Final-Project-Draft

Chris Moua

5/6/2022

Introduction

What

What this project covers is a text analysis of Lana Del Rey song lyrics in her most critically acclaimed album *Norman Fucking Rockwell!* (*NFR!*). The goal of this project will uncover a sentiment analysis of Lana's lyrics to paint a picture of her album's artistry and answer the question: **"How does the emotional/sentimental state of the album contribute to *NFR!*'s artistic and cultural commentary?"** That is the question I am attempting to answer, and results from the text analysis will inform a final opinion in the conclusion.

Why

I chose this project scope for a few reasons:

1) In the field of marketing, understanding the consumer's emotional psychology is important in constructing user journeys that help guide and inform sales and marketing campaign. I want to explore that subject but with a personal muse I find interesting. Doing a text to sentiment analysis will help me accomplish this.

2) It is also in my interest to apply a statistical lens to something that is more categorical in nature. That will stretch my statistics understanding while keeping it within a relatively comfortable understanding of the topic.

How

First off, I used a lyric website (lyricfind.com) to generate the lyrics. Then I manually collected data, with each lyric line being an observation. Next, I used R and GitHub to organize, explore, and summarize the data. Finally, this project incorporated concepts on data summary, graphical representation, and regression to analyze the data.

Body

Why is it important?

Text analysis is important because it is useful "method for turning large amounts of unstructured data into something that can be understood and analysed." This method helps to find meaning out of written communications, like song lyrics or a series of tweets from consumers. Analyzing that text data and uncovering

sentiment can help businesses better understand the user journey and design an experience for customer conversion.

Each row of text will be associated with a one or more of the six main sentiments from The Sentiment Wheel (see Figure 1). The main six are at the inner center of the circle: Sad, Mad, Scared, Peaceful, Powerful, and Peaceful.

