MA331 Sentiment Analysis

Harun Abdulqadir

25/08/2021

1 Introduction

Sentiment analysis is the process of recognizing negative or positive sentiments, detecting emotions and measuring polarity across a business or a body of text. Sentiment analysis is often used as a sort of precautionary or preemptive tool for businesses and organizations. By scaling thousands of reviews and gauging social reputation, it can be used to respond to disgruntled customers or even better understand general consensus on a particular project/product.

However, that is not all. Sentiment analysis, as we will demonstrate today can be used to decipher the disposition of a body of text such as, literature or even tweets from politicians. Various studies have been carried out for example on former POTUS Trump and his tweets, aiming to highlight a certain leaning and potentially forego a situation. Other examples include books and texts from famous literature. The applications for sentiment analysis are endless.

The two texts I aim to apply sentiment analysis on are:

- **Treasure Island**: Adventure novel by Scottish author Robert Louis Stevenson, serialized 1881-82.
- **Persuasion**: Published at the end of 1817, Jane Austen's last fully published romantic novel.

My analysis will aim to follow a particular line of thinking... both novels are from the same period and have been dramatized various times. This would suggest that they might have similar sentimental undertones.'Treasure Island' is largely an adventure novel and 'Persuasion' is a romantic novel. Although the differences are obvious and I suspect 'Persuasion' will have a more negative connotation, I am hoping to find more similarities than we would expect.

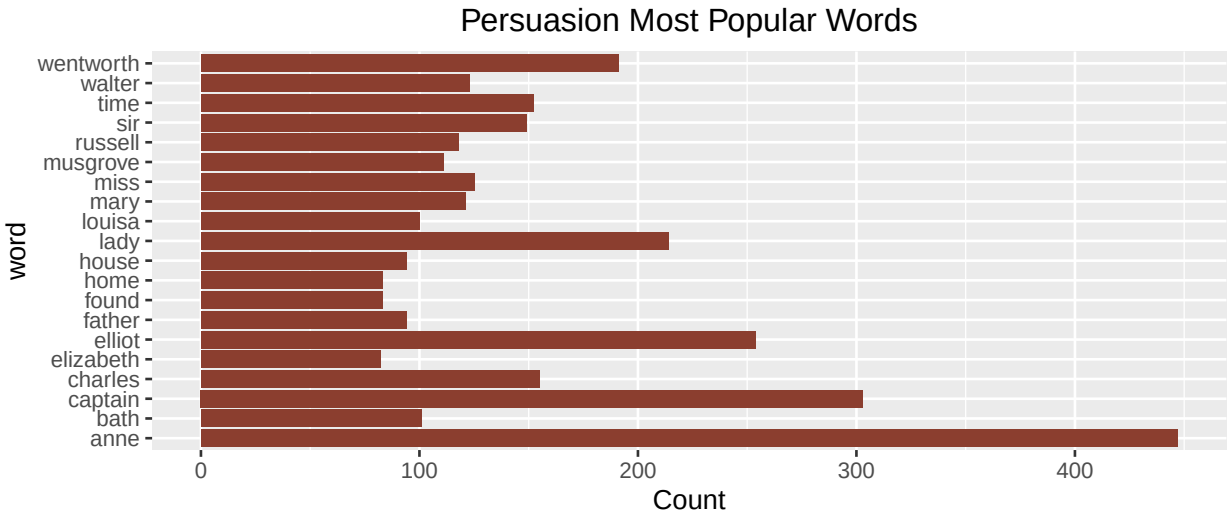
2 Methods

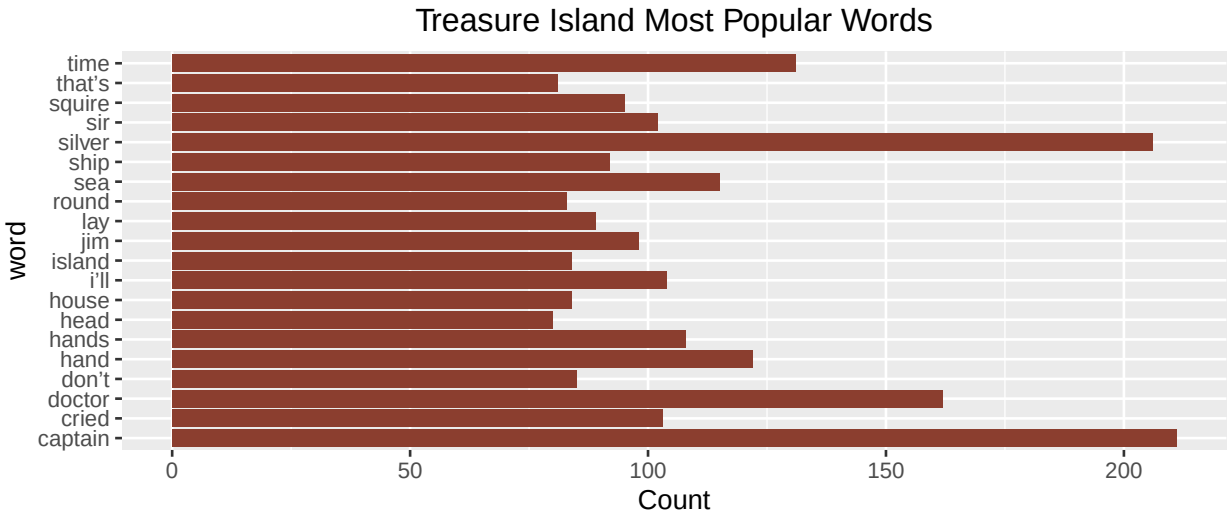
2.1 Pre-Processing

These two tables are important in providing a brief overlook of the text and the most frequent words. What preceded this was the cleaning of the data.

