MA615-Assignment4

Yifan Zhang

12/5/2021

Task 1: Pick a book

The book I picked is *David Copperfield* by Charles Dickens, which talk about the life of David Copperfield from childhood to maturity. The moral lesson that David Copperfield could be said to impart is that kindness, sympathy, and generosity are more important and perhaps more desirable than wealth, power, and social position.

Task 2: Words analysis

Words Frequency

For this part, firstly, I extract `chapters` from txt file and utilize `stop_words` lexicon to remove analytically useless words and calculate rest words' frequencies and proportions. Here is result (only frequency > 400 words will be displayed).

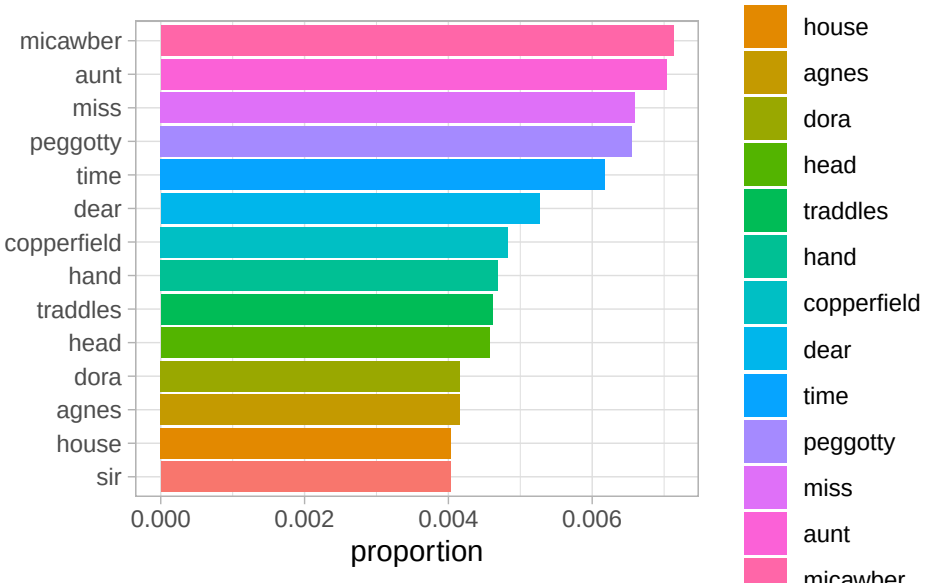

Figure 1: Word Frequency (>400)

Figure 1 shows that most common used words in *David Copperfield* are `micawber`, `aunt`, `miss` and `peggotty`, which account for over 0.6% of total words. To be more clear and get a wider view of it, I also include `wordcloud` to visualize it, which is quite a fancy plot.

Figure 2: David Copperfield wordcloud (Top 100 words)

Figure 2 shows us Top 100 frequently used words in *David Copperfield* and the sizes of words in cloud represent their frequency, the bigger, the more frequent. Actually, only calculate frequencies of a book is not enough. For the next step, I would go deeper into sentimental analysis of *David Copperfield*.

Sentimental test

As to sentimental analysis, R studio tidyverse provide three sentimental lexicon, `AFINN`, `BING` and `NRC`, for us to do words based analysis. To begin with, I utilize `BING` to label words in *David Copperfield* with binary sentimental attitudes, `positive` and `negative`. Here is the visualization of result:

