MA615 Assignment4

Text Analysis of Pierre and Jean

Yanbing Chen

2021/12/6

Task One: Pick a book

According to the task one requirement, I pick a fiction book named Pierre and jean written by Guy de Maupassant on The Gutenburg Project to do text analysis. The book size is 284kB.

Task Two: Bag of words analysis

In this part, I tidy information in the book to gain a clean data file called 'tidy_pierre'. Because there are many options for sentiment lexicons, I try three methods that are AFINN,BING, and NRC to make box plot to find which measure is the best. As figure1 shown, the BING fits better, therefor, finally I chose to apply NRC method in deeper text analysis.
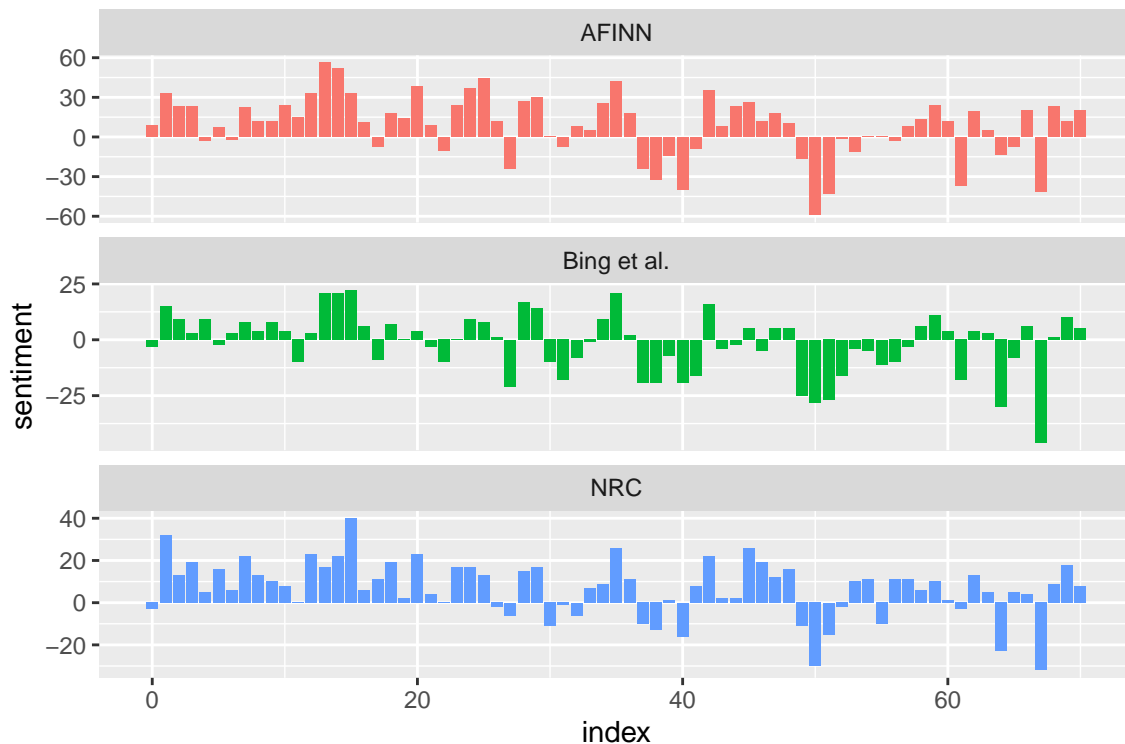

Figure 1: sentiment plot

Figure 2: negative positive words count

Figure2 counts the number of positive and negative sentiment words in `Pierre and Jean`, and lists the top ten positive and negative emotional words that appear most frequently in the book. Base on the picture's information, we can see that positive words appear more frequently than negative words. In negative words, the most frequent occurrence is "poor", appearing nearly 30 times. In positive terms, "like" appears more than 90 times, and good, well follow by it.

Figure 3: word cloud

Figure3 illustrates that the most frequently used words in this book are jean, mother, father, time, son, day and so on. The size of the font and the depth of the color can also tell that words related to a person's name or family members appear more frequently.

Figure4 shows a comparison between positive and negative words in this book, and the size of words text is in proportion to its frequency within its sentiment. Apparently, the font size of positive words are much

Figure 4: sentiment word cloud

larger than negative words, which also explains that the frequency of positive words are higher than the negative words'.

Task Three: Sentence-level analysis