THE RED AND THE BLACK

Clare Tang

2021/11/27

Backgroud

THE RED AND THE BLACK is a historical psychological novel in two volumes written by Stendhal, published in 1830. It chronicles the attempts of a provincial young man to rise socially beyond his modest upbringing through a combination of talent, hard work, deception, and hypocrisy. He ultimately allows his passions to betray him.

Julien Sorel, the Bildungsroman, is an intelligent and ambitious protagonist. He comes from a poor family and fails to understand much about the ways of the world he sets out to conquer. He harbors many romantic illusions, but becomes mostly a pawn in the political machinations of the ruthless and influential people about him. The adventures of the hero satirize early 19th-century French society, accusing the aristocracy and Catholic clergy of being hypocritical and materialistic, foretelling the radical changes that will soon depose them from their leading roles in French society.@Manual{ url = {https://en.wikipedia.org/wiki/The_Red_and_the_Black}}

Gutenberg analysis

The book id for *THE RED AND THE BLACK* is 44747 on gutenberg website. The original dataframe contains the id and text of the book in unorganized form. After using function unnest_tokens and filtering out the words that do not have real meanings from stop_word database, each observation only include a meaningful word extracted from the book. Plus, the line number and the chapter the words belong to are also included. Analysis of sentiment uses Text mining in R as reference.

```
## # A tibble: 6 x 4
##   gutenberg_id linenumber chapter word
##          <int>      <int>   <int> <chr>
## 1        44747          1       0 the
## 2        44747          1       0 red
## 3        44747          1       0 and
## 4        44747          1       0 the
## 5        44747          1       0 black
## 6        44747          3       0 a
```

Word clouds for the whole book

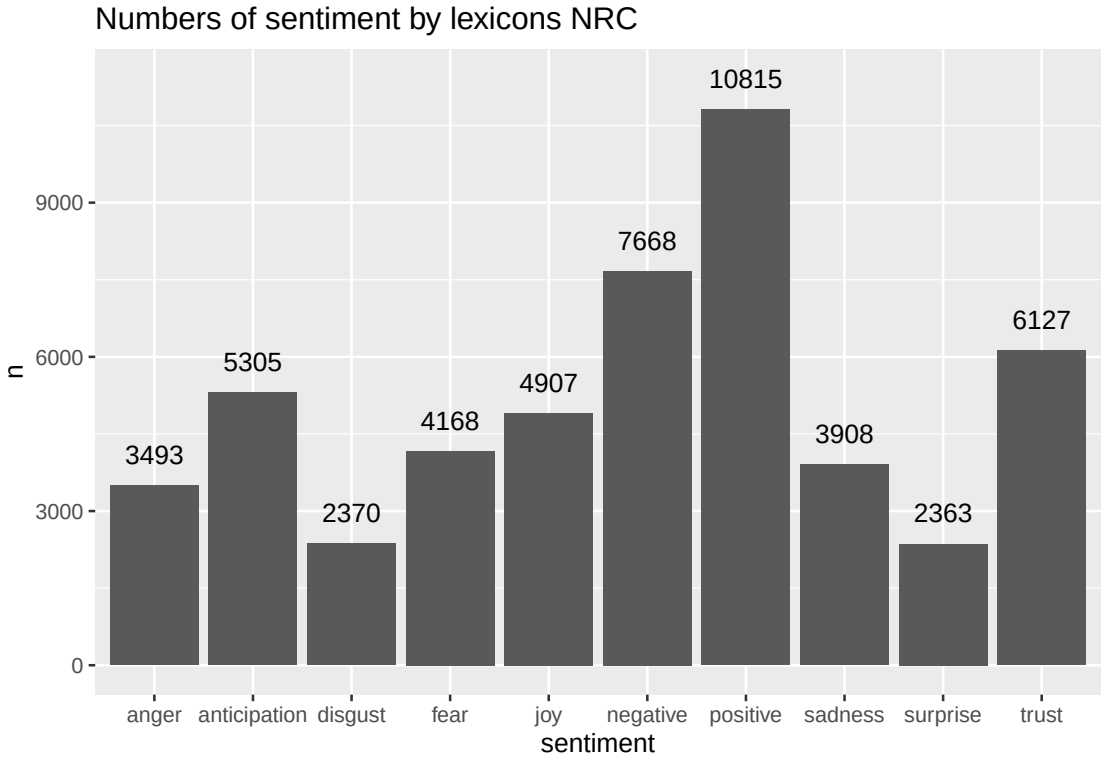
1. Wordcloud below shows 100 most common used words in *THE RED AND THE BLACK*. The size of the words in the wordcloud represents the frequency of the words shown in the book. Apparently, the hero name julien appears the most. “de” means “of” in English.
2. Wordcloud below shows most commonly used words that are positive or negative in the book. For example, ‘death’ and ‘poor’ are often used in describing the plots. ‘great’, ‘like’ and ‘love’ are frequently used and it makes sense because the book involves the love story of julien.

Sentiment analysis of words

- There are three main lexicon used when doing sentiment analysis for words. Here follow their results and another lexicon called loughran is also analyzed.

Lexicon NRC

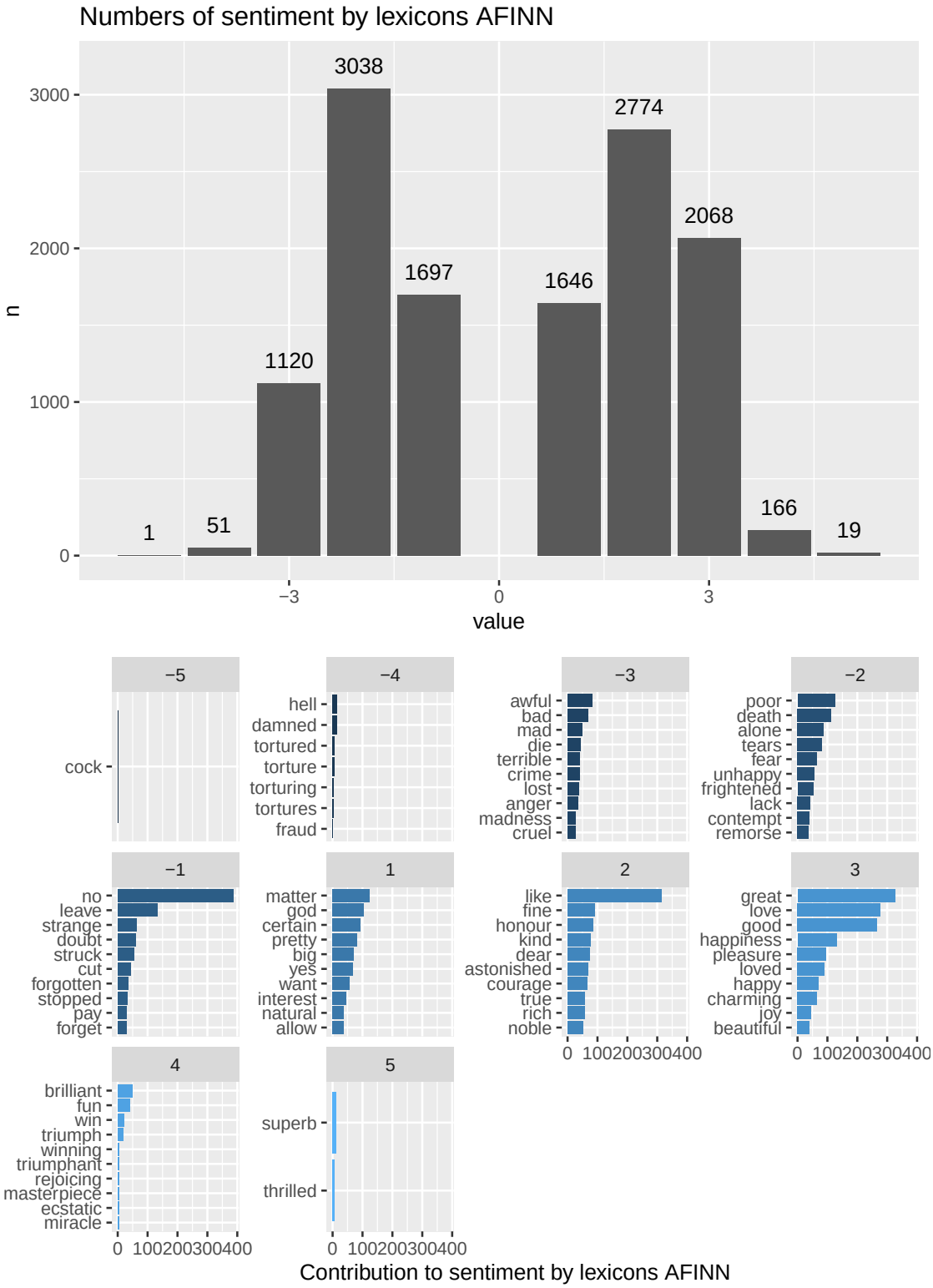
- There are 10 types of sentiment levels in NRC: trust, surprise, positive, joy, anticipation, fear, negative, sadness, anger, disgust.
- The graphs below represent:
  1. The number of words in each level of sentiment
  2. 10 mostly used words in each level of sentiment
  3. Wordcloud of 100 mostly used words in the 'joy' sentiment level
  4. Sentiment value based on index of 100 (100 promises senitment can be analyzed within the chapter)



