Text mining - Task Three

Qiannan Shen

12/7/2021

Truenumbers

Firstly, I use functions in the Tnum package to load the book into the mssp test2 number space. Using the function tnBooksFromLines(text, root) and sourcing the document "doc/Book2TN-v6A-1.R" could complete this. Tnum package is a good tool to let us explore the data. The function tnum.getDBPathList() displays a list of the subjects in the tnum space which we could check whether our book has uploaded.

Secondly, using the function tnum.query() could help to approach the truenumber database, and the function tnum.objectsToDf() could let me make data frame from the list of tnum objects. I use tnum.query(query = "homer/the_odyssey/section# has text", max = 9000) to get the text in the book by section. Then I clean the tnum data frame. I separate the "subject" column into "Author", "Book_name", "Section", "Paragraph", and "Sentence" and make it a numeric type which could make following data-processing and visualization easier. Table 1 and table 2 display the clean data frame I got from Truenumber.

Sentiment plot

Figure 1 is the aggregated sentiment plot which displays the sentiment distribution for every section and the red dots are the mean sentiment score for these sections. This graph also ranks the sentiment from highest positive to lowest negative so that we could check the sentiment level based on the whole book easily.

Figure 2 is smoothed plot for the duration of the text based on percentage, allowing for comparison between plots of different texts. This plot gives the overall decreasing shape of the text's sentiment.

Table 1: Dataframe from Tnum

Author	Book_name	Section	Paragraph	Sentence	property	string.value
homer	the_odyssey	0	8	2	text	"BOOK II."
homer	the_odyssey	0	8	3	text	"BOOK III."
homer	the_odyssey	0	8	4	text	"BOOK IV."
homer	the_odyssey	0	8	5	text	"BOOK V."
homer	the_odyssey	0	8	6	text	"BOOK VI."
homer	the_odyssey	0	8	7	text	"BOOK VII."
homer	the_odyssey	0	8	8	text	"BOOK VIII."
homer	the_odyssey	0	8	9	text	"BOOK IX."
homer	the_odyssey	0	8	10	text	"BOOK X."

Table 2: Dataframe from Tnum

Author	Book_name	Section	Paragraph	Sentence	property	string.value
homer	the_odyssey	1	3	1	text	"Tell me, O Muse, of that ingenious hero who travelled far and wide after he had sacked the famous town of Troy."
homer	the_odyssey	1	3	3	text	"Tell me, too, about all these things, oh daughter of Jove, from whatsoever source you may know them."
homer	the_odyssey	1	7	4	text	"You, sir, take no heed of this, and yet when Ulysses was before Troy did he not propitiate you with many a burnt sacrifice?"
homer	the_odyssey	1	11	1	text	"Telemachus saw her long before any one else did."
homer	the_odyssey	3	26	8	text	"Meanwhile Aegisthus here at home plotted his evil deed."
homer	the_odyssey	4	34	3	text	"He is an Egyptian, and people say he is my father; he is Neptune's head man and knows every inch of ground all over the bottom of the sea."
homer	the_odyssey	5	26	5	text	"The sea took the raft and tossed it about as Autumn winds whirl thistledown round and round upon a road."
homer	the_odyssey	6	25	10	text	"Never mind him, but go up to my mother, and lay your hands upon her knees if you would get home quickly."