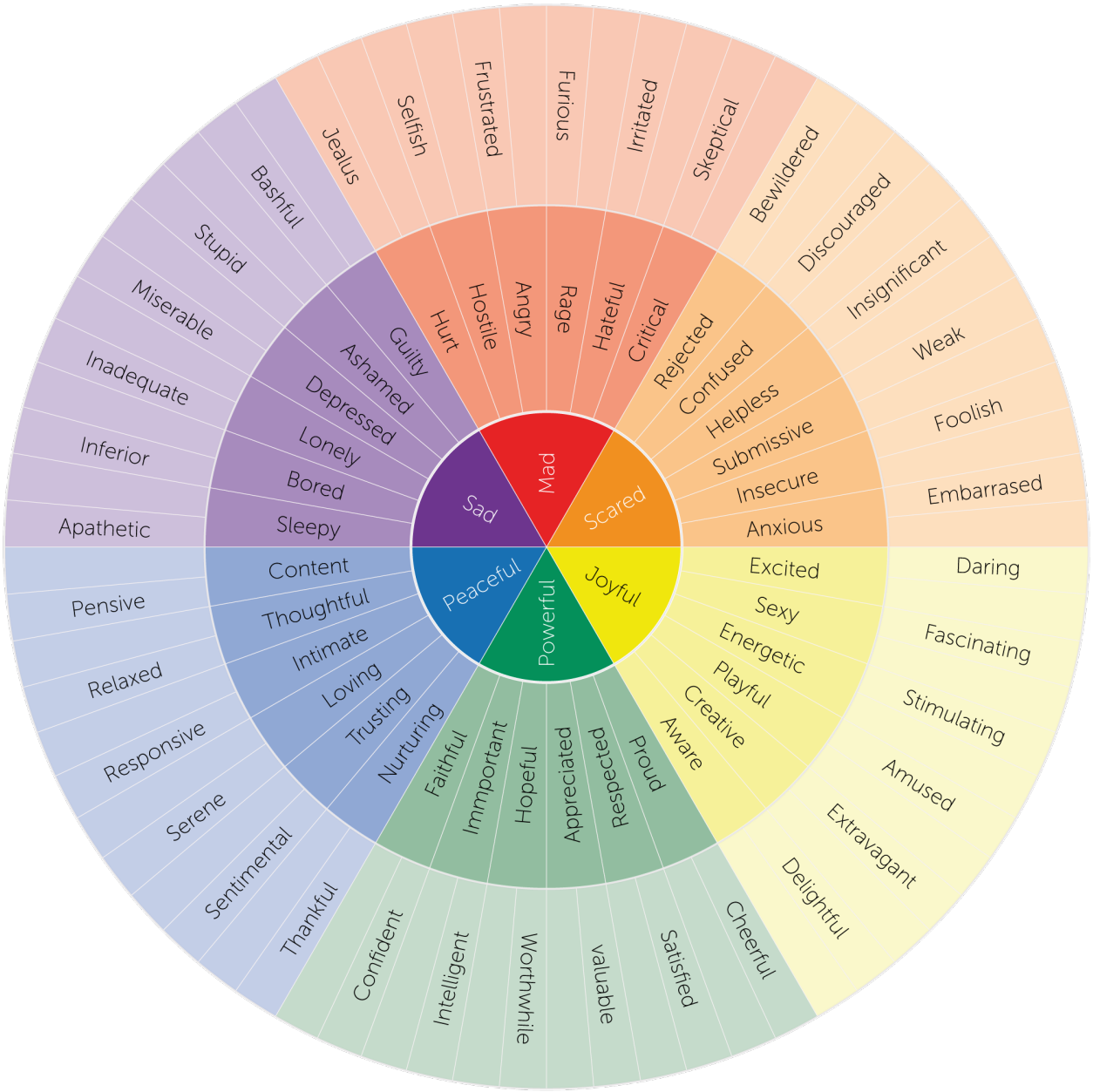


Figure 1: The Sentiment Wheel

Problems, Challenges, and Resolutions

I've encountered a few challenges in exploring this project:

1) How to structure the unstructured text data?

I initially attempted to structure the data with tidytext, using the 'janeaustinr' tutorial package as a guide. However, that file already exists in a package that can be brought over to R easily. My challenge was with figuring out how to "package" the album lyrics. My original intent was to use tidytext and diplyr to structure and clean the text. However, upon more research and discussion with peers, the method to do text analysis will require technical skills and time commitment outside the scope of this project. As such, this project will use Excel and its .csv file to organize the text data.

2) How to assign sentiment to the text?

Another challenge was figuring out how to assign the sentiment to the lines of text. At the moment, I am unsure of the approach. However, the project instead conducted a manual qualitative assessment of the lyrics and assigned sentiment based on word association. For example, if the lyric mentioned words like "happy", "love" or "party", that lyric would be assigned a "Joyful" sentiment. Thus, this project recognizes doing so adds a layer of personal bias. While I would like to avoid my own personal bias when assigning/inferring sentiment to the lyrics, but that is be a limitation in this project. In an ideal approach, a script or software would be able to scrub the lyrics.

3) Open issue:

The project had to scale back and focus on one album only. Initially, I did not realize the technical learning curve that comes with text analysis and the methodology needed for it. However, through this project I am learning that method and taking the time to be methodical with my project. The project's output may not be as polished as I had initially hoped, but regardless of the outcome, I do feel it will advance my knowledge and interest in text/sentiment analysis.

Topics from Class

Topics 1 and 2: R Markdown to summarize data, and GitHub to develop repository.

R Markdown is used to import the data and summarize it. With GitHub, this project is available for others to view. Because this is a topic that touches pop culture, it may also be of interest to those outside of statistics. GitHub's public settings will allow for me to share these findings and final opinion with other individuals who have an interest in Lana Del Rey.

```
library(readr)
lanadelrey<-read.csv("lanadelrey.csv")
```

```
dim(lanadelrey)
```

```
## [1] 643  10
```

```
names(table(lanadelrey$Song))
```

```
##  [1] "Bartender"
##  [2] "California"
##  [3] "Cinnamon Girl"
```

```
##  [4] "Doin Time"
##  [5] "Fuck It I Love You"
##  [6] "Happiness Is A Butterfly"
##  [7] "Hope Is A Dangerous Thing For A Woman Like Me To Have But I have It"
##  [8] "How to Disappear"
##  [9] "Love Song"
## [10] "Mariners Apartment Complex"
## [11] "Norman Fucking Rockwell"
## [12] "The Greatest"
## [13] "The Next Best American Record"
## [14] "Venice Bitch"
```

```
names(lanadelrey)
```

```
##  [1] "Lyric"        "Song"         "Track.Number" "Album"        "Sad"
##  [6] "Mad"          "Scared"       "Peaceful"     "Powerful"     "Joyful"
```

