Task THREE

Wuji Shan

12/7/2021

Download the book

```
library(tnum)
tnum.authorize("mssp1.bu.edu")
tnum.setSpace("test2")
source("Book2TN-v6A-1.R")

# adjust the book type
Cami_book <- gutenberg_download(1608)
#write.table(Cami_book, "Cami_book.txt")
book_fix <- read.table("Cami_book.txt", header = T)
```

Load the book Camille into the test2 number space

```
# tnBooksFromLines(book_fix$text, "Alexandre_Dumas/Camille_Book92")

tidy_Cami_3 <- book_fix %>%
  mutate(
    linenumber = row_number(),
    chapter = cumsum(str_detect(text,
                                regex("chapter",
                                      ignore_case = TRUE)))) %>%
  unnest_tokens(word, text)

df_Cami <- tnum.query('Alexandre_Dumas/Camille_Book92/section# has text', max = 10000) %>% tnum.objects'

Cami_sentence <- df_Cami %>% separate(col = subject,
                  into = c("path1", "path2","section","paragraph","sentence"),
                  sep = "/",
                  fill = "right") %>%
  select(section:string.value)

Cami_sentence <- Cami_sentence %>% mutate_at(c('section','paragraph','sentence'),~str_extract_all(.,"\\d
                                           %>% unlist()
                                           %>% as.numeric())

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

plot(sentence_out)
```

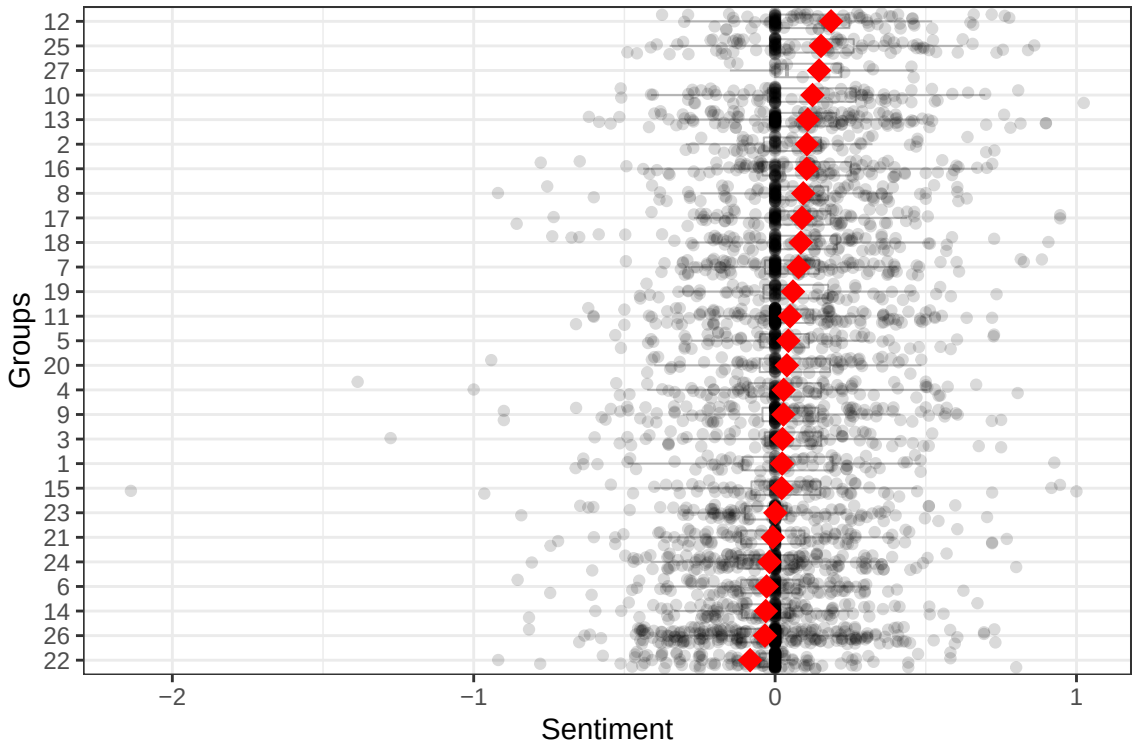

Figure 1: Sentiments Number in Each Section

Figure 1:

This graph shows the sentiments score group in each section via using sentimentr and sorted the average sentiment score from high to low. Range -1 ~ 0 of x-axis represents negative words, and range 0 ~ 1 represents positive words. We can observe that positive words appear more than negative words in book Camille. Moreover, group 12 contains most sentiment words and section 22 has the least.

