless (Taylor’s Version), while the ones that have the least positive sentiment are folklore and 1989. We can further examine the composition of the sentiment of the different albums using NRC Sentiment analysis. Here, we can see the different sentiments represented in columns, with each bar representing a different emotion. Here, we see the different levels of joy, anticipation, and other emotions, and how they vary between her different eras.



Continuing with the trend of analyzing the sentiment of different albums, we can utilize a radar chart and NRC Sentiment analysis to obtain a different point of view of how emotions compare between the different albums. In R, we can use the radarchart library to create an interactive radar chart where each album has its own points across each of the emotions.

A colorful diagram of a person's face

Description automatically generated with medium confidence

The chart, with all albums displayed, is a bit clustered and difficult to interpret. Thankfully, the functionality in R allows us to select which albums we wish to display. To demonstrate this, we can compare folklore and evermore, which are considered “sister albums” as both were written and released during the height of the COVID-19 pandemic, with little advertisement.

A diagram of a different color

Description automatically generated with medium confidence

Between the two albums, evermore features more tokens that portray sadness and anticipation, while folklore features more fear and anger. The radar chart is useful for quickly selecting different albums for analysis without needing to rerun the code.

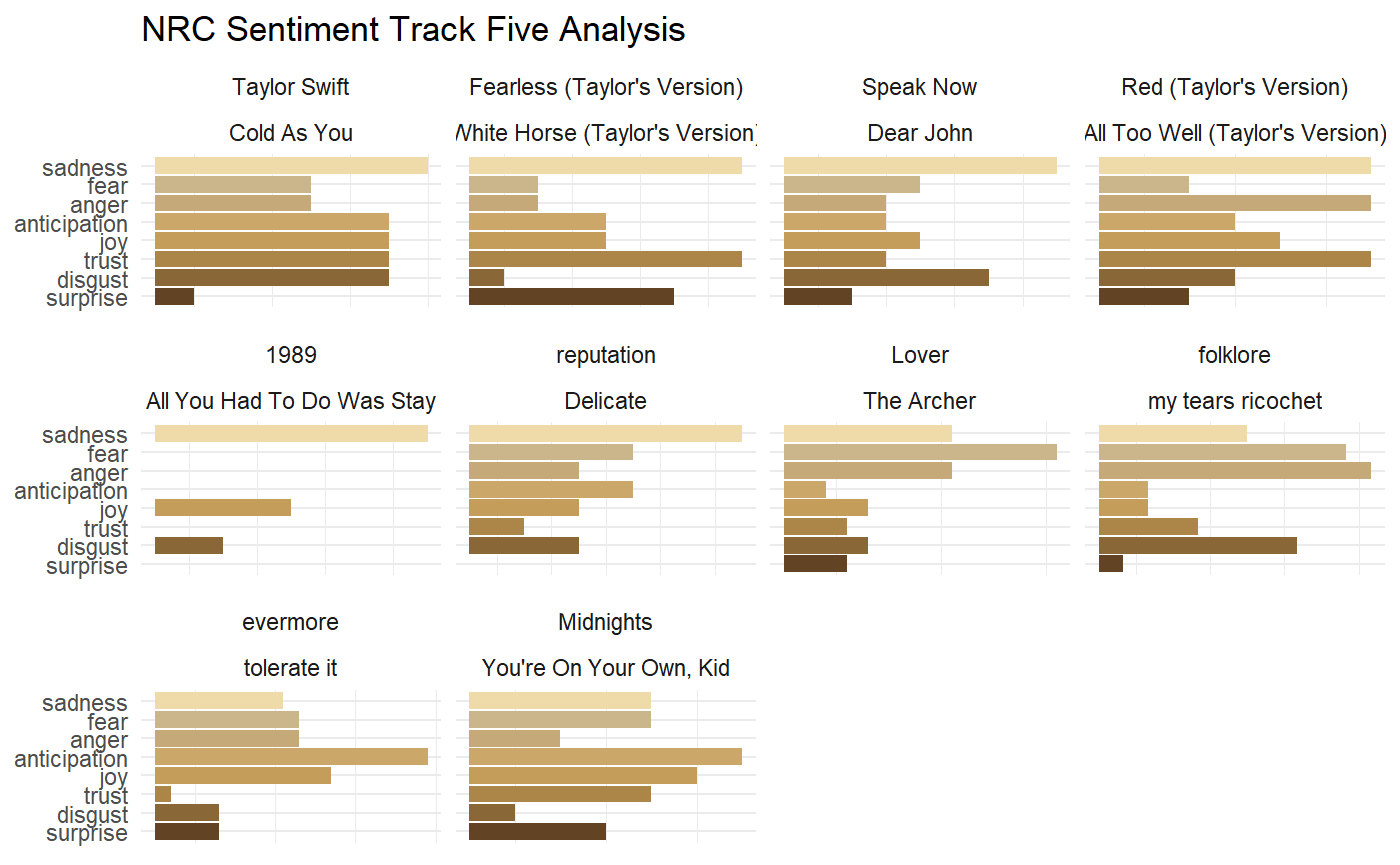
Continuing use of the NRC Emotion Lexicon, we can analyze Taylor’s discography at different levels. Creating a faceted plot of the words in songs with sentiments can give us an idea of why songs, albums, or even her discography are labeled the way that they are. It also provides further insight into what makes the songs so positive or negative. As an example, let’s look of one of her most popular sad songs, All Too Well (Taylor’s Version).

A chart with brown squares

Description automatically generated

All Too Well discusses how a beautiful and passionate relationship ended in a painful breakup. With words like promise and honest, we see words that are related to the trust that was broken, while sweet and masterpiece reference the feeling in relationship before the breakup. Utilizing a chart that shows the categorization of different words in the song lyrics allows us to understand why a piece of text is categorized the way it is. Although the overall tone may be negative, there are other pieces that make the composition more complex and share a story of both positive and negative emotions.

All Too Well is the fifth track on the album Red (Taylor’s Version). Swift is known for putting some of her most emotional and sad tracks in the fifth track in her different albums. Many fans of Swift’s music attempt to rank which are the saddest of these tracks. We can attempt to do this using sentiment analysis. Additionally, looking at the how the different emotions are portrayed within the songs is an interesting way to understand what the lyrics are conveying. Knowing this, we can see the sentiments of text of the different track fives of her albums.



Across the board, we see that there is quite a bit of sadness and fear. Compared to the sentiment of her whole discography, which had the highest number of tokens expressing joy, these tracks are categorically more negative in their emotions.