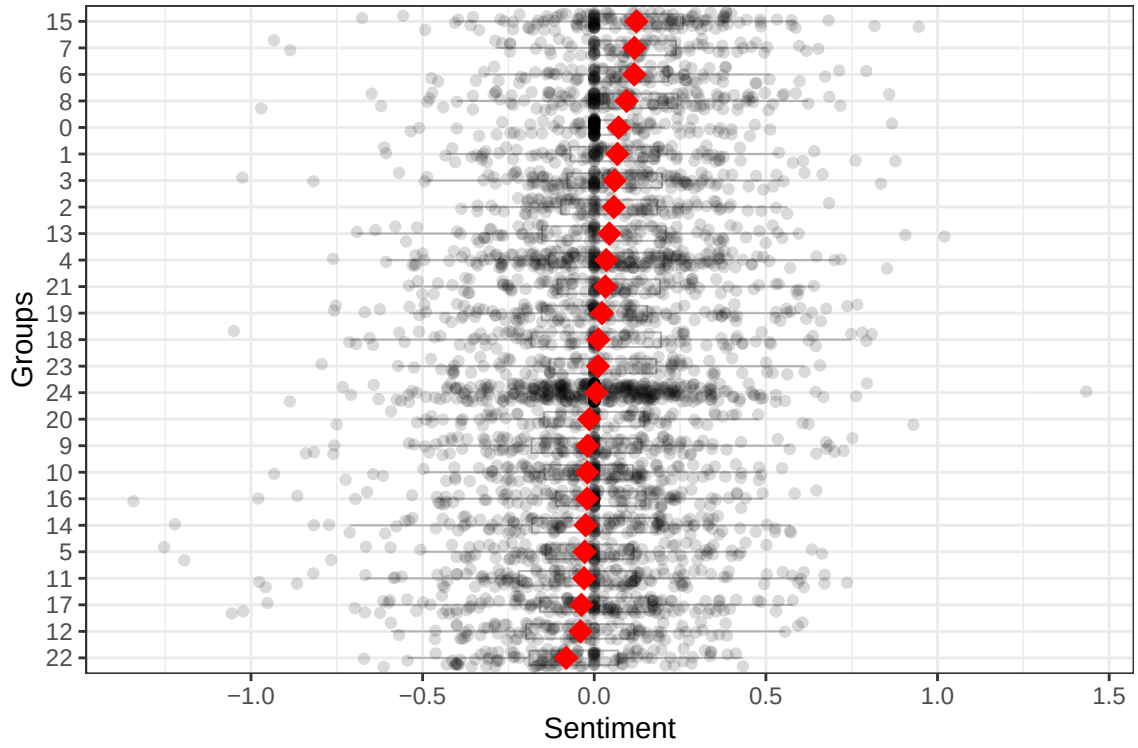

Figure 1: Aggregated Sentiment Plot

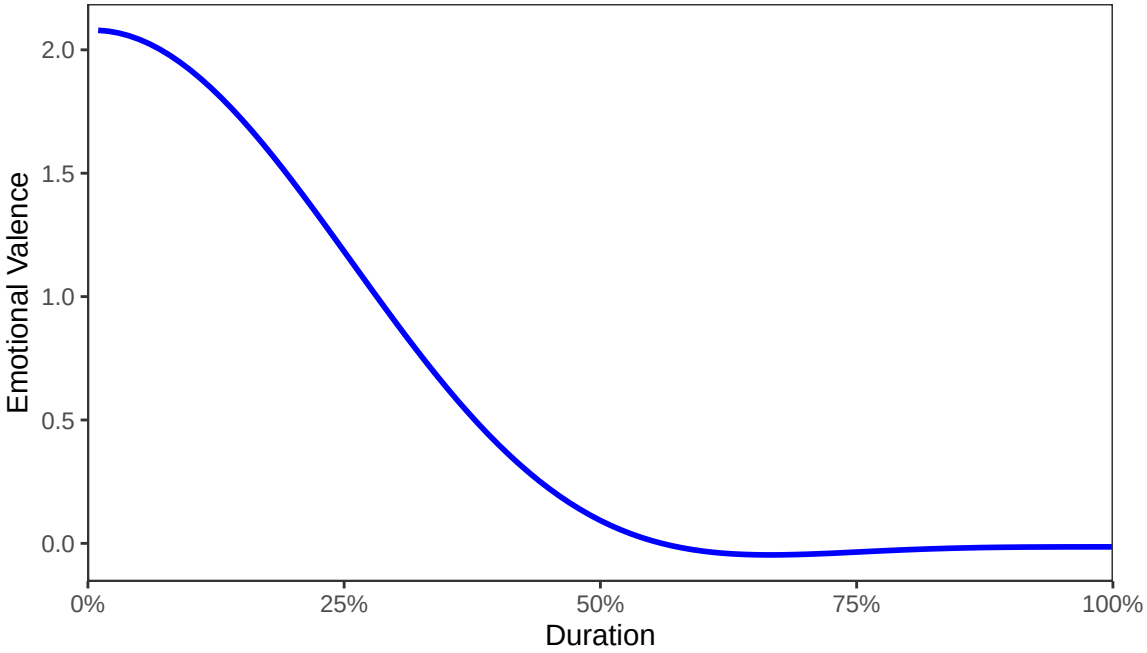

Figure 2: Sentiment Level Plot

Emotion Plot

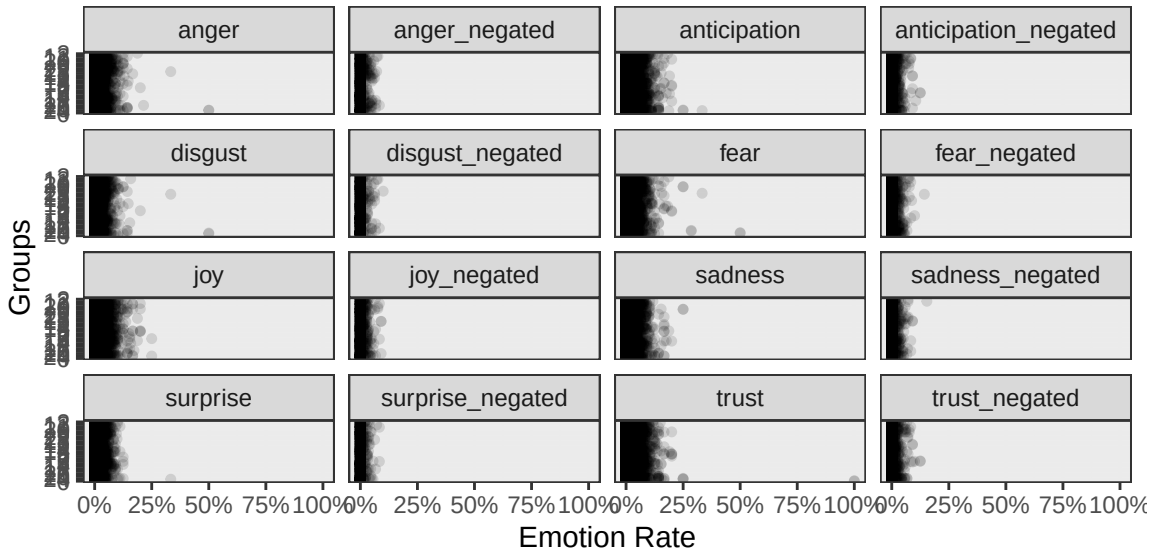

Figure 3: Emotion Plot

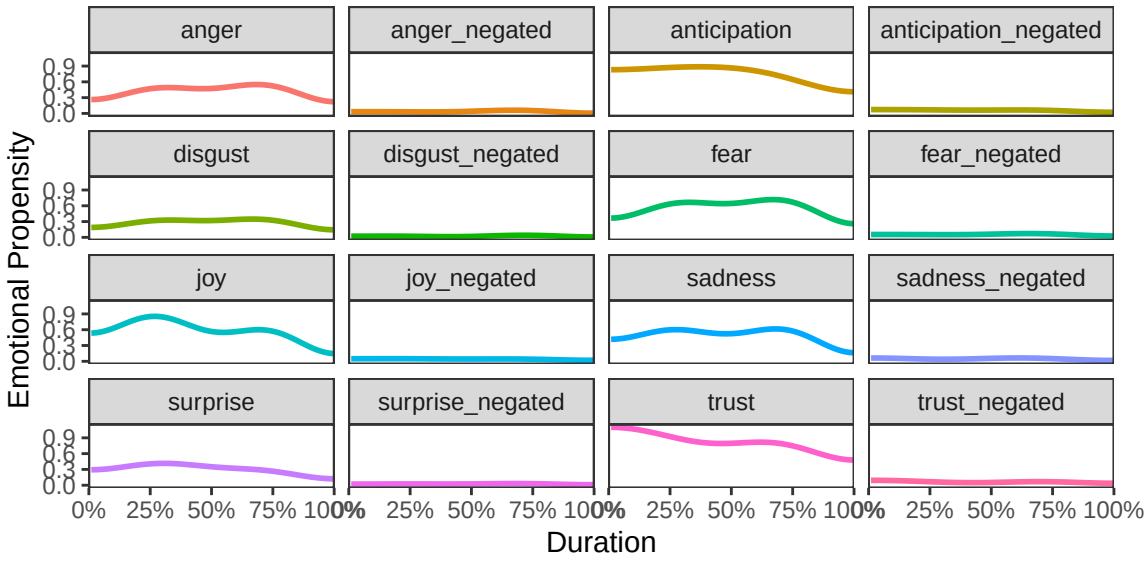

Figure 4: Emotion Level Plot

Comparison

Figure 3 compares the bag of words analysis done in Task two using Bing with the analysis in sentimentr and tnum package. Parts of them look the same in the sentiment, while some parts of the sentiment result are dramatically different. As we can see in sections 10, 17, and 23, the sentiment result is much different between these two methods.

