Text analysis Task3

Yuli Jin

2021/11/24

task 3 sentence-level analysis

tnum

First, I put the book into tnum, the following table shows evidence of my tnum database.

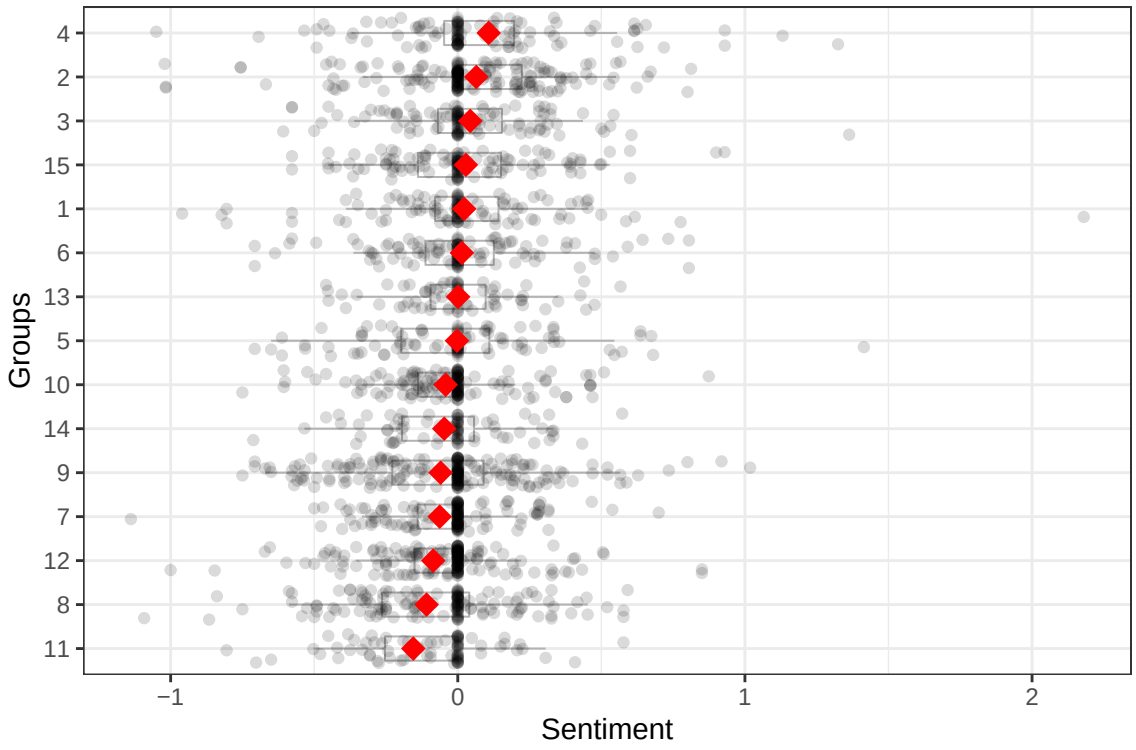
```
# query heading text to display the head
q24<- tnum.query('zweig/test2/heading# has text',max=90)
df24 <- tnum.objectsToDf(q24) # turn the object to df
knitr::kable(df24 %>% select(subject:numeric.value)%>% head())
```

subject	property	string.value	numeric.value
Zweig/test2/heading:0001	text	“”	NA
Zweig/test2/heading:0002	text	“”	NA
Zweig/test2/heading:0003	text	“”	NA
Zweig/test2/heading:0004	text	“”	NA
Zweig/test2/heading:0005	text	“”	NA
Zweig/test2/heading:0006	text	“”	NA

```
# query section and heading text to display the head
q26<- tnum.query('zweig/test2# has text',max=60)
df26 <- tnum.objectsToDf(q26) # turn the object to df
df26 %>% select(subject:string.value)%>% head()
```

```
##                                                 subject property
## 1                                 Zweig/test2/heading:0001     text
## 2 zweig/test2/section:0001/paragraph:0001/sentence:0001     text
## 3 zweig/test2/section:0001/paragraph:0001/sentence:0002     text
## 4 zweig/test2/section:0001/paragraph:0001/sentence:0003     text
## 5 zweig/test2/section:0001/paragraph:0002/sentence:0001     text
## 6 zweig/test2/section:0001/paragraph:0002/sentence:0002     text
##
## 1
## 2
## 3
## 4 "Exacerbated voices called back and forth; then, with a puffing and a chugging and another shrill s
## 5
## 6
```

Then I use sentimentr to get sentiment score group by these scores with section to get the average result.
The plot sort the average sentiment score from high to low.



Compare this analysis with the analysis you did in Task TWO

It is difficult to directly compare Sentimentr and Bing's score. Therefore, I apply `scale` function to keep two variable into the same criteria. Then I use ggplot to plot bar plot. From the Figure below, we can see that the trends, say positive and negetive direction, are mainly similar. But the exact number differs from two methods. Generally, sentimentr package is more optimictic than Bing method.

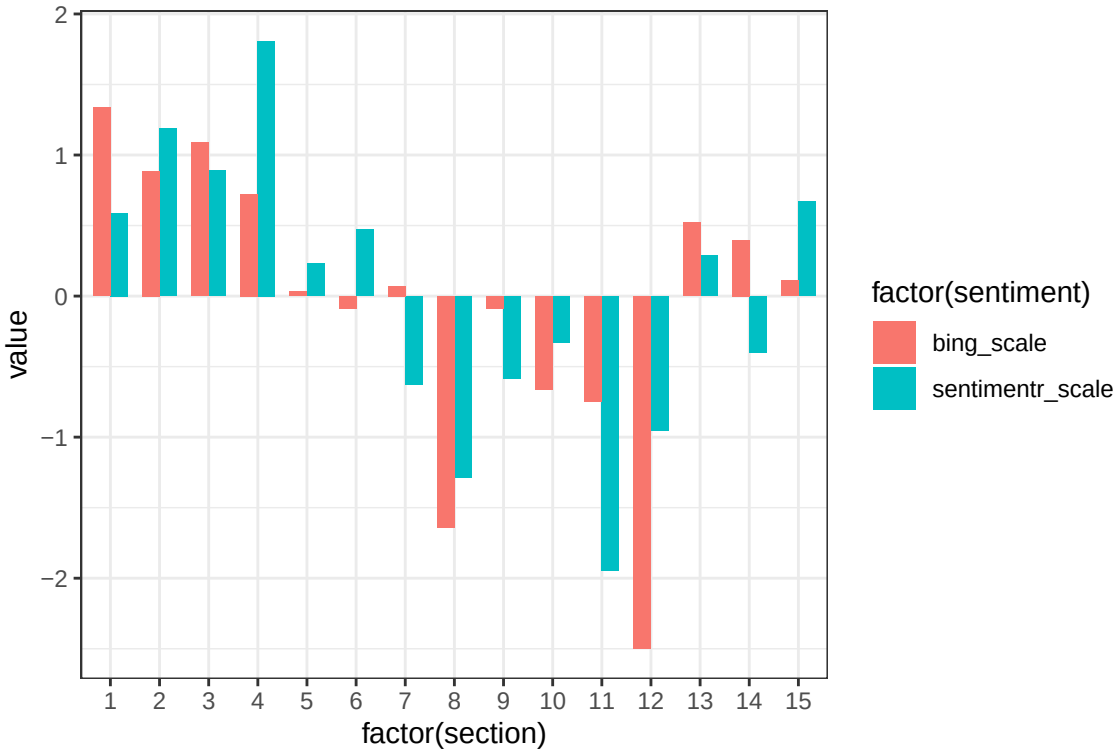


Figure 1: sentiment comparison

EXTRA CREDIT: character analysis

Baron and Edger are two main character among the fiction book. I Pick these two characters from my book. The following table in the count number of times each character appears in each chapter:

chapter	baron	edgar
1	10	2
2	22	17
3	18	16
4	13	12
5	9	5
6	16	18
7	12	13
8	15	24
9	13	19
10	8	14
11	10	12
12	5	18
13	0	11
14	1	13
15	1	14

The following table is the count of number of times both characters appear in the same paragraphs.

section	paragraph	both_appear
2	4	1
2	28	1
2	35	1
2	36	1
2	40	1
3	1	1
3	16	1
4	3	1
4	7	1
4	11	1
4	12	1
4	23	1
4	28	1
5	1	1
6	1	1
6	3	1
6	21	1
7	5	1
7	7	1
7	8	1
7	9	1
8	1	1
8	6	1
8	11	1
8	29	1
8	31	1
8	35	1
9	2	1
9	21	1
9	32	1
9	41	1
10	11	1
11	5	1
11	16	2

Reference

https://www.gutenberg.org/ebooks/45755
https://emilhvitfeldt.github.io/textdata/reference/lexicon_loughran.html