Final-Project-Draft-Chris-Moua

Chris Moua

5/3/2022

Introduction

What

What this project covers is a text analysis of Lana Del Rey song lyrics. The goal of this project will uncover a sentiment analysis of Lana's lyrics to paint a picture of her artistic evolution and draw correlation between the emotional identity of her lyrics and their respective album commercial success and reception. This project will attempt to cover at least three albums: Born to Die, Ultraviolence, and NFR!.

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 will use a lyric website to generate the lyrics. According to 'Text Mining with R', each stanza can be called a character vector that will be analyzed. Therefore, I will employ tidytext and forcat for the text analysis/mining, dplyr for the data framing, and application of statistical concepts of distribution and inference for the analysis.

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."^1 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.

An sample character vector from Lana Del Rey will look like:

```
text<-C("Why, who me, why? -"),
"Feet don't fail me now -",
"Take me to your finish line -",
"Oh my heart it breaks -",
"Every step that I take"
text
```

Each row of text or token will be associated with a one of the six main sentiments from this wheel. The main six are at the inner center of the circle.

As noted, I will use tidytext to do the text clean up and analysis. Once I have converted each lyric line into a token/tidytext format, I will then create a distribution chart of the different sentiments.

Problems and Challenges

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

1) How to structure the unstructured text data?

I am attempting to just use the simple data structure the tidytext tutorial provides, via using the 'janeaustinr'package. However, that file already exists in a package that can be brought over to R easily. My challenge is with figuring out how to "package" the album lyrics. For now, the project will have to use excel to organize the text data.

2) How to assign sentiment to the text?

Another challenge is figuring out how to assign the sentiment to the lines of text. At the moment, I am unsure of the approach. However, the project may have to do a manual qualitative assessment of the lyrics and assign sentiment that way before the whole data set is formatted. I would like to avoid my own personal bias when assigning/inferring sentiment to the lyrics, but that will be a limitation I will acknowledge in the final project.

3) Open issue:

The project may have 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

R Markdown

The project report will be generated in R Markdown. Also, tidytext will be done in R.

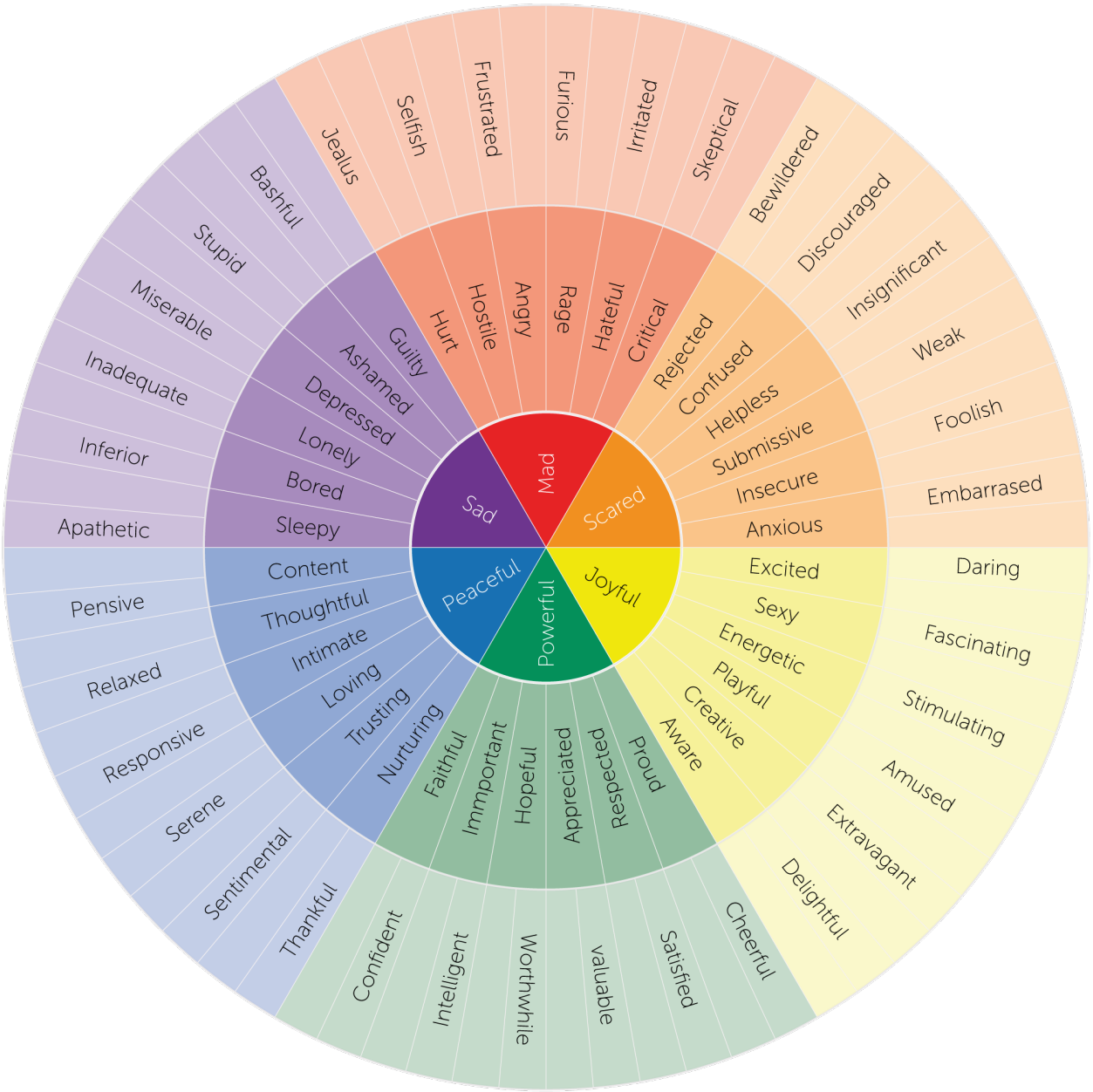

Figure 1: The Sentiment Wheel

Github

The project will be shared in the Github repository.Additionally, I am using a few learning tutorials from other users in Github to deal with the text analysis.

Text Analysis

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.

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)

Clearning 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.