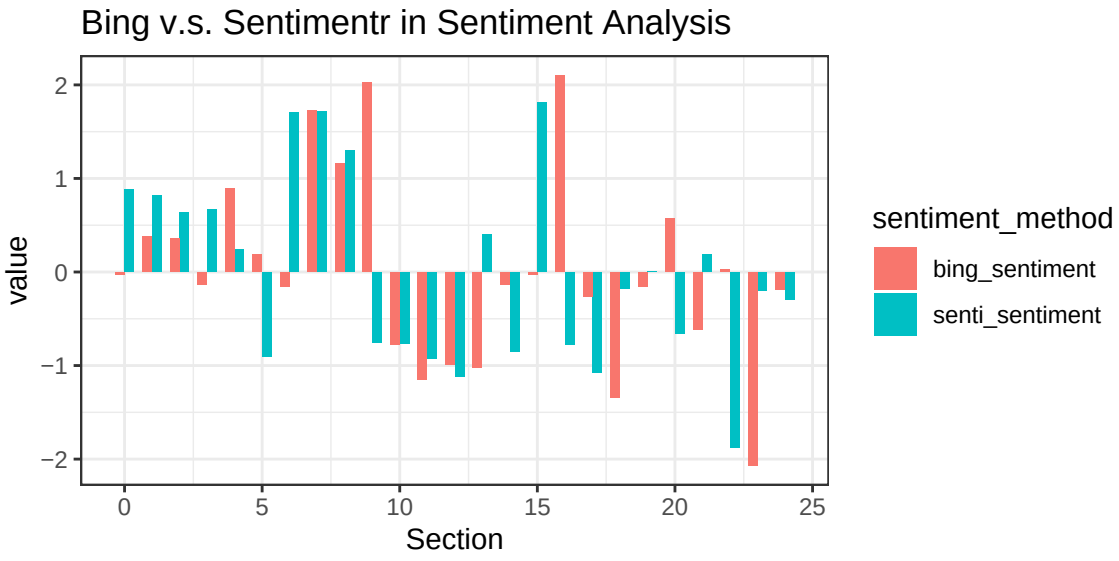

Figure 5: Bing v.s. Sentimentr in Sentiment Analysis

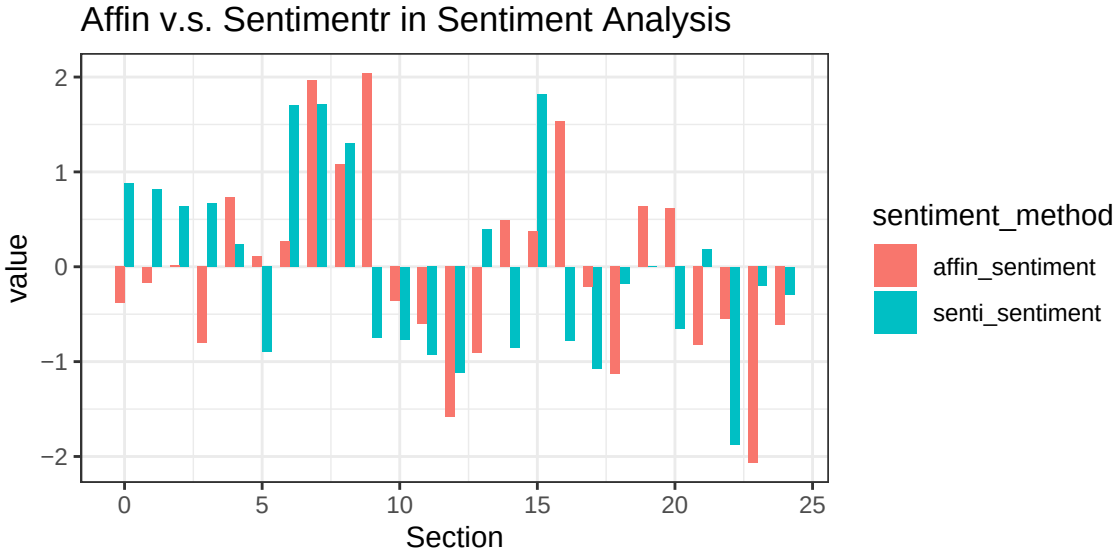

Figure 6: Affin v.s. Sentimentr in Sentiment Analysis

Figure 4 compares the bag of words analysis done in Task two using Affin with the analysis in sentimentr and tnum package. Figure 5 compares the bag of words analysis done in Task two using NRC with the analysis in sentimentr and tnum