Problem 3 – Data Capture / NLP
Sarah Lazio-Maimone

The purpose of this document is to generate a data set from web scrapping the Rolling Stone's Top 500 Album reviews found at https://www.rollingstone.com/music/music-lists/500-greatest-songs-of-all-time-151127/. Using this unstructured review data, the team wants to analyze the text written by Rolling Stone's magazine to better understand what exactly is being said about each album. After applying natural language processing techniques, we should be able to grasp the meaning of what it is to be considered a top 500 album of all time.

The team successfully scraped the Rolling Stones top 500 album reviews using Python's BeautifulSoup and Selenium packages. Simply put, BeautifulSoup allows the programmer to view and grab the underlining webpage's code, and Selenium automates clicking through the browser. The dataset produced by this routine was almost perfect, but we decided to manually clean up some of the columns in Excel. The Year column was converted to 4-digits, from 'YY' to 'YYYY'. We also hired an intern (me) who manually removed each records' leading writer, producer, and billboard information that was dumped into the beginning of all the reviews/descriptions. Finally, the team was able to search, replace, and parse out the trailing 'related' information that described which other relative music the artist had contributed to. This left us with a clean data set that was ready for natural language processing!

Before breaking down the album reviews, we wanted to get a better of an understanding of our data set. Below in Figure 1, we see that a majority of these albums were produced between 1960 and 1970. This tells us that we can expect to see a lot of our reviews to be heavily influenced by what genres were popular during those decades.

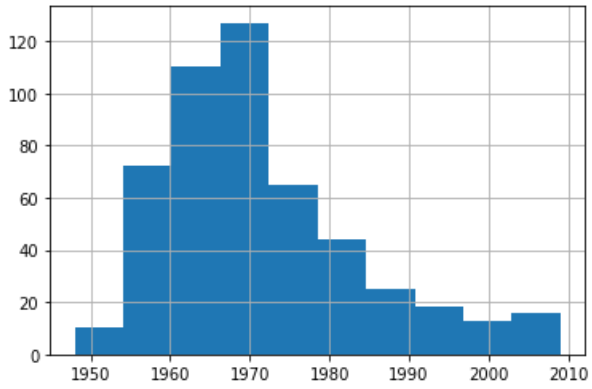

Figure 1

Problem 3 – Data Capture / NLP
Sarah Lazio-Maimone

When we visualize the data by artist. Below in Figure 2, the data set is largely represented by The Beatles, The Rolling Stones, Bob Dylan, Elvis Presley, and U2, all having more that 5 records in the top 500 albums rated by Rolling Stones Magazine.

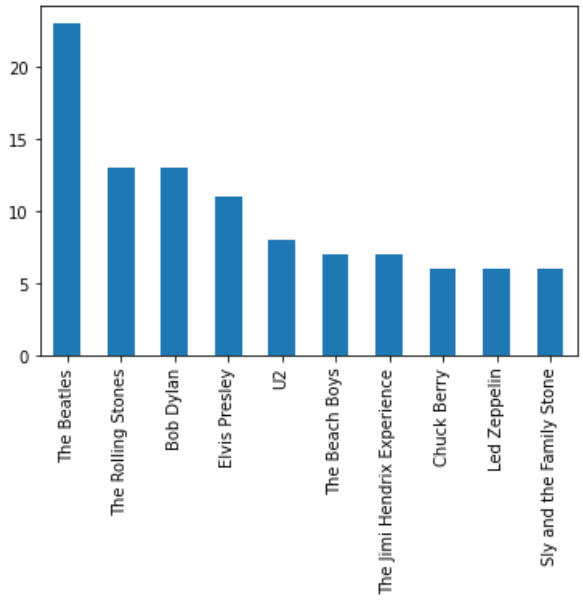

Figure 2

Finally, we take a look at the top writers of the Rollings Stone data set. In Figure 3 below, the top albums are written by John Lennon & Paul McCartney, Mick Jagger & Keith Richards, Bob Dylan, Bono (+U2), and Prince. This aligns well with the previous artist graph shown in the last section.