Lexicon AFINN

- There are 10 different value of sentiment in AFINN: -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5
- The graphs below represent:
  1. The number of words in each value of sentiment
  2. 10 mostly used words in each value of sentiment and their frequencies
  3. Wordcloud of 100 mostly used words with sentiment value of -2
  4. Wordcloud of 100 mostly used words with sentiment value of 2
  5. Sentiment value based on index of 100 (100 promises senitment can be analyzed within the chapter)

Figure 1: Wordcloud for joy words by lexicon NRC


Numbers of sentiment by lexicons AFINN
n
value
3038
2774
2068
1697
1646
1120
166
51
19
1
3000
2000
1000
0
−3
0
3
−5
−4
−3
−2
−1
1
2
3
4
5
cock
hell
damned
tortured
torture
torturing
tortures
fraud
awful
bad
mad
die
terrible
crime
lost
anger
madness
cruel
poor
death
alone
tears
fear
unhappy
frightened
lack
contempt
remorse
no
leave
strange
doubt
struck
cut
forgotten
stopped
pay
forget
matter
god
certain
pretty
big
yes
want
interest
natural
allow
like
fine
honour
kind
dear
astonished
courage
true
rich
noble
great
love
good
happiness
pleasure
loved
happy
charming
joy
beautiful
brilliant
fun
win
triumph
winning
triumphant
rejoicing
masterpiece
ecstatic
miracle
superb
thrilled
0
100
200
300
400
Contribution to sentiment by lexicons AFINN

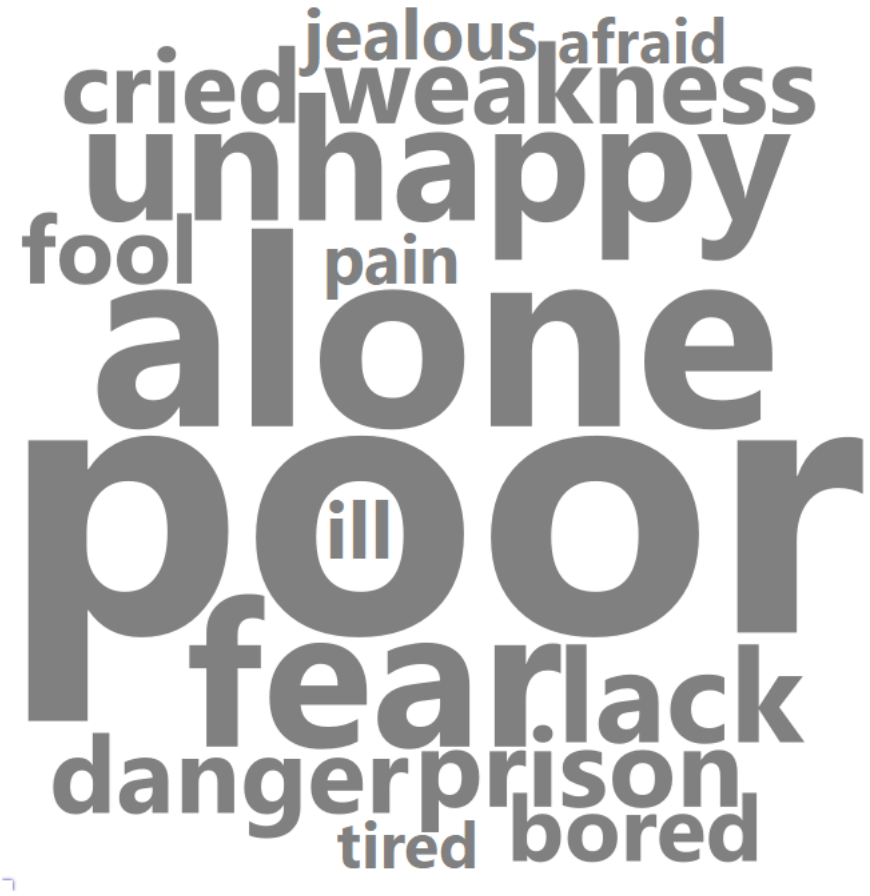
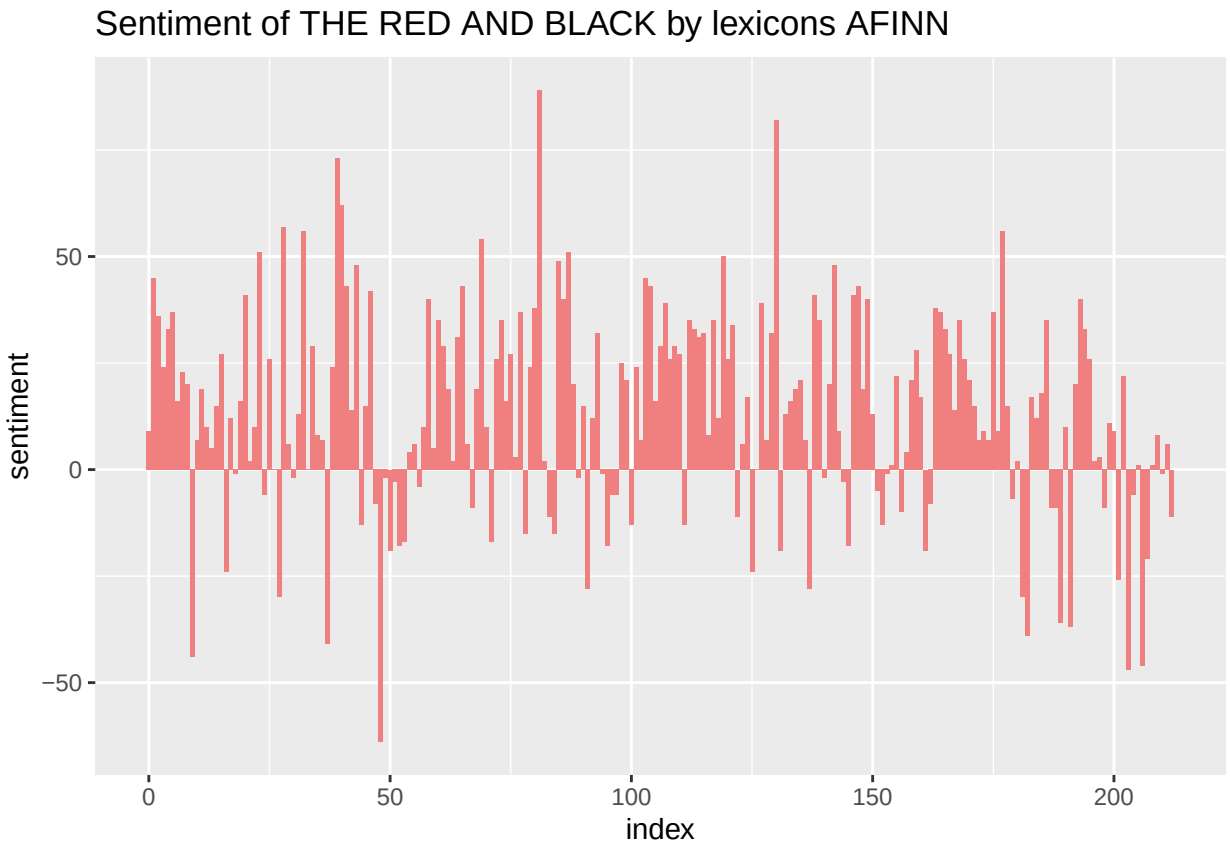