package. The results are similar to Bing and sentimentr one is somewhat different in that the sentiment analysis.

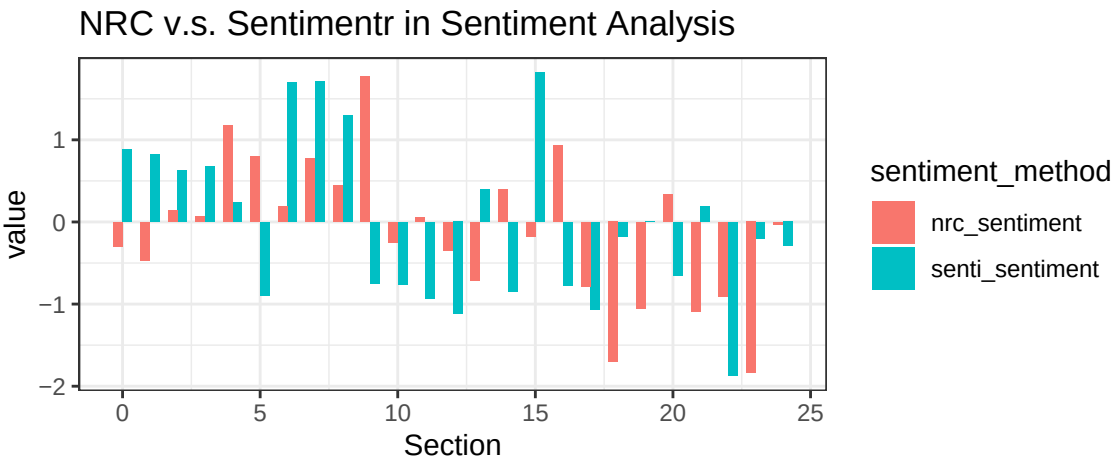

Figure 7: NRC v.s. Sentimentr in Sentiment Analysis

Extra point: Characters Tag

Table 3: The number of times Ulysses appears

Section	n
0	8
1	15
2	17
3	7
4	17
5	27
6	14
7	18
8	23
9	6
10	7
11	7
12	4
13	18
14	35
15	12
16	26
17	34
18	19
19	50
20	26
21	23
22	43
23	22
24	73