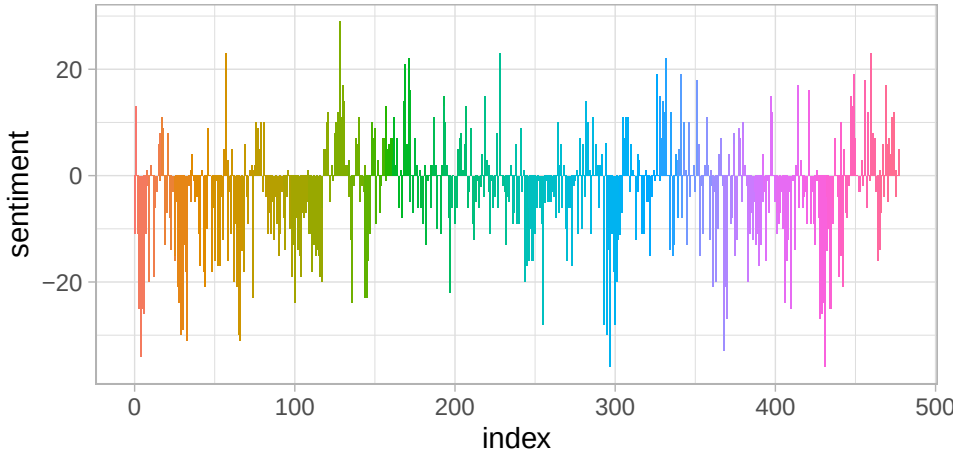

Figure 3: David Copperfield Sentiment Analysis (bing)

Figure 3 illustrates that, under the lexicon `BING`, the plot of *David Copperfield* changes toward more negative

than positive even some of plots indicate fairly positive. Then I decide to utilize more lexicons to continue sentimental analysis.