As show, there are 643 observations - these will represent each line of lyric spanning across the 14 songs in *Norman Fucking Rockwell!*. There are 10 variables in this data set; however, since this is a sentiment analysis, the applicable variables the project will analyze are the six emotions: Sad, Mad, Scared, Peaceful, Powerful, and Joyful. According to The Sentiment Wheel, while these are the main emotions, they contain a multitude of other more complex emotions. For example, feelings of "rage" or "jealousy" can be classified as "Mad."

```
typeof(lanadelrey$Sad)
```

```
## [1] "integer"
```

These are also categorical, non-ordinal variables. Binary code will represent whether any of the sentiments exist in a line of lyric. It is also possible a line of lyric may contain more than one sentiment. "1" will represent the presence of the sentiment, and "0" will represent the absence. Now, we will look at the numerical frequency of each sentiment.

```
sum(lanadelrey$Sad)
```

```
## [1] 123
```

```
sum(lanadelrey$Mad)
```

```
## [1] 33
```

```
sum(lanadelrey$Scared)
```

```
## [1] 68
```

```
sum(lanadelrey$Peaceful)
```

```
## [1] 78
```

```
sum(lanadelrey$Powerful)
```

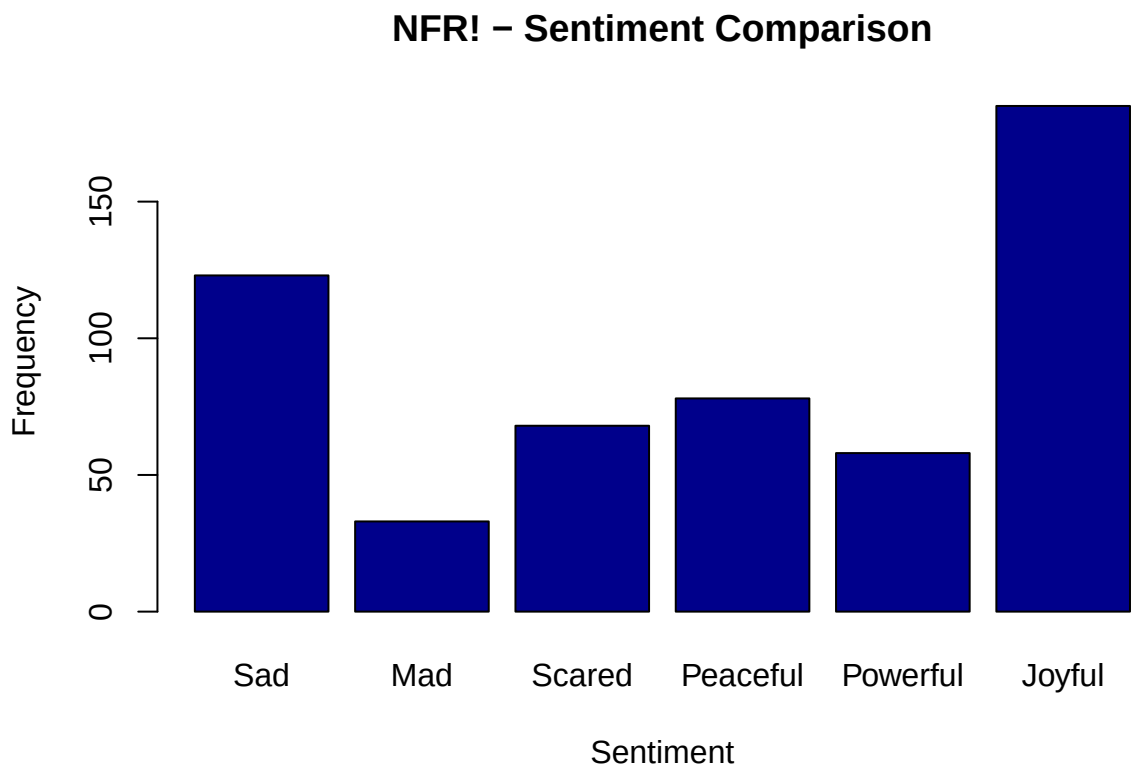
```
## [1] 58
```

```
sum(lanadelrey$Joyful)
```

```
## [1] 185
```

```
Sentiments<-c(123, 33, 68, 78, 58, 185)
```

```
barplot(Sentiments,
        main = "NFR! - Sentiment Comparison",
        xlab = "Sentiment",
        ylab = "Frequency",
       names.arg = c("Sad", "Mad", "Scared", "Peaceful", "Powerful", "Joyful"),
      col = "darkblue",
         horiz = FALSE)
```



