MA615 Assignment4

Text Analysis of Pierre and Jean Task Three

Yanbing Chen

2021/12/6

In this task, I used Truenumbers to do text analysis. Firstly,I adjusted book type and uploaded the book I chose to tnum. Then I plotted a picture to show order of chapters in this book, and calculated the positive words and negative words in each chapter to show a comparison(as Picture 1 shown).Finally, according to the task requirement,I compared this analysis with the analysis I did in Task Two(Bing method) and used Figure 2 to display the result.

Download the book

```
# read the book into R
#gutenberg_works(str_detect(author,"Guy de Maupassant"))
book<-gutenberg_download(c(3804))
```

```
#devtools::install_github("Truenumbers/tnum/tnum",force = TRUE)
library(tnum)
tnum.authorize("mssp1.bu.edu")
tnum.setSpace("test2")
source("Book2TN-v6A-1.R")
```

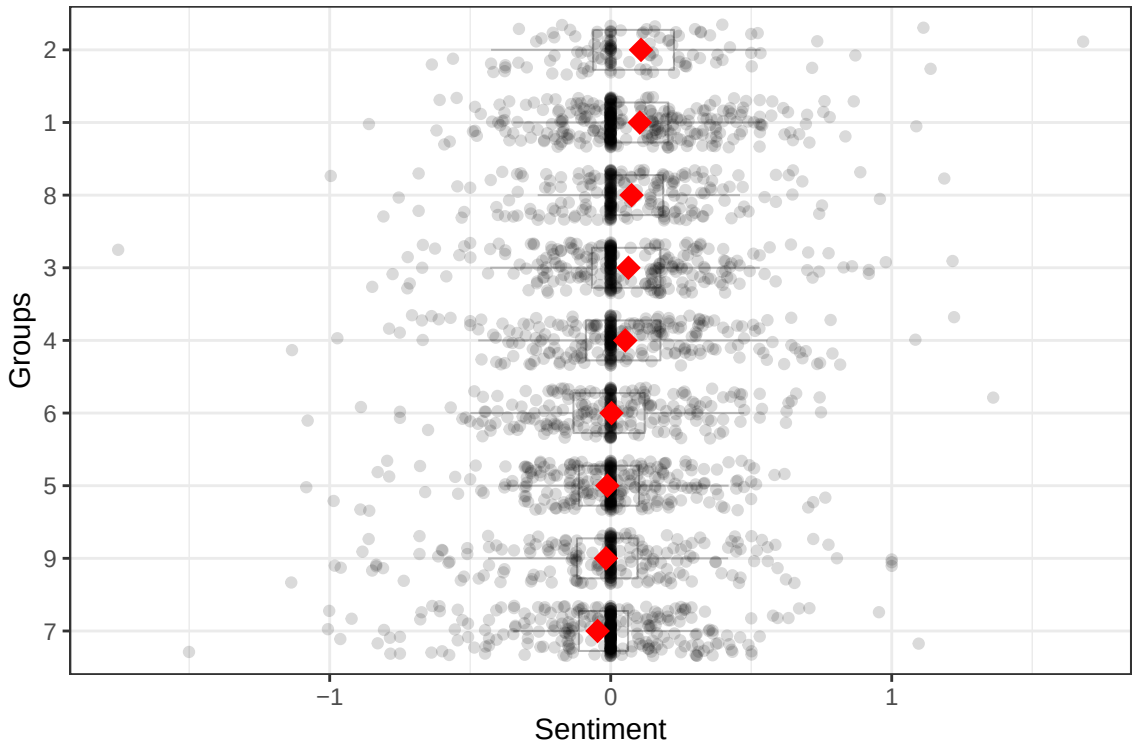
Load 'Pierre and Jean' into the test2 number space

```
DF6<- tnum.query('Maupassant/Pierre_Jeans/section# has text',max=10000) %>% tnum.objectsToDf()
DF6 %>% view()
pierre_sentence<-DF6 %>% separate(col=subject,
                  into = c("path1", "path2","section","paragraph","sentence"),
                  sep = "/",
                  fill = "right") %>%
  select(section:string.value)
```

```
#book_sentence$section<-str_extract_all(book_sentence$section,"\\d+") %>% unlist() %>% as.numeric()
pierre_sentence<-pierre_sentence %>% mutate_at(c('section','paragraph','sentence'),~str_extract_all(.,"
```