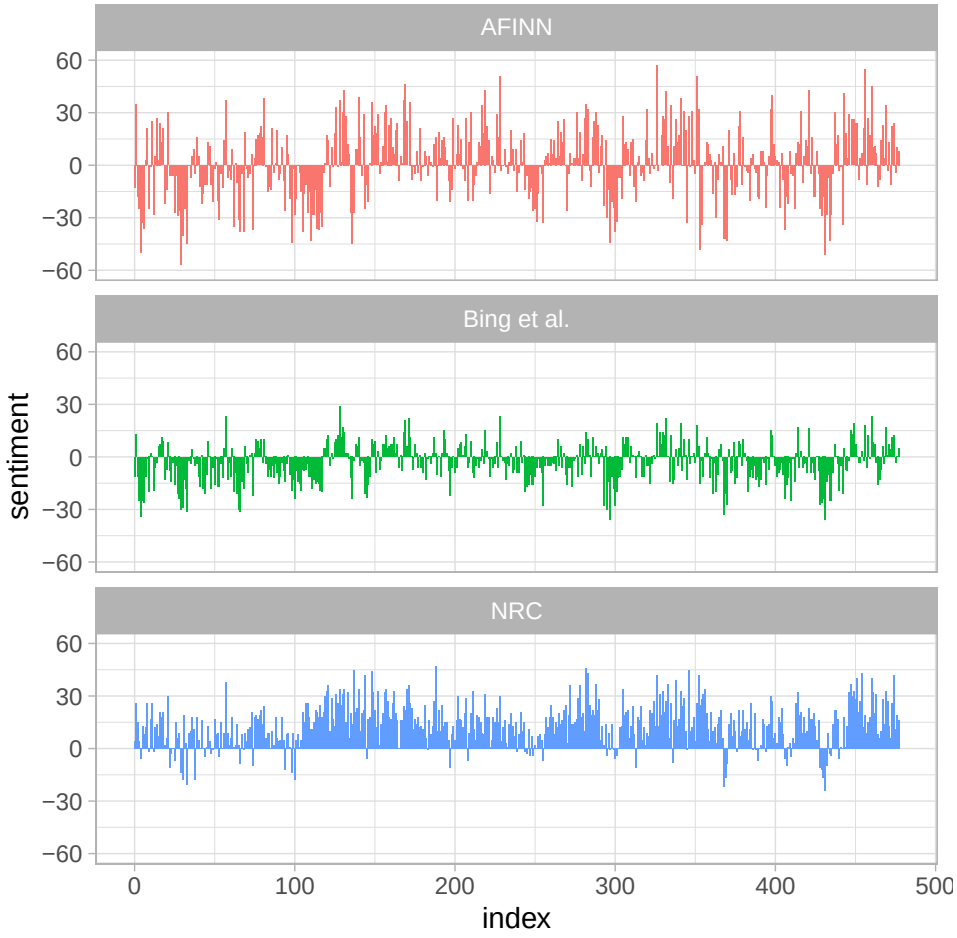

Figure 4: Comparision of three lexicons on David Copperfield

The three different lexicons for calculating sentiment give results that are different in an absolute sense but have similar relative trajectories through the novel. We see similar dips and peaks in sentiment at about the same places in the novel, but absolute values are significantly different. The AFINN lexicon gives the largest absolute values, with high negative values, while the BING lexion shows similar patterns but absolute values are relatively smaller. Comparatively, the NRC lexicon give a for more positive sentimental result. To figure out why this happens, I decide to look into words in these three lexicons.

The following table shows proportions of positive and negative words in lexicons:

```
#>   lexicon sentiment proportion
#> 1   afinn  negative     64.55%
#> 2   afinn  positive     35.45%
#> 3     nrc  negative     58.98%
#> 4     nrc  positive     41.02%
#> 5    bing  negative     70.45%
#> 6    bing  positive     29.55%
```

All three lexicons have more negative than positive words, but the ratio of negative to positive words is higher in the Bing lexicon than the NRC and AFINN lexicon. For the next step, I find out most common positive and nagative words in the book then visualize that in fig 6 and fig 7.

Figure 5: Top 10 Negative and Positive Words

Both of plots indicate that the most common used negative word is `miss` while the most frequently used positive one is `love`. We can use this visualization to see the most important positive and negative words, but the sizes of the words are not comparable across sentiments.

Figure 6: Comparison Cloud (BING)

Extra Credits

Apart from `AFINN`, `NRC` and `BING`, I include lexicon `loughran`. here is the proportion of negative and positive words in this lexicon.

```
#> # A tibble: 2 x 3
#>   sentiment     n proportion
#>   <chr>     <int> <formttbl>
#> 1 negative   2355 86.93%
#> 2 positive    354 13.07%
```

About this lexicon, it is weird that over 85% of words in this lexicon are `negative`. Next, I utilize it to do the similar research on book *David Copperfield*.

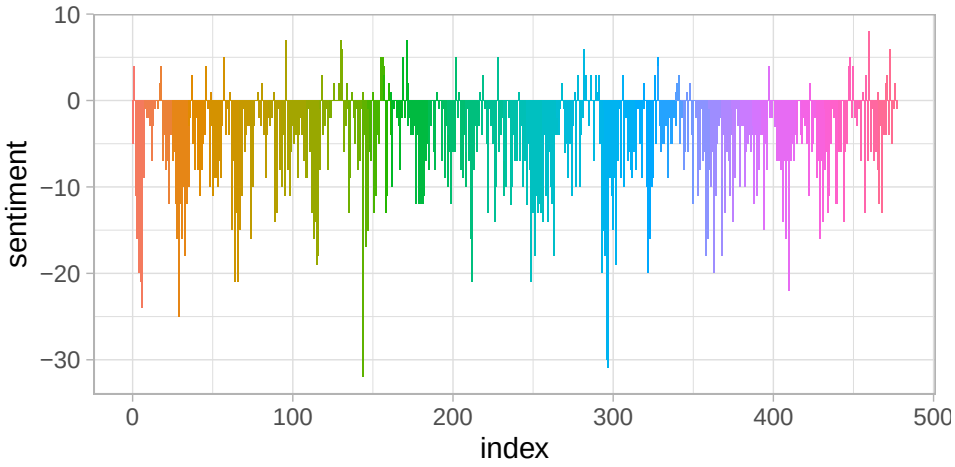

Figure 7: David Copperfield Sentiment Analysis (loughran)

No surprise, no matter how plots going on, the sentiment result shows greatly negative emotion, which can be explained by over 85% of words are `negative` in this lexicon.