Figure 2: Wordcloud of positive words by lexicon AFIN

Figure 3: Wordcloud of negative words by lexicon AFINN



Lexicon BING

- There are 2 types of sentiment levels in BING: positive and negative
- The graphs below represent:
  1. The number of words in each level of sentiment
  2. 10 mostly used words in each level of sentiment
  3. Wordclouds of 100 mostly used words for each sentiment level
  4. Sentiment value based on index of 100 (100 promises sentiment can be analyzed within the chapter)

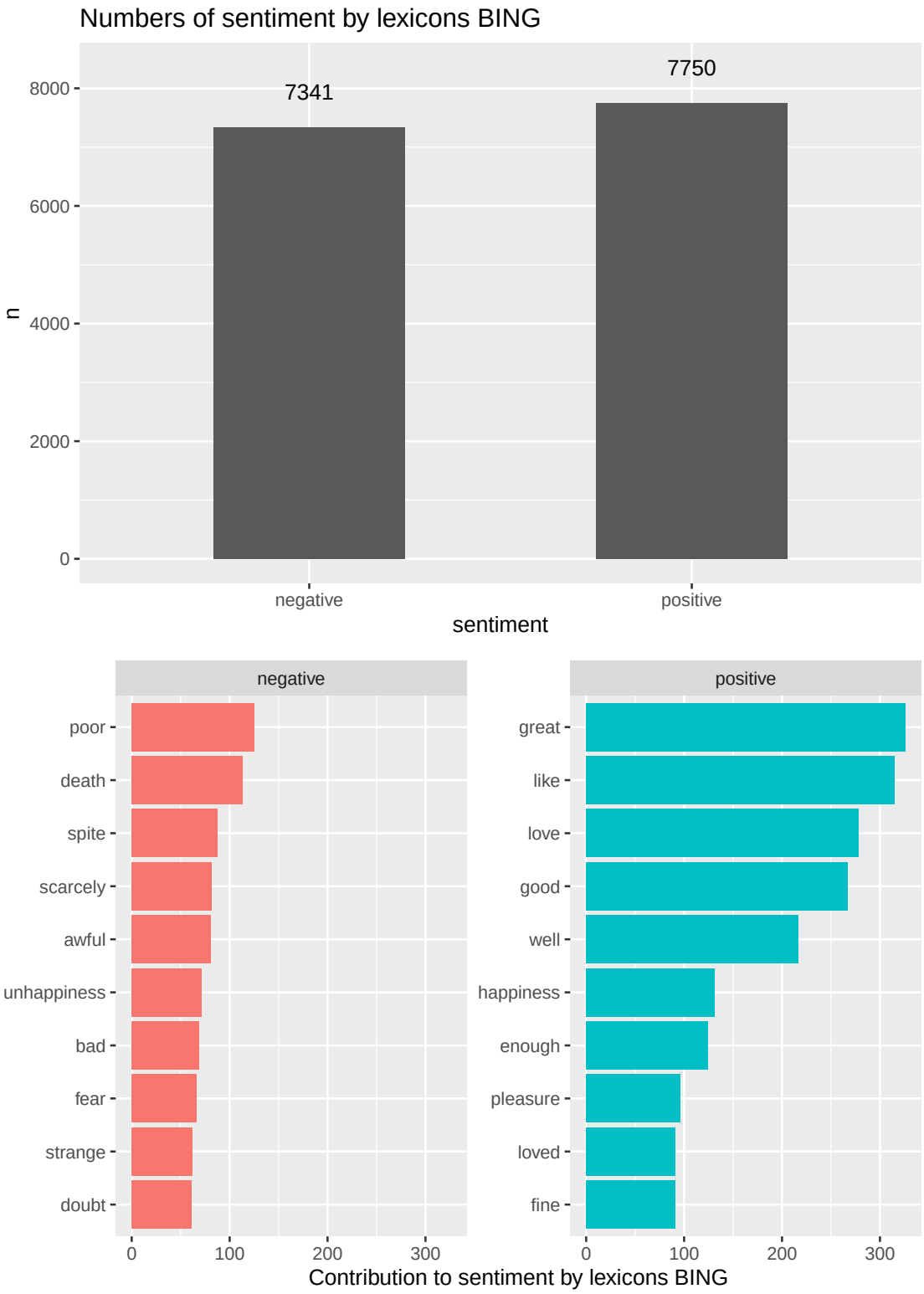
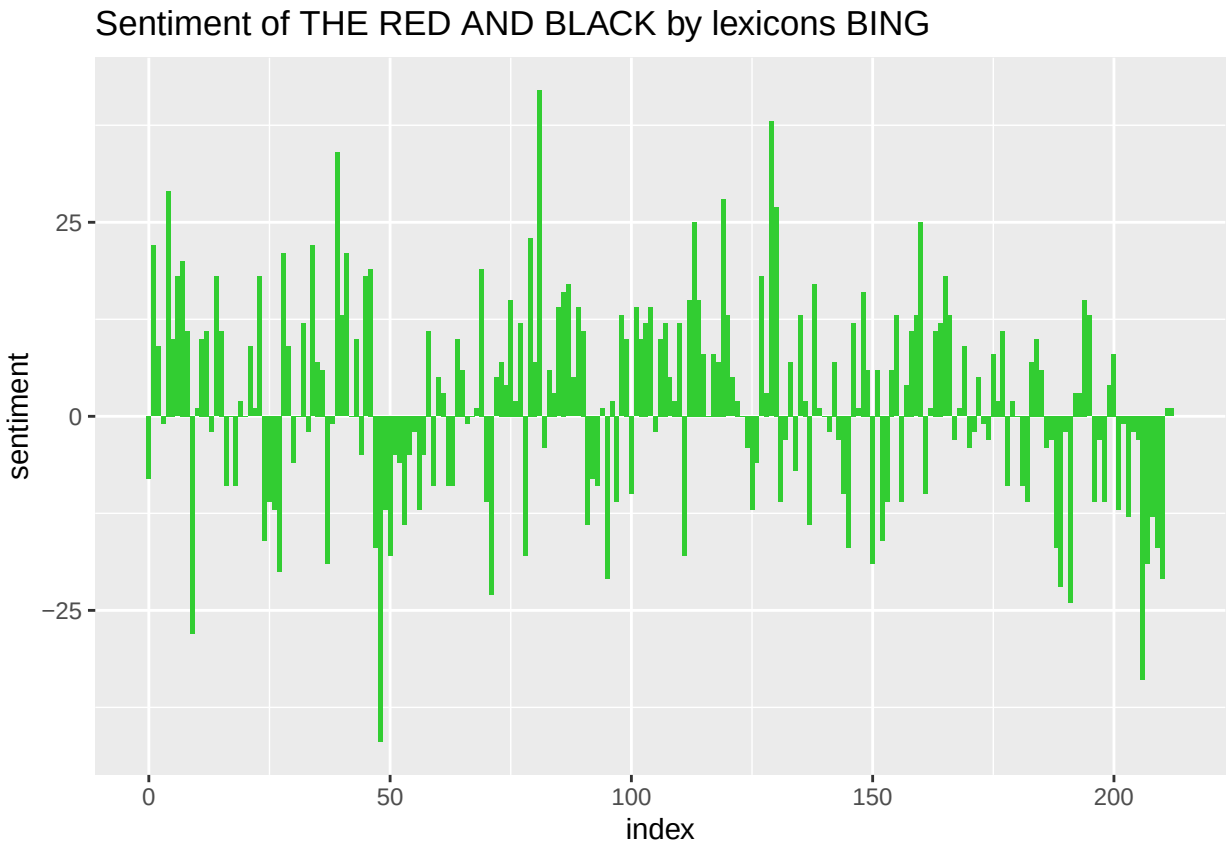

