Assignment4

Andrew Sisitzky

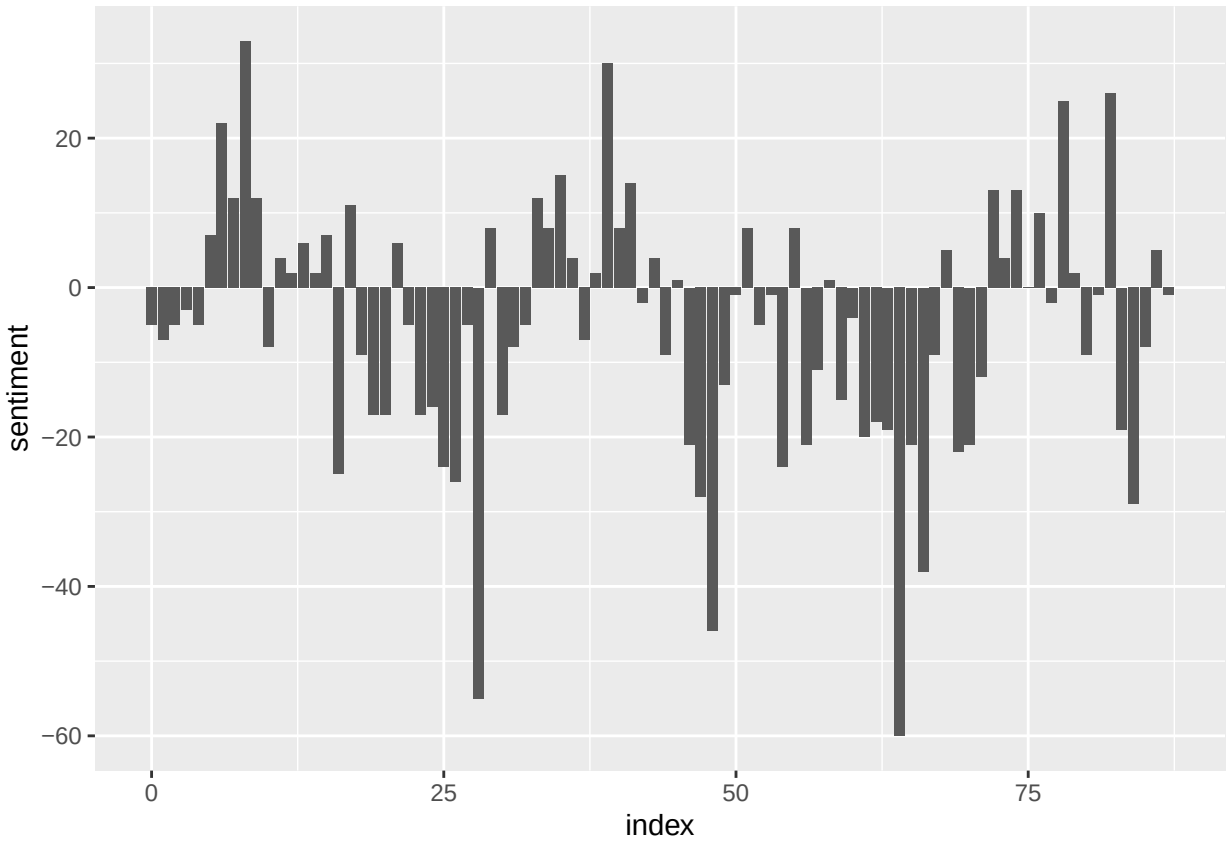
12/8/2021

PARTS I AND II

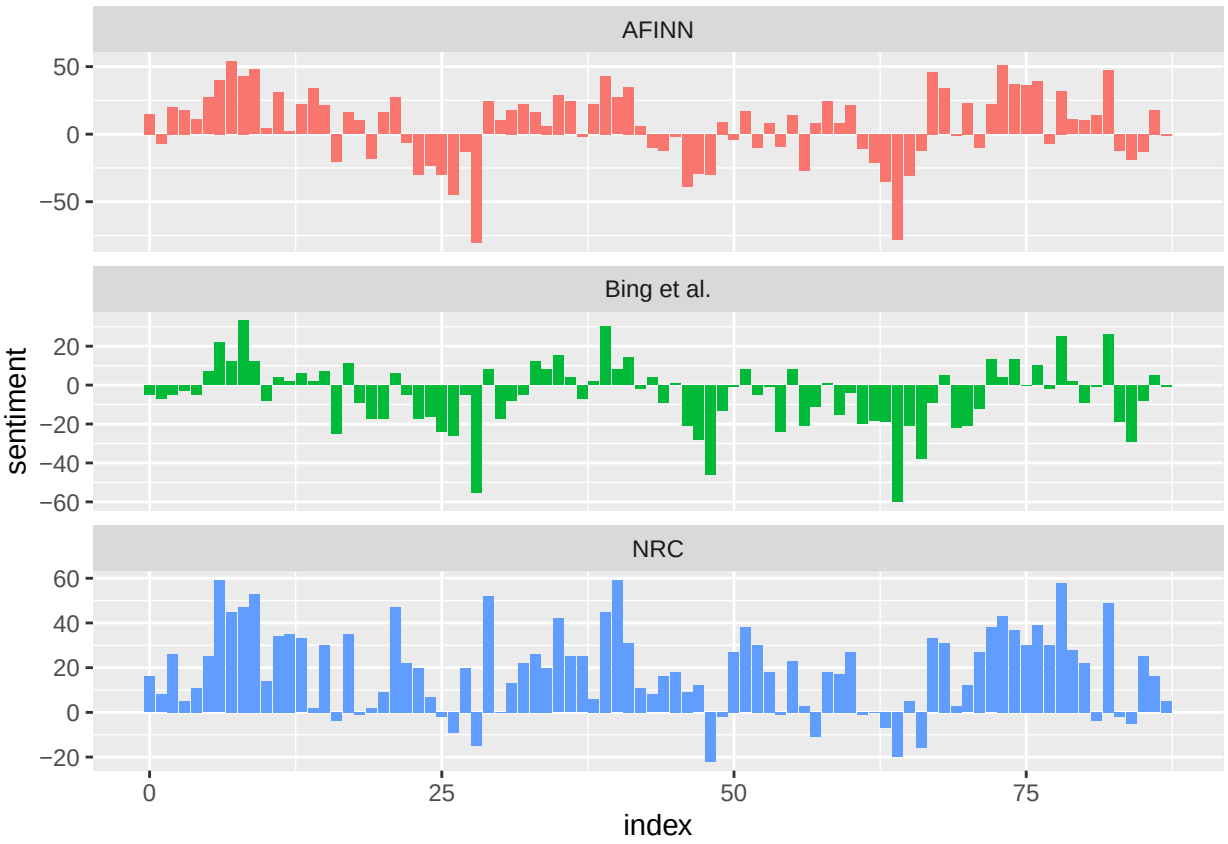
The book that I chose to analyze is The Scarlet Letter by Nathaniel Hawthorne.

```
## # A tibble: 495 x 2
##    word         n
##    <chr>    <int>
##  1 sin         44
##  2 death       42
##  3 evil        42
##  4 shame       40
##  5 infant      32
##  6 prison      30
##  7 scaffold    30
##  8 change      27
##  9 grave       26
## 10 god         22
## # ... with 485 more rows
```