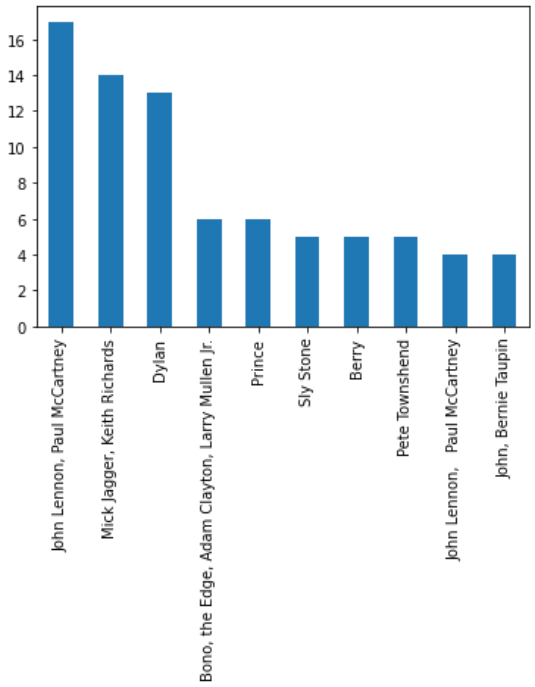

Figure 3

Problem 3 – Data Capture / NLP
Sarah Lazio-Maimone

Since we want to understand the language of each review describing the albums, we now begin our natural language processing techniques. The first step is to remove any non-numeric characters, like: !@#$%. Next, we remove any stop-words. These are very common spoken words used in text/language that don't really add to the meaning of the sentence, examples are: a, an, we, you, and, the. The team decided to add additional stop-words to skip over that are specific to musical data sets, such as: 'song', 'songs', 'appears', 'make', 'title', 'track', etc. Finally, we convert all of our characters to lower-case, which negates any differences between cases. After this step, our data set should only have words from each review that provides actual meaning to what The Rolling Stones are conveying about each album.

Using a Word2Vector model, we can now see which specific words are most similar to each other. Since this data set only considers the top albums of all time, every review should be a good one. It might make the most sense to determine which qualities of certain music can attribute to an album becoming a top 500 record. In Figure 4 below, we are analyzing which words are most similar to top artists Hendrix, Queen, and Springsteen. The percentages after each row indicate how similar to each word is to what you're comparing. The top 3 words similar to Hendrix is: guitar, studio and McCartney (Figure 4). The top 3 associated with Queen is: cut, McCartney and band (Figure5) . Finally, Springsteen is similar to: classic, rolling, and inspired (Figure 6).

```
[('guitar', 0.9945306777954102),
 ('studio', 0.9942450523376465),
 ('mccartney', 0.9934495091438293)]
```

Figure 4

```
[('cut', 0.9776227474212646),
 ('mccartney', 0.9771108627319336),
 ('band', 0.9768770337104797)]
```

Figure 5

```
[('classic', 0.9678933620452881),
 ('rolling', 0.9663482904434204),
 ('inspired', 0.966289222240448)]
```

Figure 6

For fun we wanted to see if our data set agrees that Prince was the king of pop. Below in Figure 7, Prince is 99.03% similar to the word "pop".

```
albums_w2v.wv.similarity('prince', 'pop')
```

```
0.9903786
```

Figure 7

Problem 3 – Data Capture / NLP
Sarah Lazio-Maimone

When we apply a T-SNE model to our data set, we can now visualize the word similarity percentages in a plot. The overall plot is shown in Figure 8 below, but it's very hard to see since there are a lot of words in our data set. After circling Hendrix in red, Queen in black, and Springsteen in blue, we can see quickly that the distance between Queen and Springsteen are grouped closer together than Hendrix. This distance is a visual representation of how similar the words are based on our model.

Figure 8

We zoom into our three artists Hendrix (Figure 9), Queen (Figure 10) and Springsteen (Figure 11) below. Specifically, now we can visualize which words are similar to each artist