Numbers of sentiment by lexicons BING
7341
7750
8000
6000
4000
2000
0
n
negative
positive
sentiment
negative
positive
poor
death
spite
scarcely
awful
unhappiness
bad
fear
strange
doubt
great
like
love
good
well
happiness
enough
pleasure
loved
fine
0
100
200
300
Contribution to sentiment by lexicons BING

Figure 4: Wordcloud of negative words by lexicon BING

Figure 5: Wordcloud negative words by lexicon BING



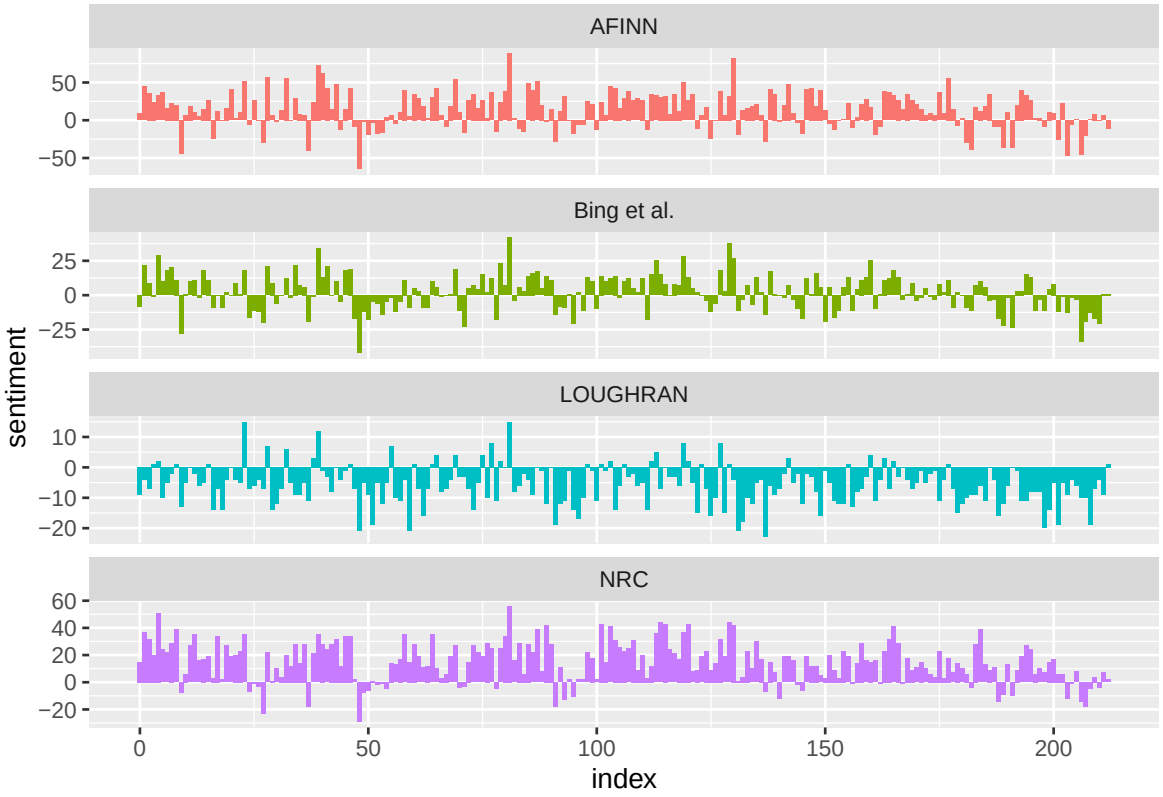
Lexicon LOUGHRAN

- There are 6 types of sentiment levels in LOUGHRAN: negative, positive, uncertainty, litigious, constraining, superfluous
- The graphs below represent:
  1. The number of words in each level of sentiment
  2. 10 mostly used words in each level of sentiment
  3. Wordclouds of 100 mostly used words in positive
  4. The frequency change of each sentiment level based on chapters
  5. Sentiment value based on index of 100 (100 promises sentiment can be analyzed within the chapter)
- We can notice that the levels of sentiment in LOUGHRAN are mostly negative words, and we can also understand from plot 5 that the values of sentiment are mostly below the zero line.

Compare lexicons

- Graphs below shows the comparison of sentiment for four lexicon methods.
- The NRC sentiment is high, the AFINN sentiment has more variance, the Bing et al. sentiment has longer stretches of similar text, but all three agree roughly on the overall trends in the sentiment through a narrative arc. However, sentiment of LOUGHRAN are mostly below the zero line which can be explained by that most sentiment level in lexicon LOUGHRAN are negative classifications. As we can seen from the graphs of numbers of words in each sentiment level above.

Figure 6: Wordcloud of positive words by lexicon LOUGHRAN



- Although the book ends with Julien death, the whole book mainly talks about how he pursed his occupation and love. And Stendhal shows how Julien lost his reputation and got sentenced in the last several paragraphs, which is the main part that most negative words should be used. Given the above reasons, I think using lexicon BING is more proper for this book because there are a concentrated negative value of sentiment in the BING graph and the value of sentiment presents evenly in the previous chapters.