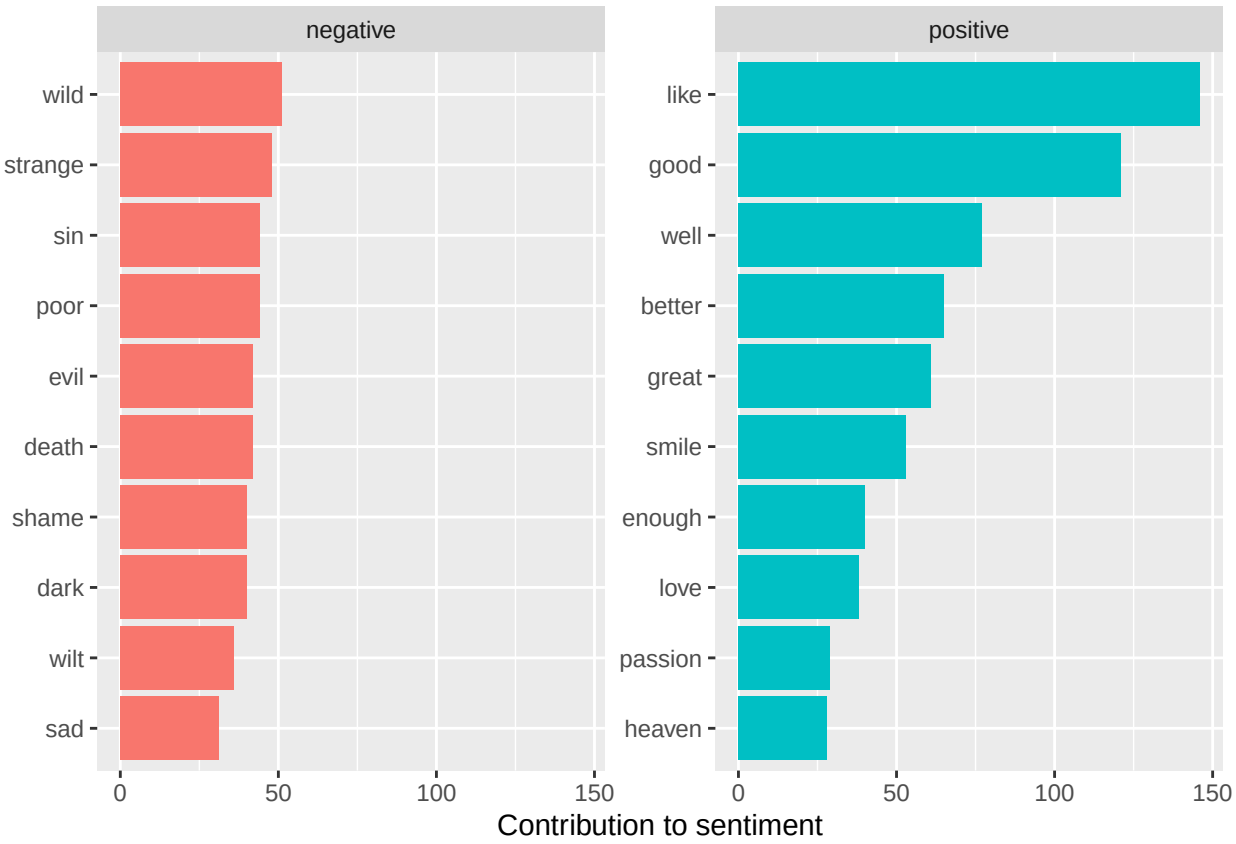
Using the nrc lexicon, I chose to highlight words that were classified under 'fear' as I found this to match up well with the plot of the novel. The words that came up the most in the book were sin, death, evil, and shame.



Then using the bing lexicon, I plotted the sentiment over the course of the book, indexing by 100 lines. In this plot, we see that this book generally using negative words and has three major peaks during which many negative words were used.



I then plotted the sentiemnt analysis of this book using all three lexicons. Here we can see that they each show different levels at every stage. While the bing lexicon we looked at before grades the words as mostly negative throighout the book, we see that AFINN and NRC both tend to show more positive levels. While all three show three distinct negative peaks throughout the story, they all show these peaks to a different degree. Overall, NRC seems to show a generally positive grade with the three peaks looking rather small compared to what is shown by the other lexicons.



When looking at the bing word count, we can see that positive words such as like and good tend to have a higher contribution to the sentiment than other positive words. In terms of negative words, it seems that they tend to have less of an extreme contribution to the sentiment, but rather a higher quantity tend to have an equally high contribution to sentiment.

```
## Joining, by = "word"
```

This word cloud that I created highlights the 100 most commonly used words and organizes them based on their sentiment, positive or negative.

PART III

For part III, I began by authorizing the mssp1.bu.edu server and setting the tnum test space to "test 3"

```
tnum.authorize("mssp1.bu.edu")
```

```
## Available spaces: testspace, MEPED, alion-rf, shared-testspace, test2, alion, NCNM, ED-900-Workshop,
```

```
## Numberspace set to: testspace
```

```
tnum.setSpace("test3")
```

I then downloaded the Scarlet Letter from gutenberg, made some edits in a text file, and re-read it into R.

```
scarlet_letter2 <- gutenberg_download(gutenberg_id=25344)
scarlet_letter2 <- readLines("pg25344.txt")
```