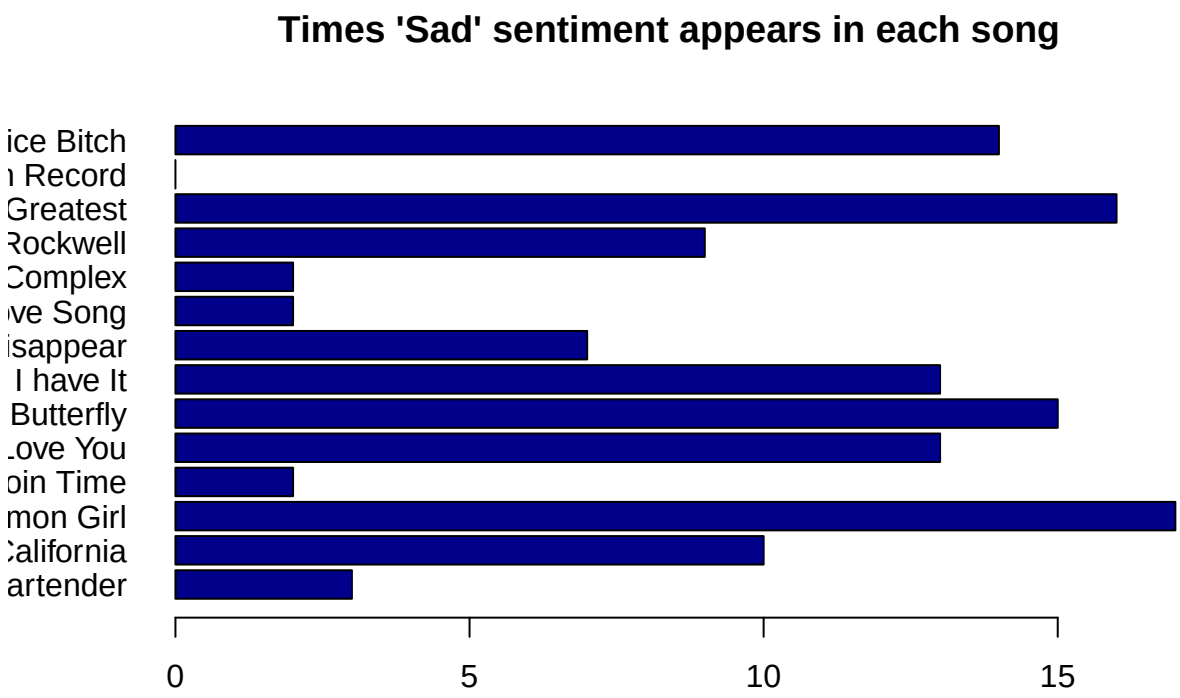
The frequency of sentiment on *NFR!* is as follows: Sad (123), Mad (33), Scared (68), Peaceful (78), Powerful (58), and Joyful (185).

