Task2Report

Mi Zhang

12/5/2021

Love at Paddington

"Love at Paddington" is written by W. Pett Ridge and pubished in 1912. It is a story of a middle class London girl, Gertie, involved in a romance with a man of the upper class, we later know his name is Henry. For the first several chapters, they did not exchange their name because they knew the gap in their social classes cannot let them go far, they haven't exchanged names. As the story goes on, things became more complicated and they almost engaged. It is interesting that how Gertie was able to stay so awaken about her identity and deal the relationship between Henry and herself. At end of story, Gertie decided to move on and back to her social class in which she showed herself as an independent and mature woman.

Tidy data

After download the book from gutenberg package, I converted the text to tidy format using unnest_tokens funtion and using group_by and mutate function to set up columns for linenumbers, chapters, and words. The table below shows a part of data frame of the book.

```
## # A tibble: 6 x 4
##   gutenberg_id linenumber chapter word
##          <int>      <int>   <int> <chr>
## 1        26135          1       0 love
## 2        26135          1       0 at
## 3        26135          1       0 paddington
## 4        26135          4       0 by
## 5        26135          6       0 w
## 6        26135          6       0 pett
```

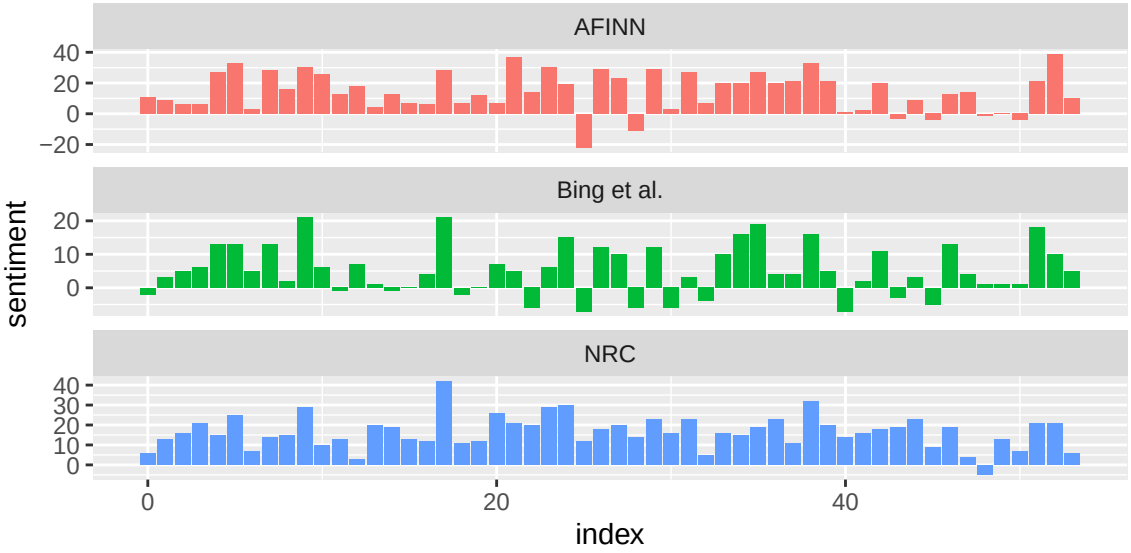
Using Lexicons to analyze sentiments

After I made the tidy text of the book, I started to use three different lexicons to analyze the sentiments of words in my text. It is interesting to see that both nrc and bing lexicon categorize words in the binary fashion into positive or negative, but nrc lexicon did more fine categories for words by categorizing them into different sentiments, like anger, disgust, fear, joy and so on. Afinn lexicon is more unique than above two in which it assigns words with a score between 5 to -5 and positive score is with respect to positive sentiment and vice verso.

we count up how many positive and negative words there are in defined sections of each book. We define an index here to keep track of where we are in the narrative; this index (using integer division) counts up sections of 80 lines of text.