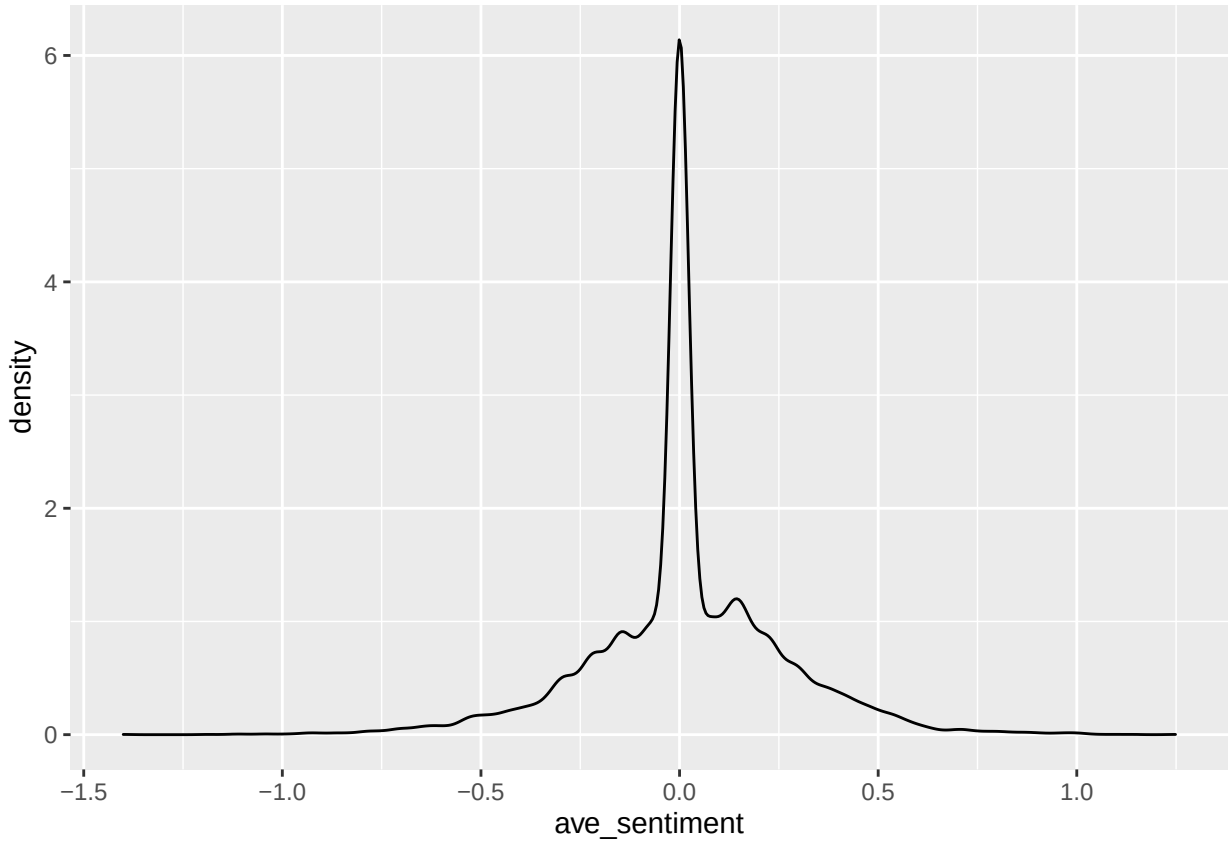
Then, I ingested the text file into the number space. For the purpose of this code running, I commented out this line of code. (it took 4 hours to run) It is saved under sisitzky/scarlet.

```
## tnBooksFromLines(scarlet_letter2, "sisitzky/scarlet")
tnum.getDBPathList(taxonomy="subject", levels=2)
```

```
##  [1] ""                                "lewiscarrol/alice"
##  [3] "carrol/alice"                    "carroll/alice"
##  [5] "elisa/the_call_of_the_wild"      "handing/hw4"
##  [7] "zara/hw4"                        "handing/sea"
##  [9] "zara/A4"                         "zara/a4"
## [11] "elisa/wild"                      "dostoevsky/hw4"
```

```
## [13] "dostoevsky/crime_and_punishment" "handing/game1"
## [15] "handing/game2"                   "sisitzky/scarlet"
## [17] "william/test3"
```

After ingesting the file into the number space, I then used the package sentimentr to do some more sentiment analysis of the text, creating this plot.



This final plot shows the avergae sentiment on a sentence level for the book. Here we see that on a sentence level, the postive and negative sentiments tend to be rather equal, with a slight edge going to positive in terms of density, while neutral sentiments seem to be the most common by a large margin.