Topic 3: Distribution of Data

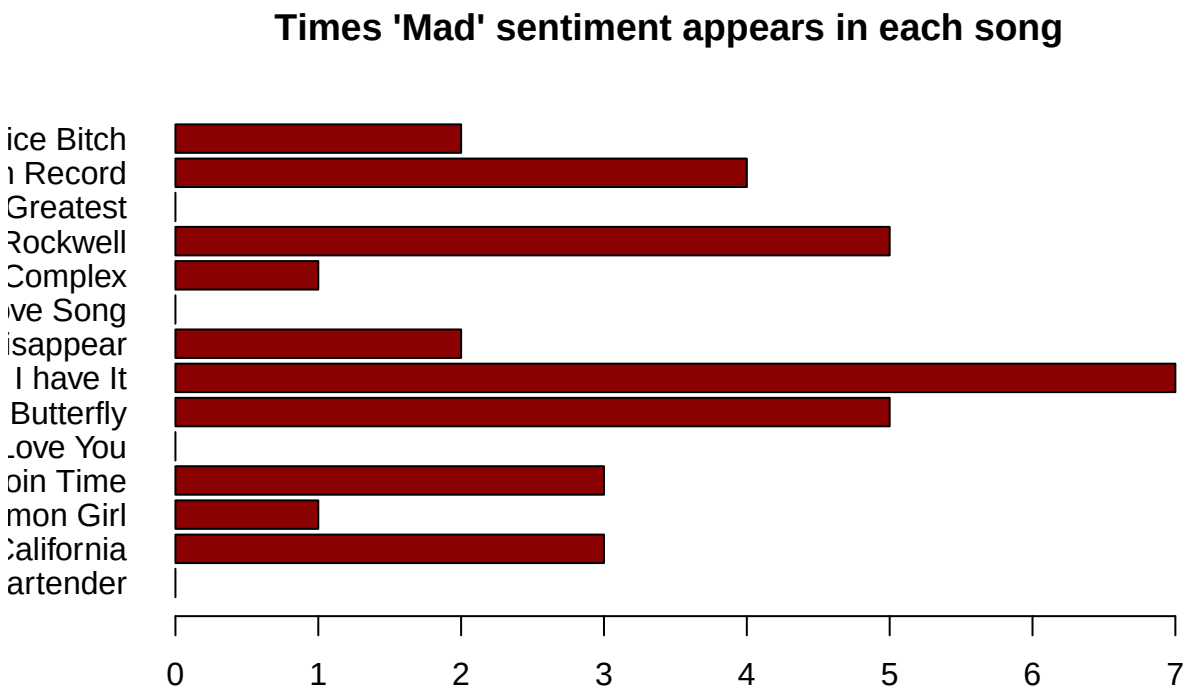
The text analysis relates to Chapter 2 from the class textbook on Summarizing Data, specifically categorical data. Although the data is not numerical, the project will still look at the "distribution" of sentiment by looking at sentiment frequency.

```
sadtable<-table(lanadelrey$Song, lanadelrey$Sad)
madtable<-table(lanadelrey$Song, lanadelrey$Mad)
scaredtable<-table(lanadelrey$Song, lanadelrey$Scared)
peacefultable<-table(lanadelrey$Song, lanadelrey$Peaceful)
powerfultable<-table(lanadelrey$Song, lanadelrey$Powerful)
joyfultable<-table(lanadelrey$Song, lanadelrey$Joyful)
```