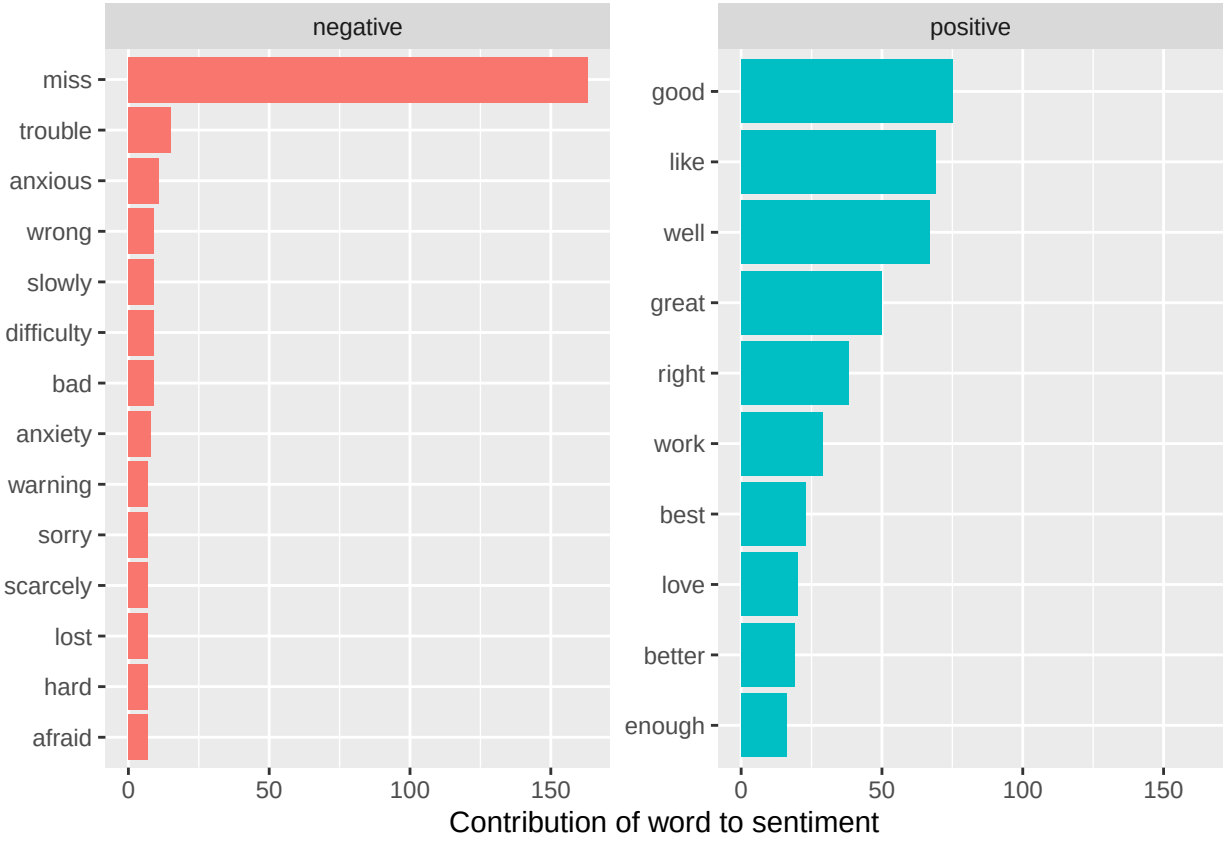
We then use pivot_wider() so that we have negative and positive sentiment in separate columns, and lastly calculate a net sentiment (positive - negative).

The plot below displays all three sentiment lexicons. We can see them displaying peaks and dips at about the same sections in the book, but the absolute values are significantly different. The AFINN and NRC sentiment have more variances, NRC has higher value among the three, and the Bing et al. sentiment show more movement of positive and negative sentiments.



Word contribution to Sentiment

Now, I want to know if the words in the book has more negative or positive sentiment. Since bing lexicon categorize words into binary fashion of positive and negative sentiment, by using inner_join() to merge tidy book and bing dictionary will help me to answer the question. We can see that words of positive sentiment appeared more frequently. And "Love at Paddington" is a love romance fiction which does make sense to me that it has more postive sentiments. I also notice that the word "miss" appeared the most often in the sentiment, but "miss" can also be a title for lady.



Different visualization

Since I have my tidy text, I can use it to do more visualization. The wordcloud below so a mixture of sentiment words in the book.

In order to make better visualization, I add color to identify the positive and negative sentiments and the font size is also corresponding to contribution of the word to sentiment, the bigger the font size means it has more contribution.

Extra Credit

In the Chapter 5 of the book "Text Mining With R", it mention another lexicon which is related to financial sentiments which is "loughran". According to the book,the loughran data divides words into six sentiments: "positive", "negative", "litigious", "uncertain"", "constraining", and "superfluous". In order to make better comparason, I just going to use the two sentiments, "positive" and "negative". Since loughran is used for finance, it makes sense that it look different than above lexicons. Indeed, I do not think loughran lexicon is useful for my book.

Reference

- https://www.tidytextmining.com/dtm.html
- https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf