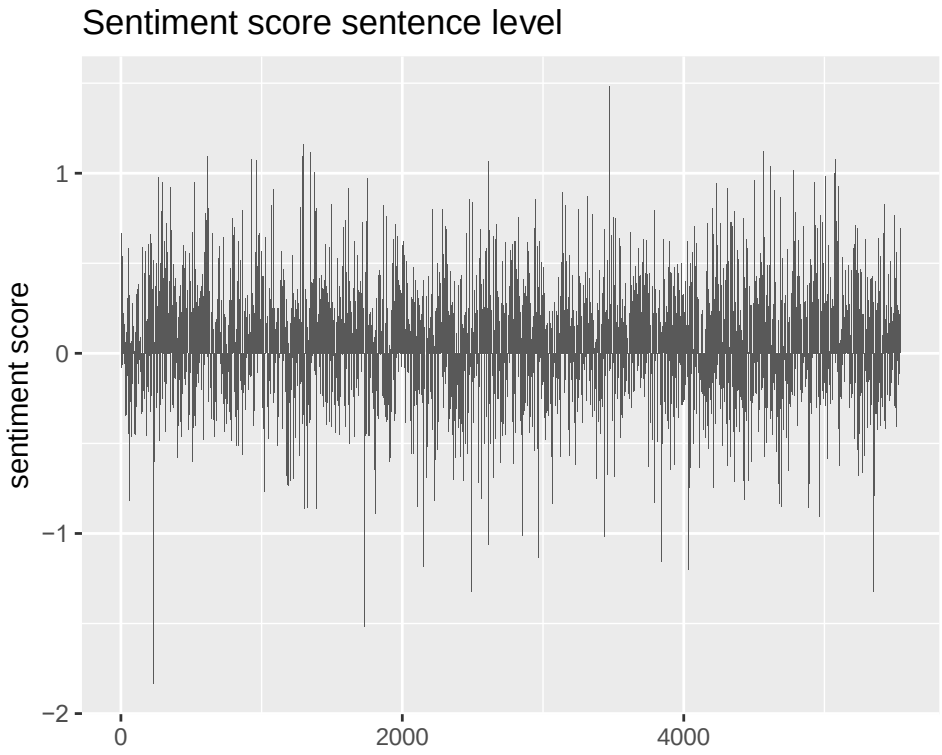
MA615: Text Analysis Tnum

Yongrong Chai

11/29/2021

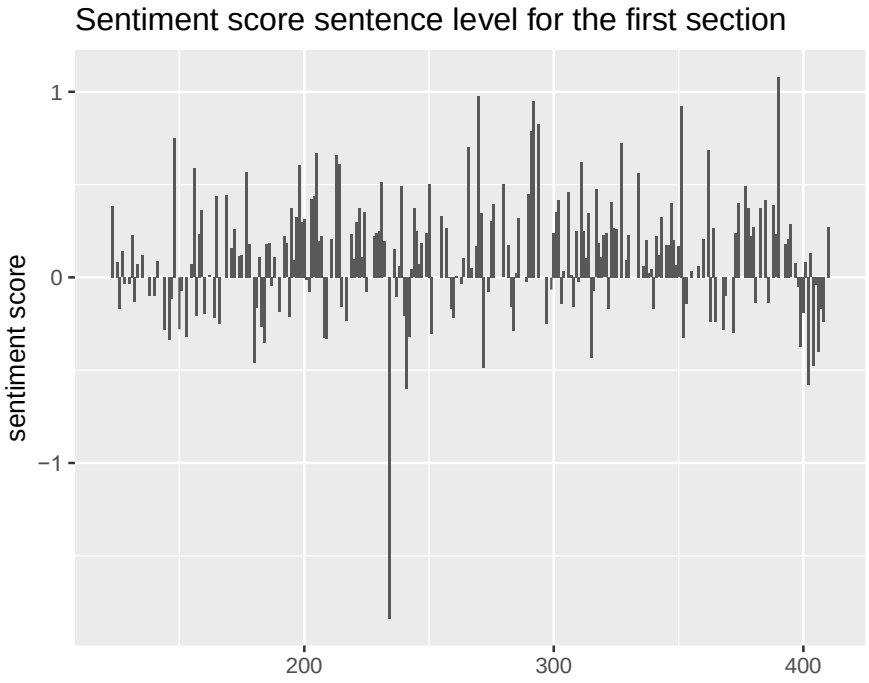
Truenumbers provides data organization and tools that I can analyze my book by sentences.