The pre-processing process often involves removing special characters, removing digits and blank spaces. The tidytext package makes it easier for us to eliminate the columns we don't need and remove the punctuation, digits and spaces, leaving just the text we need with a few lines of code.

Then, we used **stop_words** alongside the **anti_join** function to remove the unwanted words from the tidy set.





2.2 Sentiment Analysis

Now that we have the data in a tidy format, sentiment analysis can be performed with **inner join** from tidy text once again. As much as removing stop words is an anti join operation, performing sentiment analysis is an inner join operation.

Next we used the tidy format **'unnest_tokens()'** to convert the text to words in single items. This is endlessly important as it provides us with the basis to prepare us for our sentiment analysis. One we are ready, we can begin to insert the sentiment lexicons, such as BING, AFINN and NRC to find which words texts hold which emotions and the positive/negative nature of them.

3 Results

3.1 Sentiment Scores

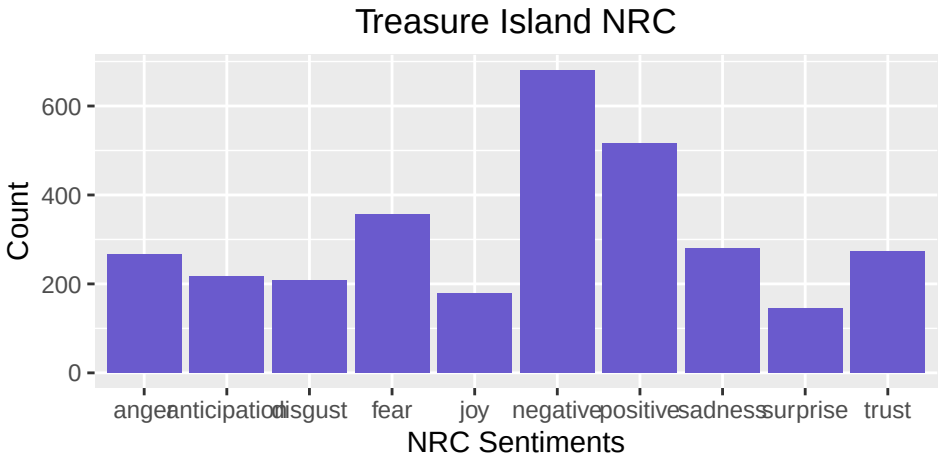
To achieve the dictionary based sentiment results, I run an inner join on the text as I previously mentioned. This only retains the 'emotion words' in the **bing** dictionary. Then I generate a couple things: firstly a string of the positive and negative words that were extrapolated and secondly a count of the total emotion words and the proportion of negative/positive.

3.2 Ratios

With the extracted words we are able to detail a visualisation of both the texts and their positive/negative sentiments before revealing the actual split and the context. As we can see persuasion seems to have more negative undertones, which is what we predicted but the frequency of words like 'miss', 'object' and 'poor' are interesting. As an adventure novel, we can see that 'treasure' being the most frequent and positively popular word gives an indication of the overall sentiment.

As we can see the, negative sentiment for persuasion was overarching. 697 negative sentiments to 505 positive ones which is 57.9%.

Treasure Island proved to be a bit more interesting as the BING sentiment was overwhelmingly negative, although the visual tended to give a different thought, even different to our preliminary prediction. As such we carried out NRC sentiment analysis for a wider range of perspective.



Interestingly we received similar sentiments from the NRC lexicon which was also mostly negative, with strong sentiments of words such as 'anger', 'sadness', 'fear'. However, when we also ran the same thing for Persuasion we received different results and this can be explained by the context within the lexicons and how they differ.

```
## # A tibble: 2 x 2
##   sentiment     n
##   <chr>     <int>
## 1 negative   4781
## 2 positive   2005
```

3.3 Insights

The three different lexicons for calculating sentiment give results that are different in an absolute sense although they have similar relative trajectories through the novel. We see similar dips and peaks in sentiment at the same places in the novel but the absolute values are significantly different. The lexicon from Bing et al. has lower absolute values and seems to label larger blocks of contiguous positive or negative text. The NRC results are shifted higher relative to the other two, labeling the text more positively, but detects similar relative changes in the text.

From what we can observe in the table above, the only table we are observing is BING as there isn't too much of a disparity between positive/negative words in NRC lexicon.There are far more negative words in the BING dictionary, which suggests to us that there is a bit of an unfair advantage for novels or texts that are on the fringe as they would pick up more negative sentiments.

4 Conclusion

To conclude, this paper explored reading text data into R and the pre-processing/cleaning process of the data. Demonstrated how to prepare the data for analyses and then continued to create visualisations to draw inferences for greater understanding of the novels. Our initial predictions were ultimately wrong in the sense that they weren't as closely related as we might have thought because they were both novels from the late 19th century.

In the end the dramatic romance was more negative as would have been the prevailing thought, and the adventurous novel had a more positive sentiment. The intricacies were however, very interesting as to provide context for some disparities.