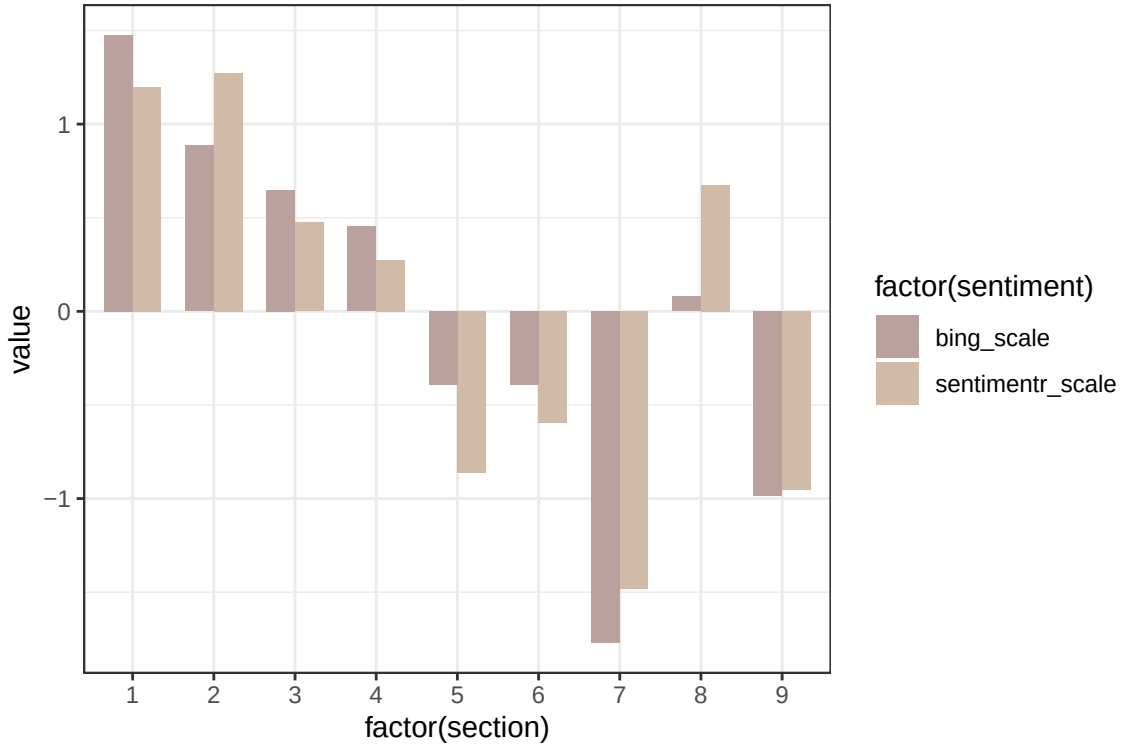
```
sentence_out<-pierre_sentence %>% dplyr::mutate(sentence_split = get_sentences(string.value)) %$%
    sentiment_by(sentence_split, list(section))

plot(sentence_out)
```



This picture shows the sentiments number in each section and lists the number of sentiments words in each section. The range of x-axis is from -1 to 1. Dots in -1 to 1 mean negative words, in the contrary, range 0 to 1 contains positive words. Based on the density of dots, the result is clear that there are more positive words than negative words in this book, which corresponds to the word cloud in task two. In addition, sention 7 contains the most number of sentiment words and section 1 has the less.

Compare two methods that were ultilized in Task Two and Task Three.



Due to these are two different methods, it is not easy and reasonable to compare them directly. Therefore, I limited the range of these sentiment words, just similar to what I did in the previous diagram in task 2. After defining the scale, I made a bar plot to explain the result. In each session, the sentiment trends in vocabulary are roughly the same, but the specific values are different. However, I think setimentr method is better the Bing.

Reference:

1.https://github.com/MA615-Yuli/MA615_assignment4_new 2.https://www.gutenberg.org/ebooks/3804
3.https://www.tidytextmining.com/sentiment.html