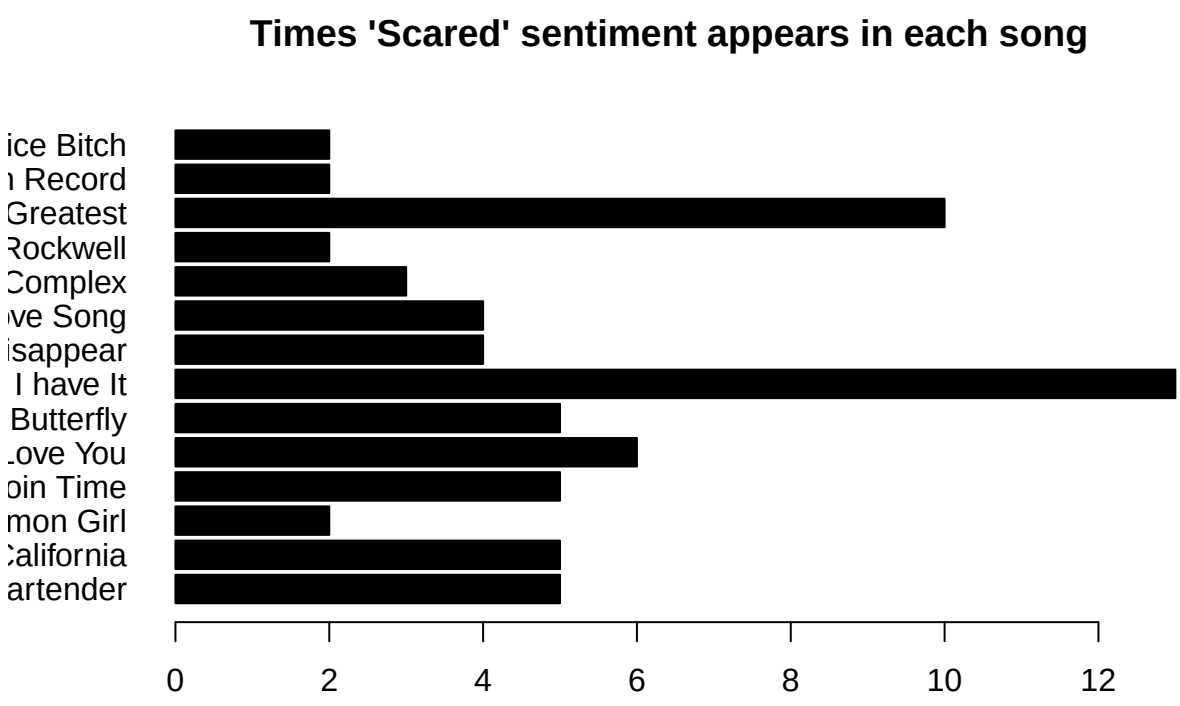
```
barplot(sadtable[,2],
        main = "Times 'Sad' sentiment appears in each song",
        horiz = TRUE,
        col = "darkblue",
        las = 1)
```



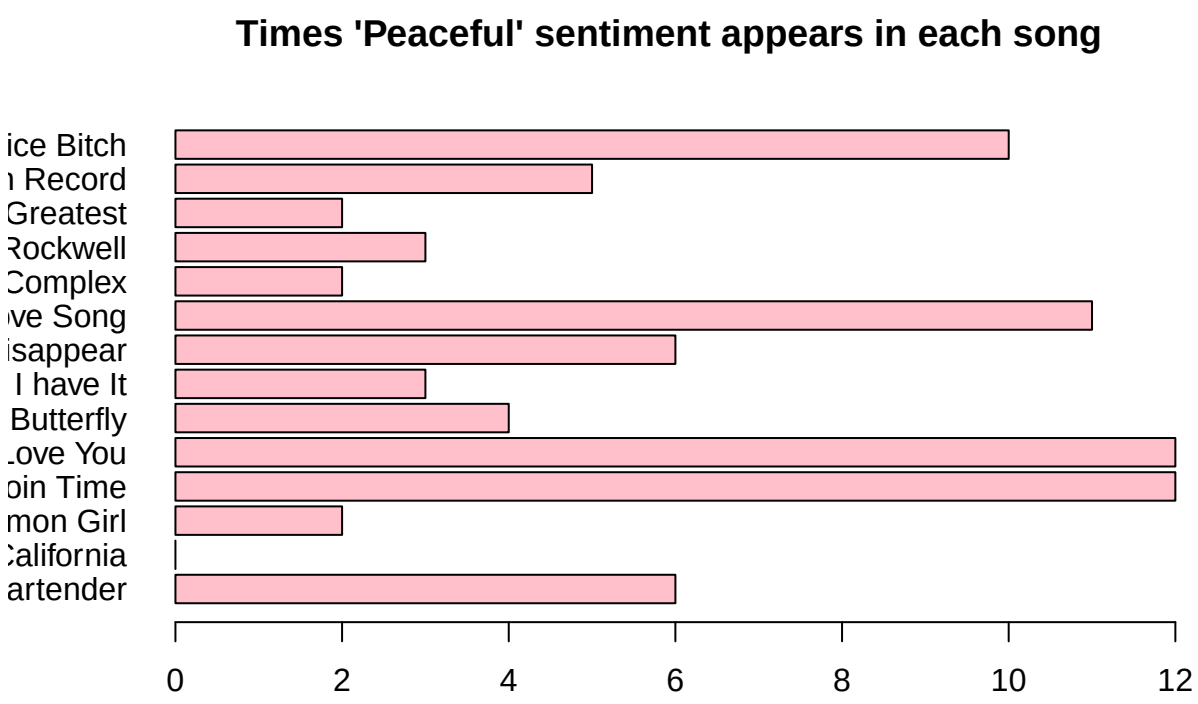

```
barplot(madtable[,2],
        main = "Times 'Mad' sentiment appears in each song",
        horiz = TRUE,
        col = "darkred",
        las = 1)
```