Tnum

- The table below lists first ten rows of the location of the first ten chapters. There are 72 chapters in *THE RED AND THE BLOACK*.

```
##                subject   string.value numeric.value
## 1  randb/heading:0011    "CHAPTER I"            127
## 2  randb/heading:0013   "CHAPTER II"            190
## 3  randb/heading:0016  "CHAPTER III"            246
## 4  randb/heading:0018   "CHAPTER IV"            370
## 5  randb/heading:0020    "CHAPTER V"            448
## 6  randb/heading:0024   "CHAPTER VI"            626
## 7  randb/heading:0026  "CHAPTER VII"            828
## 8  randb/heading:0028 "CHAPTER VIII"           1046
## 9  randb/heading:0030   "CHAPTER IX"           1202
## 10 randb/heading:0033    "CHAPTER X"           1365
```

Analysis of sentenses, paragraphs with example(Chapter 17)

- Here shows the number of words in each sentence in paragraph 17, chapter 17.

```r
#   TAKE CHAPTER 17 AS AN EXAMPLE
## focus on one paragraph -- note the word count for each sentence
q4 <- tnum.query("randb/section:0017/paragraph:0017# has count#")
```

```
## Returned 1 thru 10 of 15 results
```

```r
df4 <- tnum.objectsToDf(q4)
df4 %<>% filter(date == "2021-11-29")
df4 %>% select(subject, property, numeric.value)
```

```
##                                            subject   property numeric.value
## 1 randb/section:0017/paragraph:0017/sentence:0001 count:word           112
## 2 randb/section:0017/paragraph:0017/sentence:0002 count:word            38
## 3 randb/section:0017/paragraph:0017/sentence:0003 count:word            99
## 4 randb/section:0017/paragraph:0017/sentence:0004 count:word            82
## 5 randb/section:0017/paragraph:0017/sentence:0005 count:word           226
```

- Here shows the first sentence of paragraph 17 in chapter 17.

```r
## and now look at the text in a sentence
q5 <- tnum.query("randb/section:0017/paragraph:0017/sentence:0001# has text")
```

```
## Returned 1 thru 3 of 3 results
```

```r
df5 <- tnum.objectsToDf(q5)
df5 %<>% filter(date == "2021-11-29")
df5[, 1:3]
```

```
##                                            subject property
## 1 randb/section:0017/paragraph:0017/sentence:0001     text
##
## 1 ""Well, gentlemen, I shall be the third cure of eighty years of age who has been turned out in thi
```

- Here shows the sentiment score and the number of words in each sentence in paragraph 17 in chapter 17.

```r
## To extract a paragraph of text
q6 <- tnum.query("randb/section:0017/paragraph:0017# has text", max = 10)
```

```
## Returned 1 thru 10 of 15 results
```

```r
df6 <- tnum.objectsToDf(q6) %>% filter(date == "2021-11-29")
para_text <- df6 %>% pull(string.value) %>%
                      str_replace_all("\"","") %>%
                      str_flatten(collapse = " ")

rb1 <- get_sentences(para_text)

## to get sentiment scores by sentence
sentiment(rb1)
```

```
##    element_id sentence_id word_count  sentiment
## 1:          1           1         21  0.3927922
## 2:          1           2          8  0.0000000
## 3:          1           3         20  0.0000000
## 4:          1           4         15  0.2581989
## 5:          1           5         48 -0.8858718
```

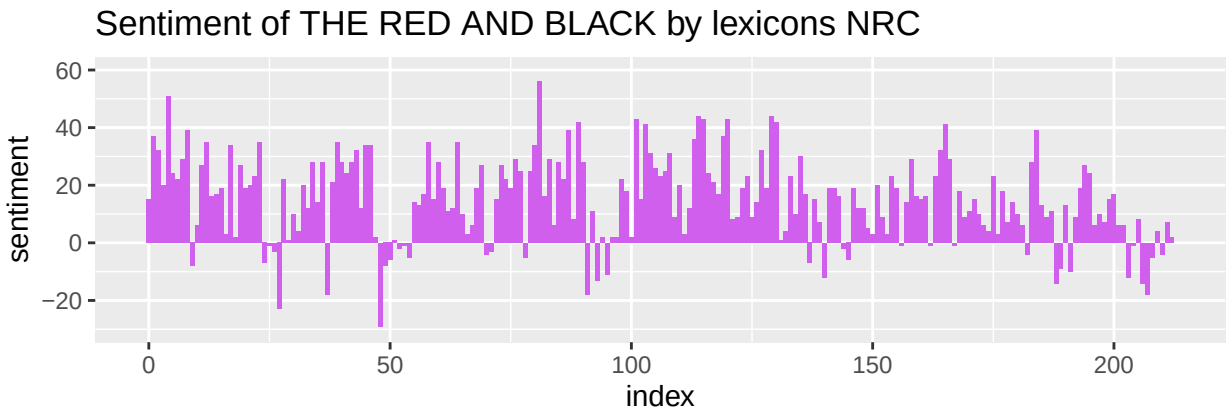
- This is the sentiment score for the whole paragraph 17 in chapter 17.

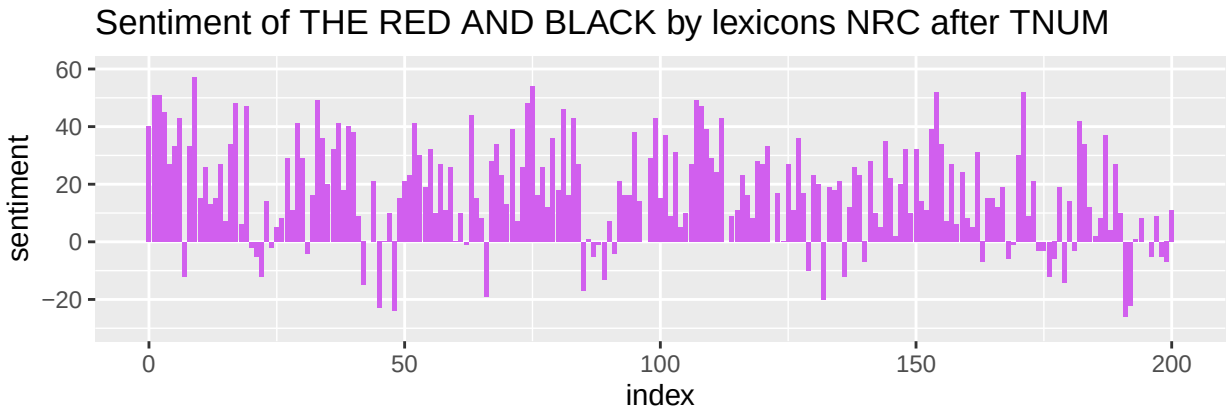
```
## to get sentiment scores aggregated by paragraph
sentiment_by(rb1)
```

```
##    element_id word_count        sd ave_sentiment
## 1:          1        112 0.4986701   -0.05802179
```

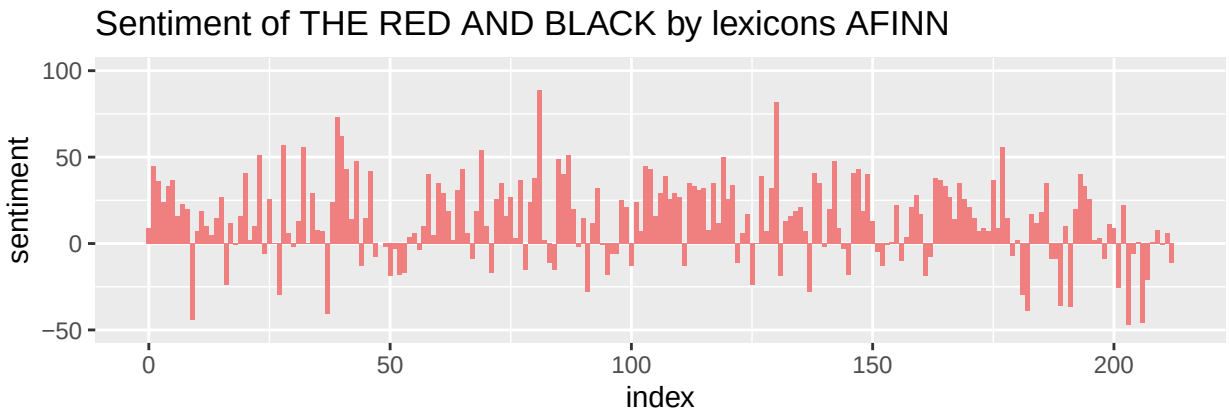
Word analysis after tnum for the whole book