Problem 3 – Data Capture / NLP
Sarah Lazio-Maimone

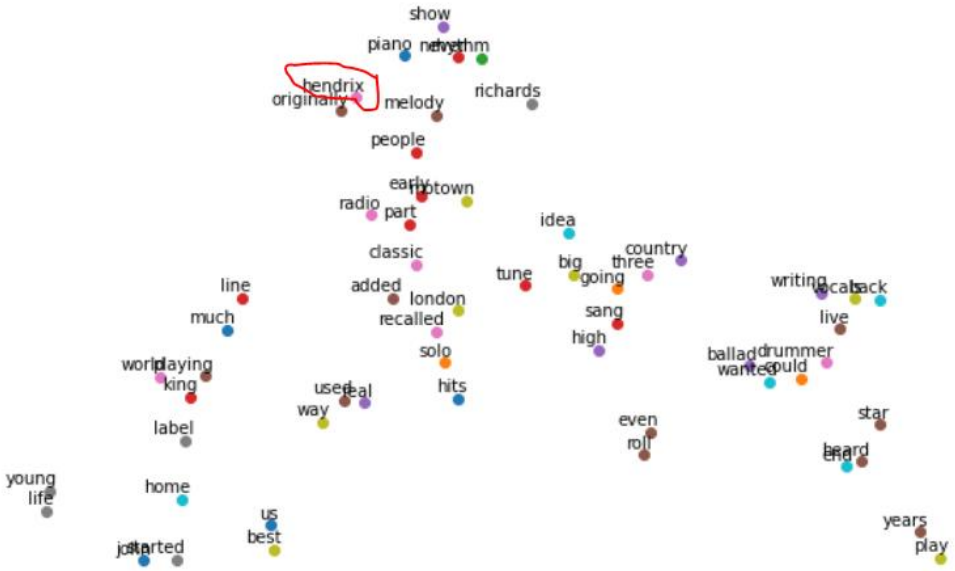

Figure 9

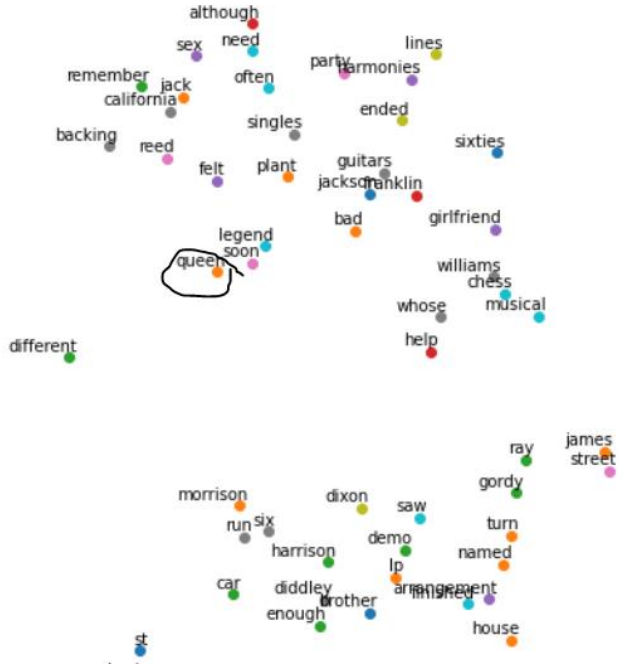

Figure 10

Problem 3 – Data Capture / NLP
Sarah Lazio-Maimone

Figure 11

We're noticing that there's a lot overlap with some of the words when analyzing similarities with specific artists. In order to visualize this, we created an embedded T-SNE cluster plot to farther define those similarities, seen in Figure 12 below.