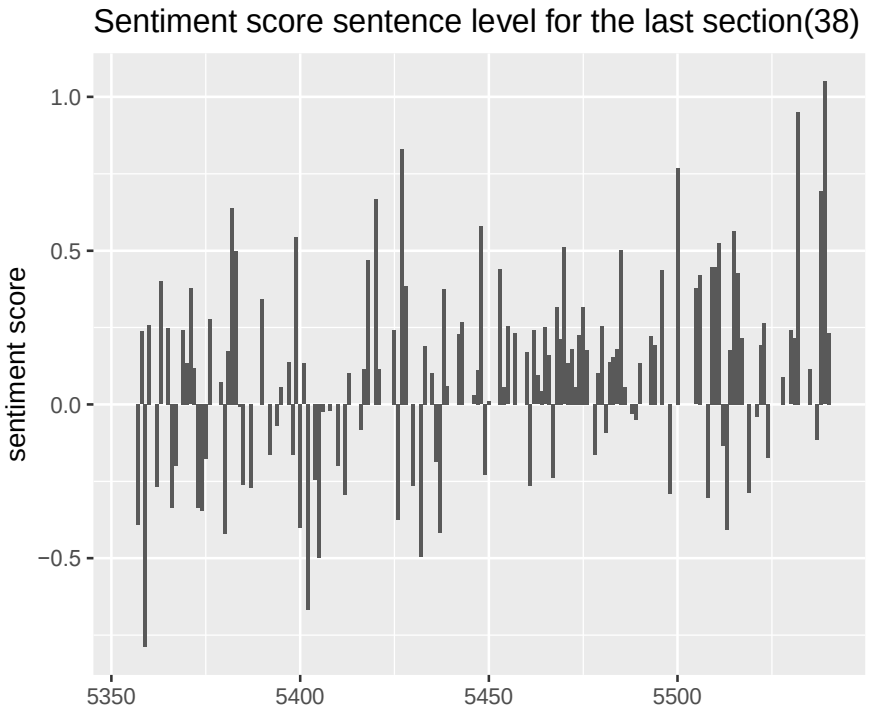
Whole book analysis



This is the plot for the whole book sentences analysis. The positive sentences are much more than negative sentences, it is as same as the result from task2.

Analysis of first section and the last section of the book





In section1, protagonist, Anne, just came to Green Gable. Although there was a lot of joy, there was also a bad situation that she was not welcome, so extreme negative sentences would appear here In the last section, the ending was perfect, full of joy and gratitude, but there were also stories of family deaths and friends parting, but overall there were more positive sentences.

Comparison sentiment_by with nrc and bing

I used a for loop to cut out every word in each sentence and then find the sentiment score in the corresponding NRC and Bing. Then sum them by sentence and divide by the total number of words in each sentence to get the avg_sentiment score in the same form as sentiment_by.

element_id <int>	word_count <int>	sd <dbl>	ave_sentiment <dbl>	bing_sum <dbl>	nrc_sum <dbl>
1	285	NA	0.0533113990	-0.03157895	-0.003508772
2	89	NA	-0.0818848362	-0.01123596	0.022471910
3	60	NA	0.2904737510	0.03333333	0.033333333
4	49	NA	0.3589285714	0.02040816	0.040816327
5	37	NA	0.6658158986	0.05405405	0.081081081
6	57	NA	0.0397359707	-0.01754386	0.000000000
7	14	NA	0.0000000000	0.00000000	0.000000000
8	36	NA	0.0000000000	0.00000000	0.000000000
9	9	NA	0.0000000000	0.00000000	0.000000000
10	35	NA	0.5408987230	0.08571429	0.114285714

1–10 of 5,540 rows | 4–9 of 9 columns

Previous 1 2 3 4 5 6 … 100 Next

Bing:

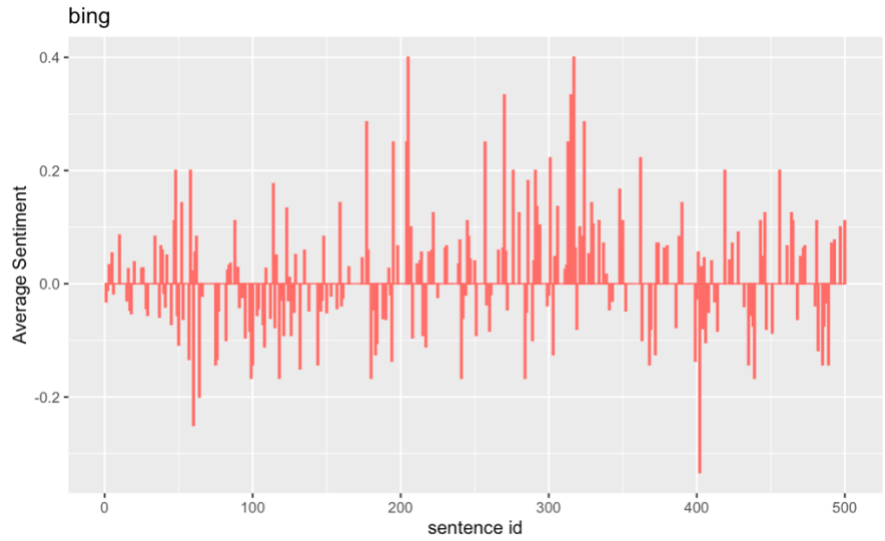
The vocabulary is relatively small, and the overall storyline is similar to that of the NRC. The sentence has positive peaks at 200 and 300.