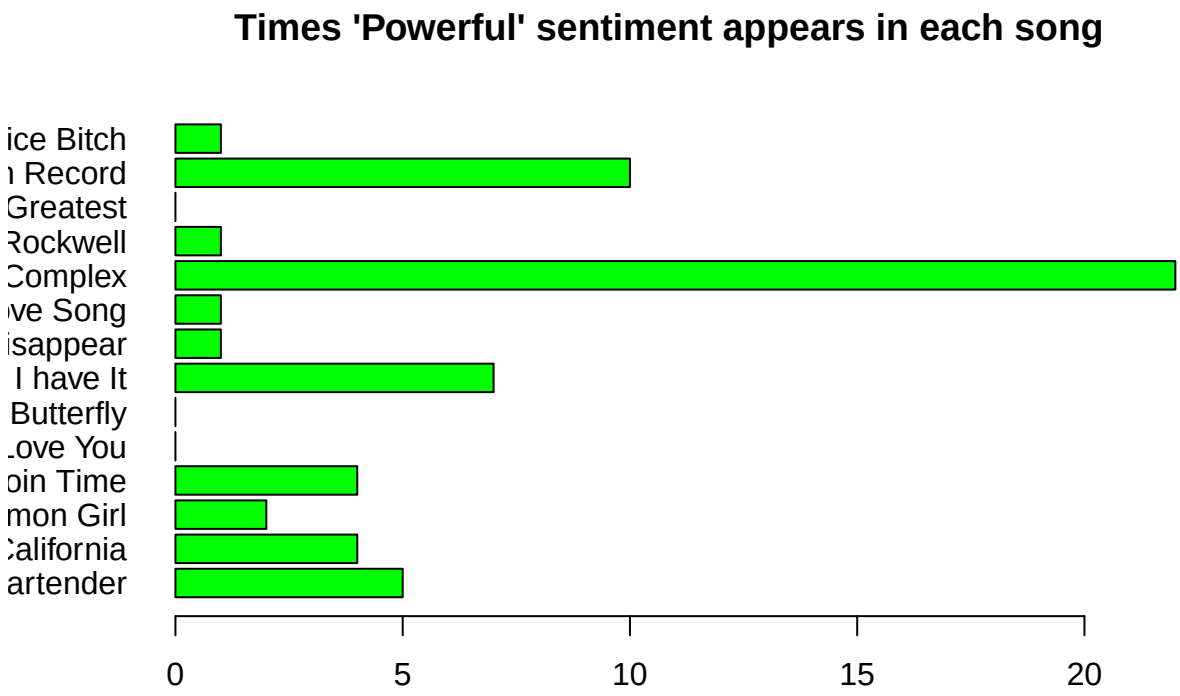

```
barplot(scaredtable[,2],
        main = "Times 'Scared' sentiment appears in each song",
        horiz = TRUE,
        col = "black",
        las = 1)
```