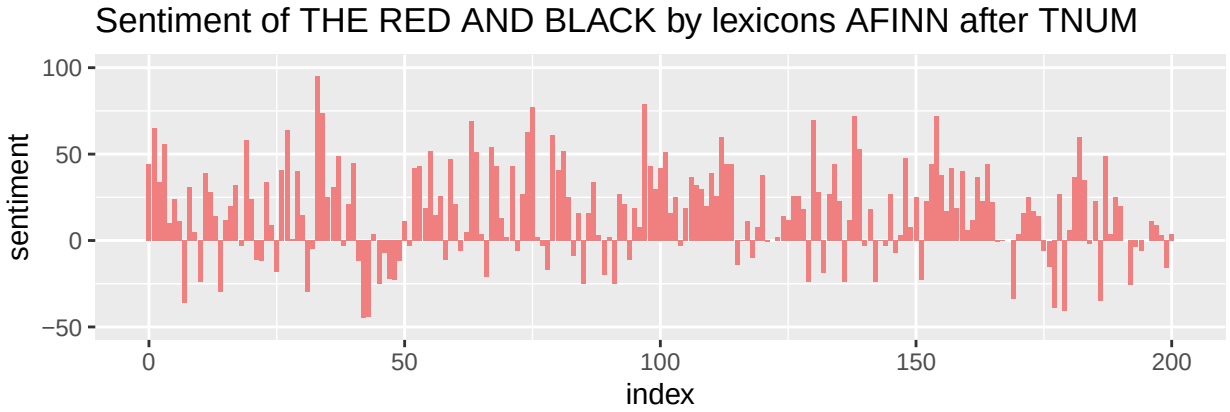
- Change the index to 60 in order to compare the word sentiment analysis in task 2
- The following graphs shows the comparison of sentiment of four kinds of lexicon between the one done before and the one done using TNUM.
- After TNUM, the trends of sentiment looks similar to the original one under the lexicon NRC.



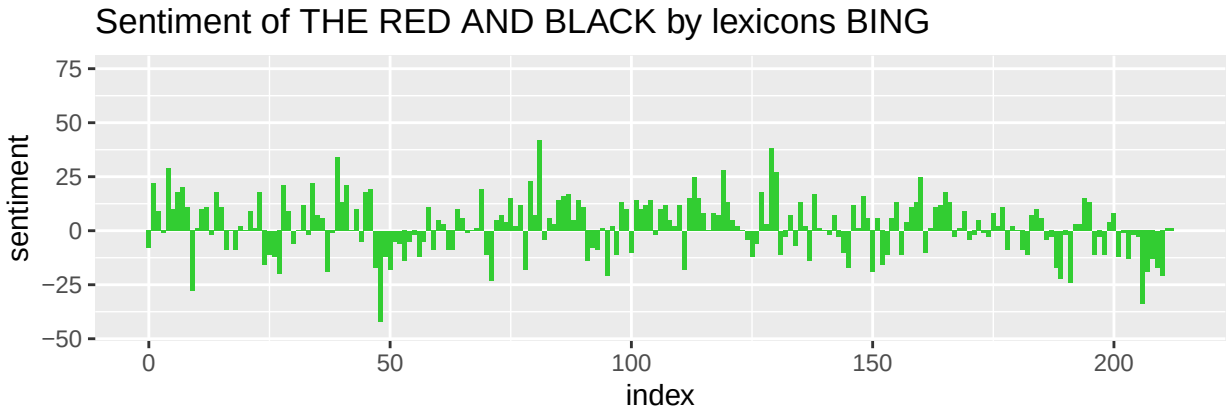


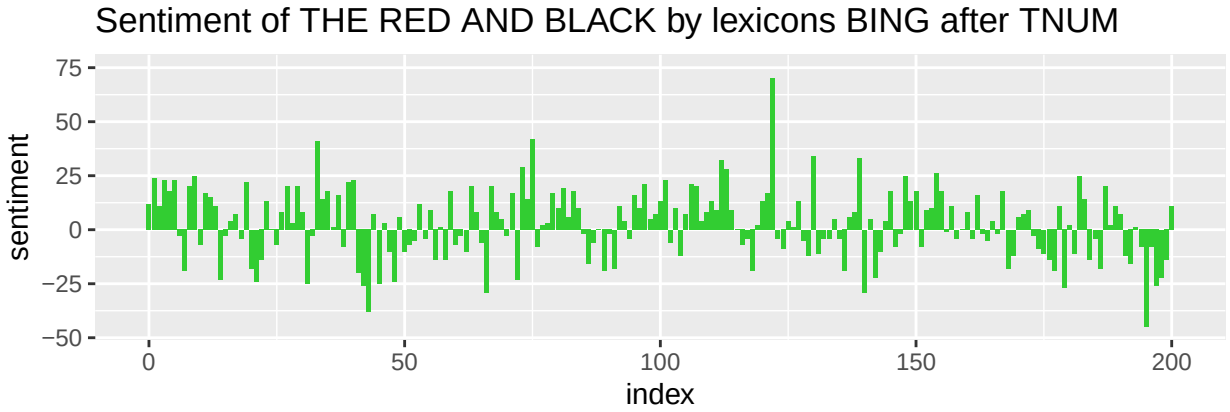
- After TNUM, the trends of sentiment also looks similar to the original one under the lexicon AFINN.