tnum:

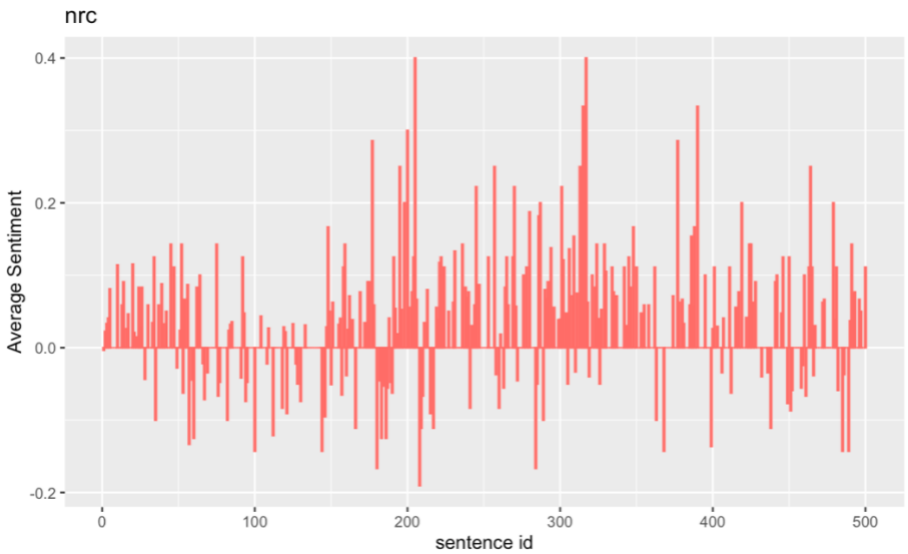
The vocabulary is relatively large, and the overall plot trend is very different from the other two. One of them is negative.

NRC：

The vocabulary is relatively larger, and the overall plot trend is similar to Bing. The sentence has positive peaks at 200 and 300. The reason Bing and NRC are more similar is probably because the avg_sentiment score algorithm is the same.