Compare the bag of words analysis in Task TWO with that from TN

```
# create a new bing with index=chapter
new_bing <- tidy_Cami_3 %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al.") %>%
    count(method, index = chapter, sentiment) %>%
  pivot_wider(names_from = sentiment,
              values_from = n,
              values_fill = 0) %>%
  mutate(sentiment = positive - negative)

# scale sentiment to keep unit same
new_bing2 <- new_bing %>%
  mutate(bing_scale = scale(sentiment)) %>%
  select(method, index, bing_scale)
```

```
# change colname in order to join by section
colnames(new_bing2)[2]='section'

# scale sentiment to keep unit same
sentence_out <- sentence_out %>% mutate(sentimentr_scale = scale(ave_sentiment))

# join two df
sentence_out_2method <- left_join(sentence_out, new_bing2, by='section') %>%
  select(section,bing_scale,sentimentr_scale)

# use pivot longer for ggplot
sentence_out_2method_plot <- sentence_out_2method %>%
  pivot_longer(cols = c('sentimentr_scale','bing_scale'), names_to = 'sentiment')

# create barplot to compare
sentence_out_2method_plot %>% ggplot(aes(y = value,x = factor(section))) +
  geom_bar(aes(fill = factor(sentiment)), stat = 'identity', position = "dodge",width = 0.7) +
  theme_bw()
```

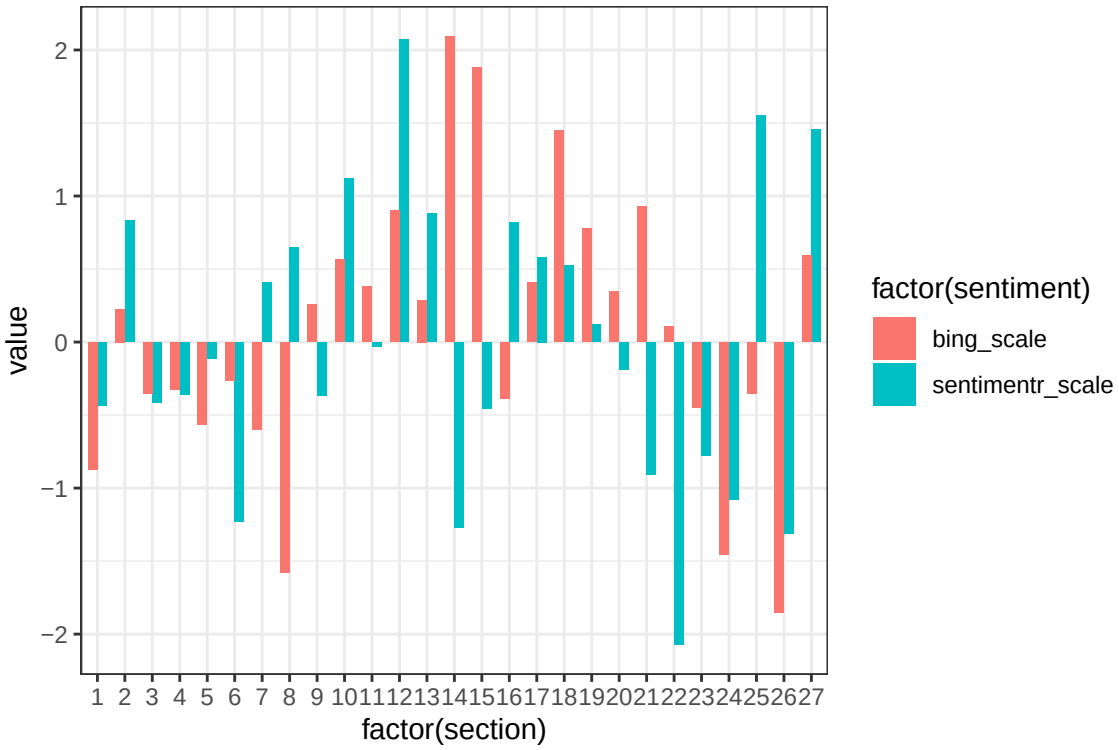

Figure 2: Compare two methods used in Task TWO and Task THREE

Figure 2:

Because of two different methods, it is hard to compare sentimentr and BING lexicon directly. Hence, I put two variable into the same criteria via using scale function. After refining the scale, I plotted a bar plot. From Figure 2, we can observe that among 27 chapters,the 11 chapters' sentiment outcome of BING scale sentimentr scale are opposite.

At the beginning, the sentiment's bias is negative because Marguerite was from countryside and her original family was very poor. During the mid-development of the story, Marguerite and Armand traveled together

living a life with respect and love, so the sentiment now is positive. At the end part, this couple were forced to forced to leave each other with misunderstanding, leading to the negative sentiment trend.

Therefore, I believe BING lexicon method fits the novel storyline better than sentimentr.

Extra Credit

Marguerite and Armand are female and male main characters of the book Camille.

This table shows the number of how many times each character appears in each chapter:

section	Marguerite	Armand
1	3	0
2	20	0
3	7	2
4	10	6
5	10	12
6	8	20
7	25	5
8	19	3
9	32	1
10	13	0
11	27	3
12	16	0
13	25	0
14	24	0
15	15	1
16	25	1
17	31	7
18	28	2
19	16	1
20	9	5
21	20	2
22	17	1
23	27	0
24	32	6
25	4	14
26	11	8
27	3	5

This table is the number of how many times both characters appear in the same paragraphs:

section	paragraph	both
4	25	1
4	41	1
4	62	1
5	1	1
5	18	1
5	36	1
5	37	1
6	27	1

section	paragraph	both
6	36	1
6	71	1
7	2	1
7	6	1
8	1	1
8	74	1
9	7	1
17	31	1
17	37	1
18	5	1
20	26	1
20	43	1
24	71	1
24	75	1
25	1	1
26	76	1
27	7	1

Reference:

1. Jin, Yuli, https://github.com/MA615-Yuli/MA615_assignment4_new
2. Gutenberg, https://www.gutenberg.org/ebooks/1608
3. Text Mining in R, https://www.tidytextmining.com/sentiment.html