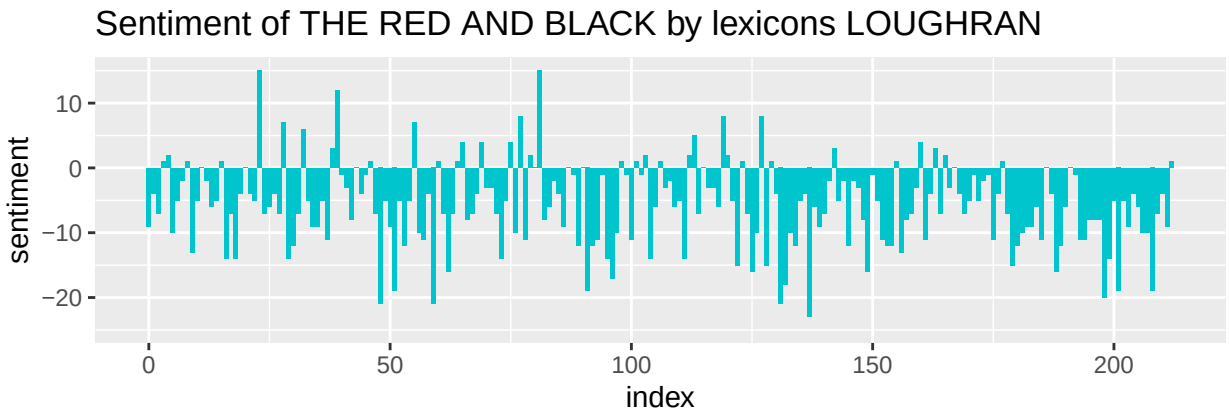


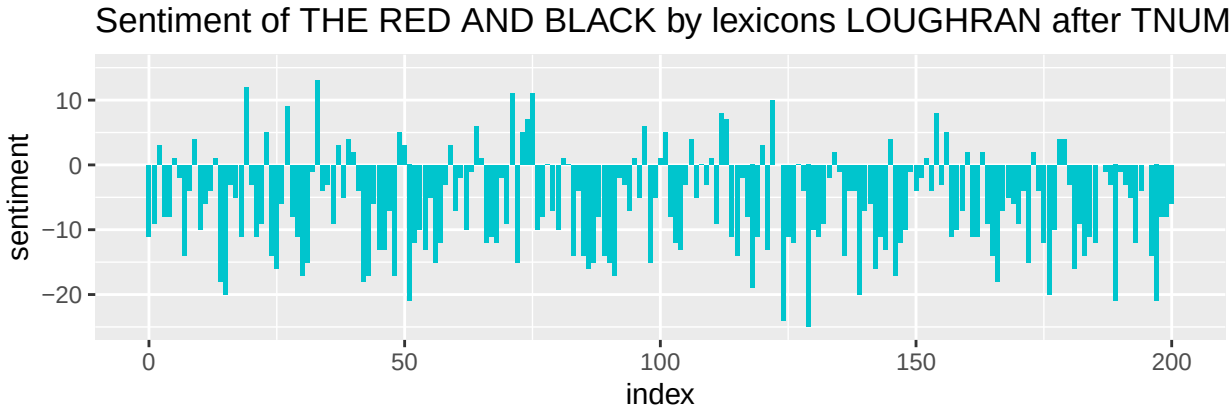
- After TNUM, the absolute value of sentiment changes little compared to the original one under the lexicon BING. Higher value of sentiment of index can also be noticed.





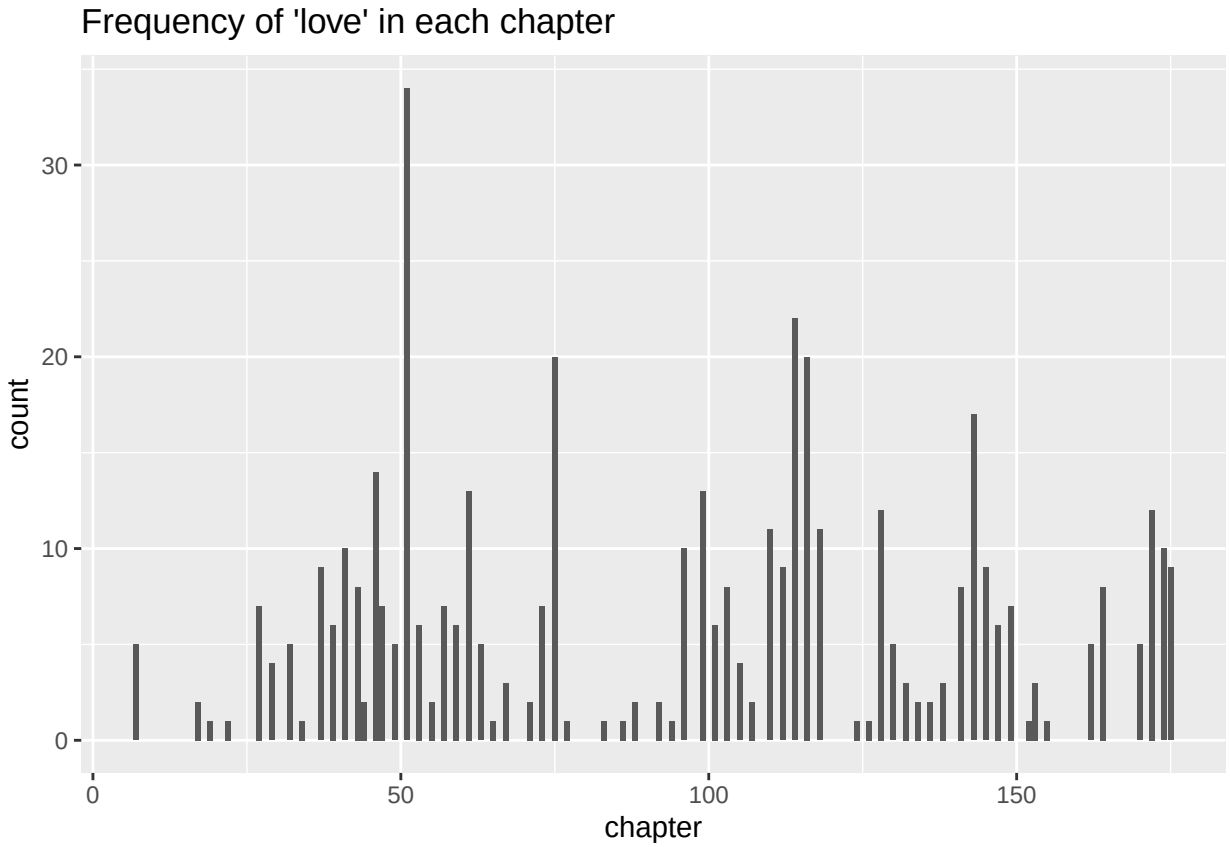
- After TNUM, the trends of sentiment also looks similar to the original one under the lexicon LOUGHRAN.