Table 4: The number of times Telemachus appears

Section	n
0	1
1	14
2	24
3	23
4	19
5	2
11	2
13	1
14	1
15	30
16	28
17	20
18	8
19	7
20	15
21	9
22	17
23	6
24	17

Table 5: The number of times both of Telemachus and Ulysses appear

Section	n
0	1
1	3
2	10
3	4
4	3
5	2
14	1
15	9
16	23
17	11
18	2
19	7
20	9
21	4
22	21
23	3
24	19

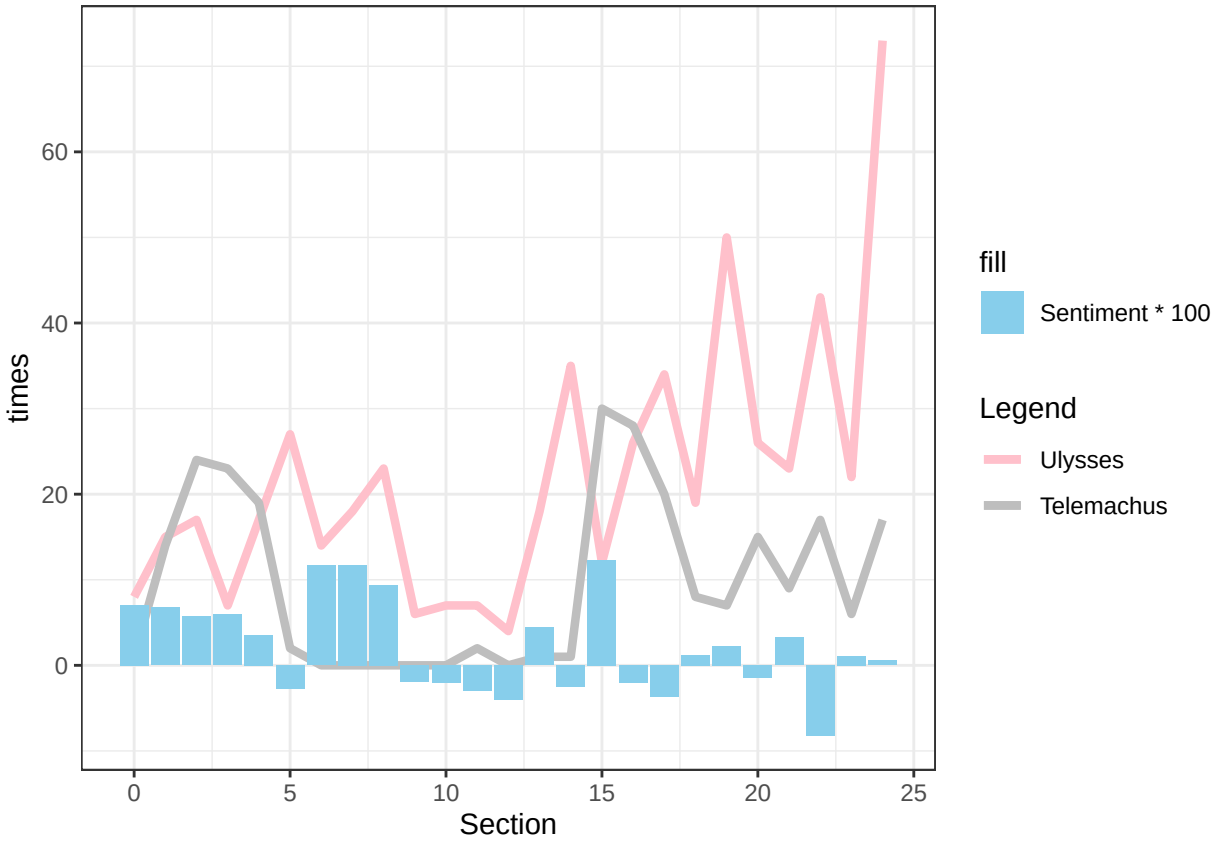

Figure 8: times of appearance with sentiment

Citation

Text Mining with R https://www.tidytextmining.com https://github.com/trinker/sentimentr https://github.com/MA615-Yuli/MA615_assignment4