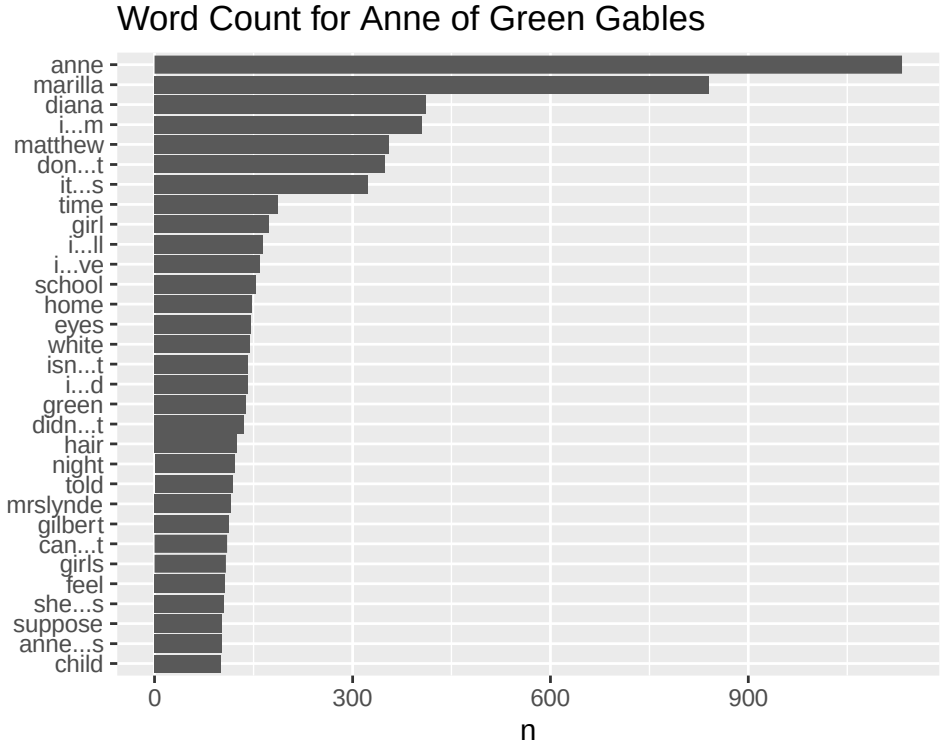

bing
0.4
0.2
0.0
-0.2
Average Sentiment
0
100
200
300
400
500
sentence id

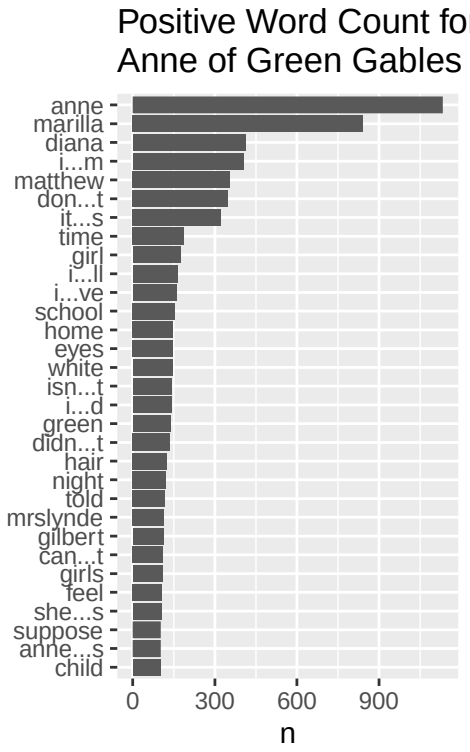

tnum
1
0
-1
Average Sentiment
0
100
200
300
400
500
sentence id

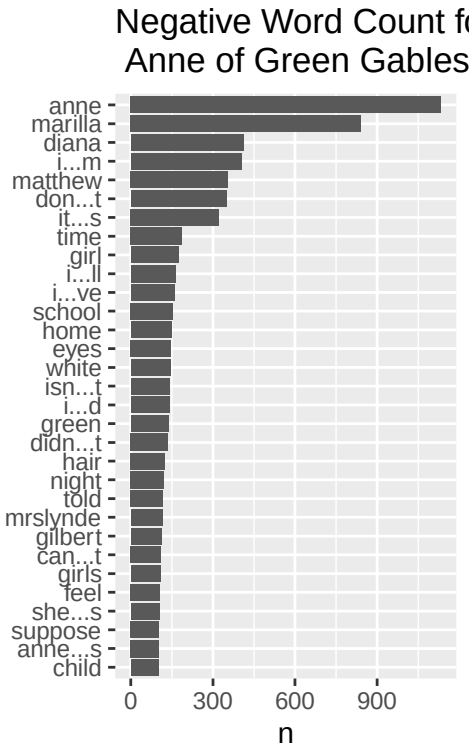

nrc
0.4
0.2
0.0
-0.2
Average Sentiment
0
100
200
300
400
500
sentence id

Extra Credit

```
# A tibble: 2,308 × 2
   word           sentiment
   <chr>          <chr>
 1 abba           positive
 2 ability        positive
 3 abovementioned positive
 4 absolute       positive
 5 absolution     positive
 6 absorbed       positive
 7 abundance      positive
 8 abundant       positive
 9 academic       positive
10 academy        positive
# … with 2,298 more rows
```




Positive Word Count fo
Anne of Green Gables
anne
marilla
diana
i...m
matthew
don...t
it...s
time
girl
i...ll
i...ve
school
home
eyes
white
isn...t
i...d
green
didn...t
hair
night
told
mrslynde
gilbert
can...t
girls
feel
she...s
suppose
anne...s
child
0
300
600
900
n


Negative Word Count f
Anne of Green Gables
anne
marilla
diana
i...m
matthew
don...t
it...s
time
girl
i...ll
i...ve
school
home
eyes
white
isn...t
i...d
green
didn...t
hair
night
told
mrslynde
gilbert
can...t
girls
feel
she...s
suppose
anne...s
child
0
300
600
900
n

Reference：

https://www.r-bloggers.com/2020/04/sentiment-analysis-in-r-with-sentimentr-that-handles-negation-valence-shifters/

https://www.gutenberg.org/ebooks/45

https://learn.bu.edu/ultra/courses/_80585_1/cl/outline

https://www.tidytextmining.com/sentiment.html