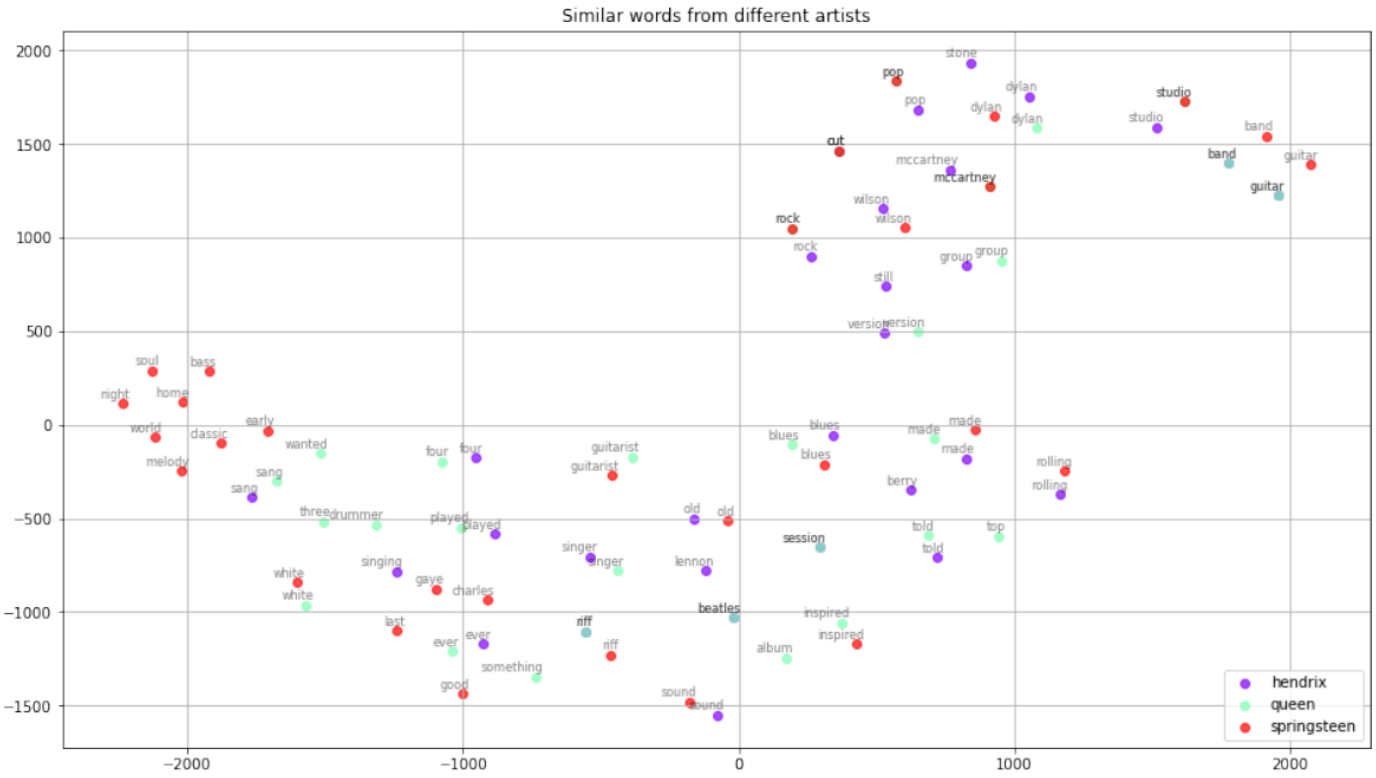

Figure 12

Noticeably, in Figure 13, we see that all three artists have similarities with [Bob] Dylan and the blues genre. Surprisingly, the word 'guitarist' is associated with Queen and Springsteen, but not Hendrix. While the word 'singer is associated with Hendrix and Queen. 'Riff' appears to be associated with all three. Finally, we notice a rather large cluster for Springsteen that includes: 'night', 'soul', 'bass', and 'classic'.

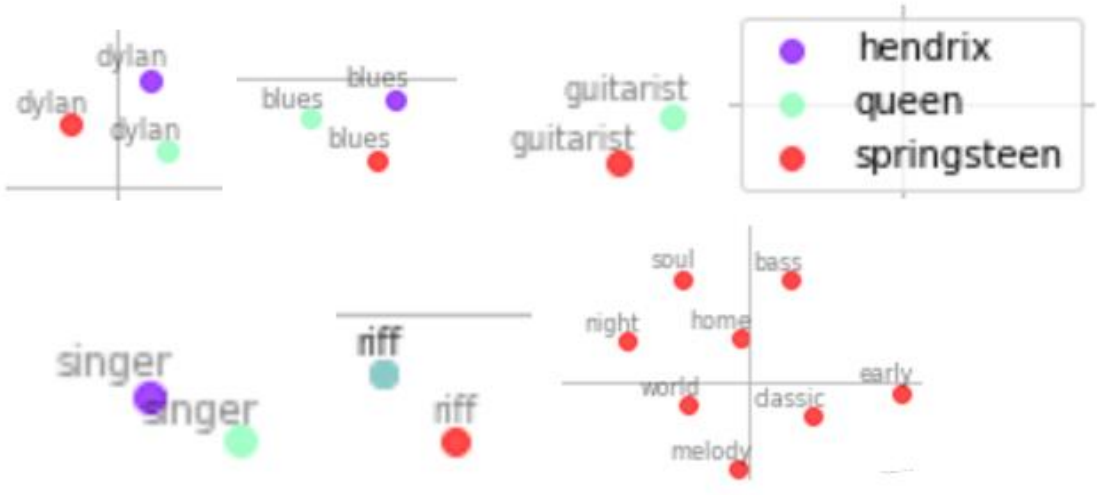

Figure 13

Problem 3 – Data Capture / NLP
Sarah Lazio-Maimone

One major flaw to this data is that all reviews are most likely written by the same person, or group of people with similar ideas. It makes it very difficult to gauge differences among the context of each review. We also know that all album descriptions are going to be positive because they are on the Rolling Stone's Top 500 for a reason. Another thing to note is that some of the reviews have a lot more words to them and contain a few sentences, while others are only one sentence. This could make it seem like there are more associations with the longer album reviews than with the shorter ones. I also found it very strange that 'guitar' was associated with Hendrix, but not 'guitarist'. We would most likely want to go back an create more definitions that would group similar words together, to minimize these discrepancies.

Problem 3 – Data Capture / NLP
Sarah Lazio-Maimone

Python Code Used to Scrape the Data

Python Code Used to NLP the Reviews