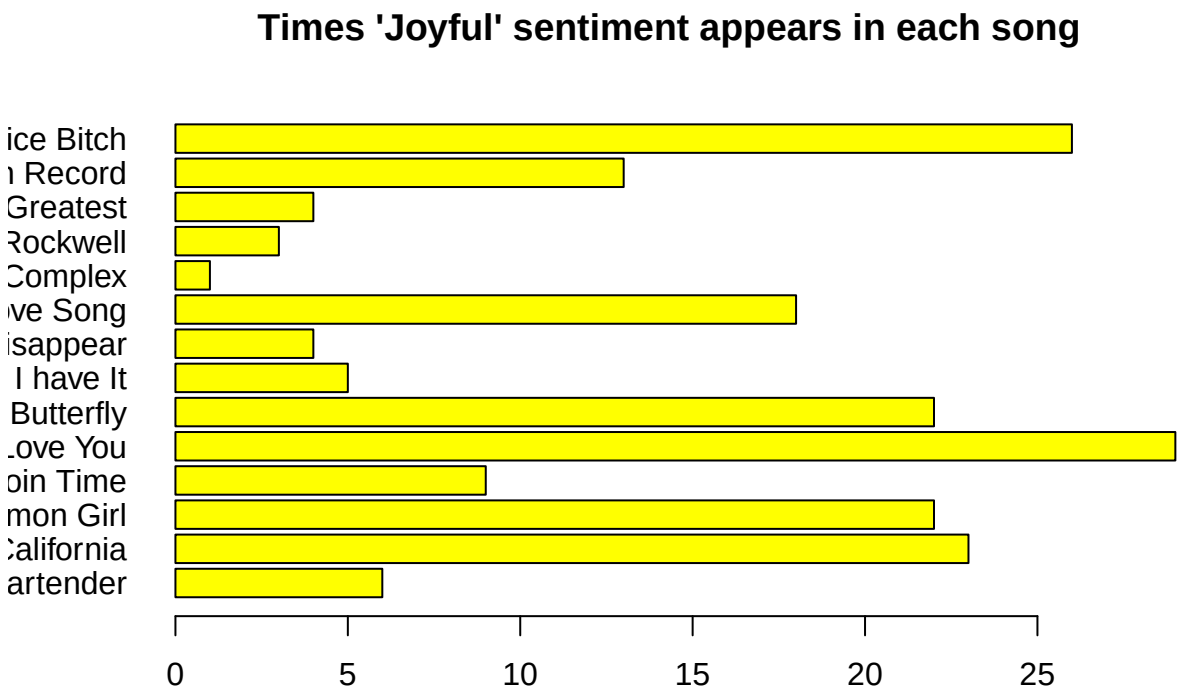

```
barplot(peacefultable[,2],
        main = "Times 'Peaceful' sentiment appears in each song",
        horiz = TRUE,
        col = "pink",
        las = 1)
```




```
barplot(powerfultable[,2],
        main = "Times 'Powerful' sentiment appears in each song",
        horiz = TRUE,
        col = "green",
        las = 1)
```




```
barplot(joyfultable[,2],
        main = "Times 'Joyful' sentiment appears in each song",
        horiz = TRUE,
        col = "yellow",
        las = 1)
```



Topic 4: Logistic Regression

Once the distribution of the sentiment has been generated, this project will use logistic regression to understand the effect of lyric sentiment on the song/album's commercial success and reception (this information will be pulled from Billboard.com)

Topic 5: Cleaning text data

Tidytext will allow for cleaning the text data by removing "stop words" and punctuation.

Conclusion

At this time, I do feel this project will advance my knowledge and curiosity around text analysis. I was excited that R had the capability to do that type of analysis.