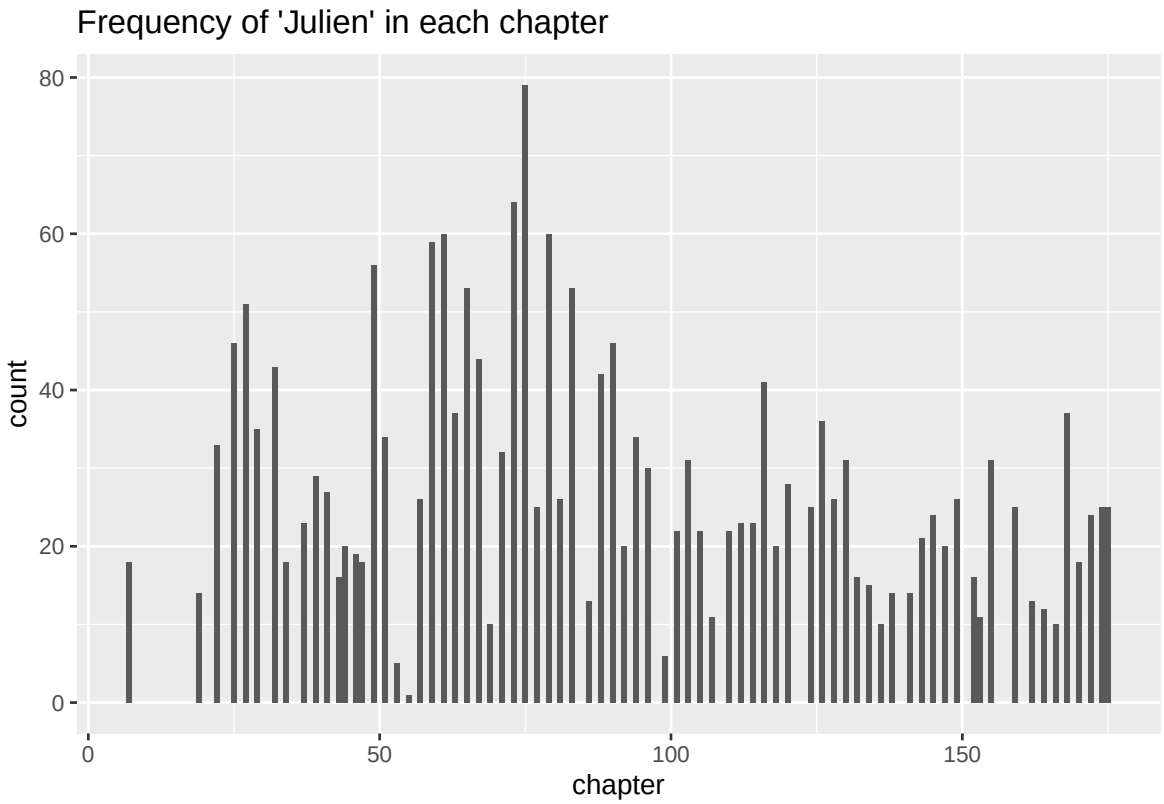


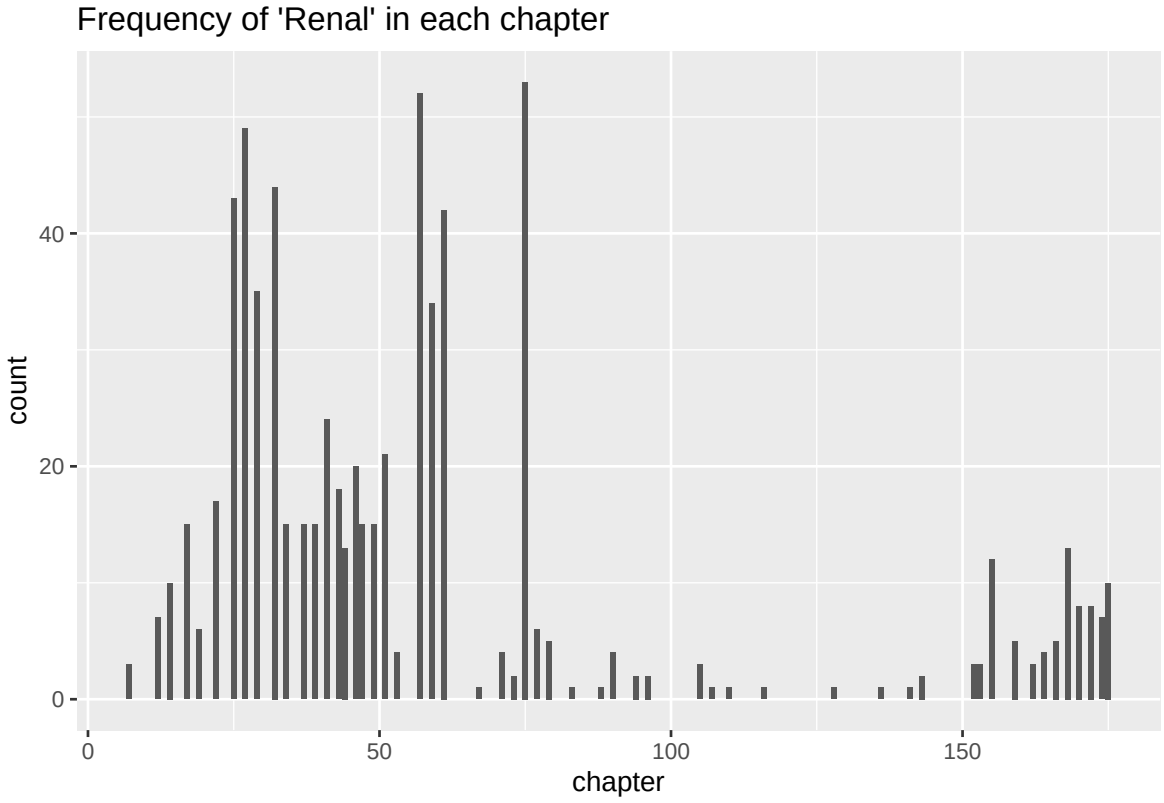
Create tags

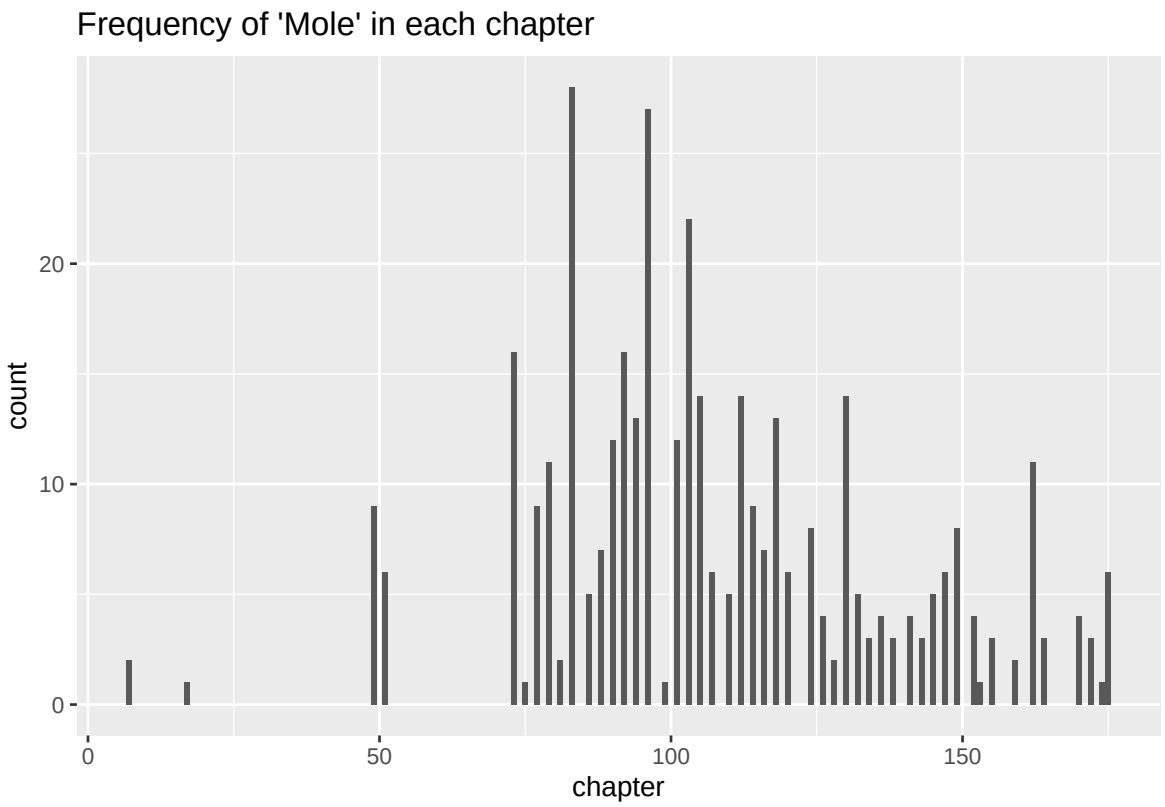
- Since *THE RED AND THE BLACK* includes two love stories: Julien and Madame de Rênal, Julien and Mathilde de la Mole, tag 'love' is created. From the above analysis, word love appears in the book for 278 times.
- Plot of numbers of 'love' in each paragraph shows the love story goes through the whole book.



- Here follows plots of number of words in each chapter for three main character: Julien, Renal and Mole.
- An interesting point can be noticed that Renal mostly presents in the first half of all the chapters and it makes sense because Julien first fell in love with Renal. Similarly, Mole presents in the second half of the whole chapters and it is because then Julien met Mole.


Frequency of 'Julien' in each chapter
count
chapter
0
20
40
60
80
0
50
100
150


Frequency of 'Renal' in each chapter
count
chapter
0
20
40
0
50
100
150


Frequency of 'Mole' in each chapter
count
20
10
0
0
50
100
150
chapter