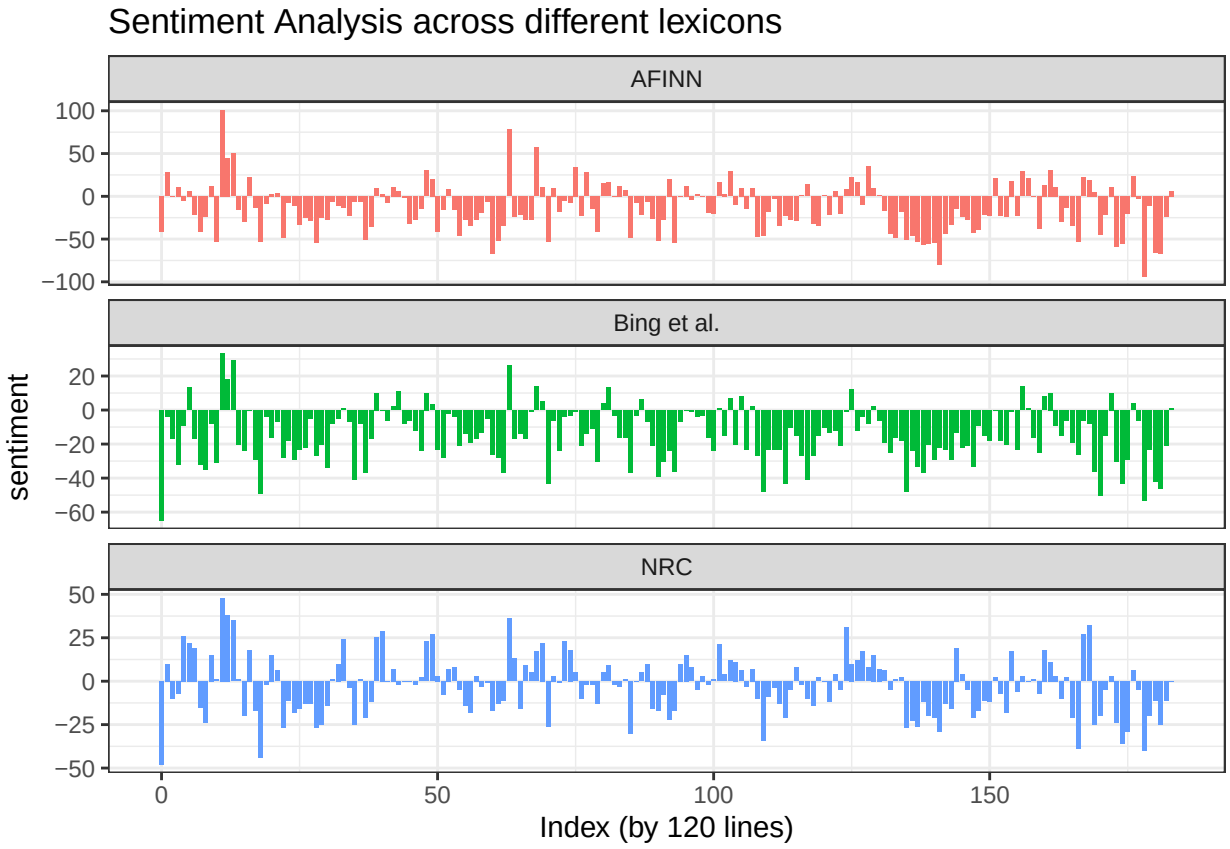
Crime and Punishment Text Analysis

Kayla Choi

12/8/2021

Task 2: Bag of Words

I chose to analyze Crime and Punishment by Fyodor Dostoevsky. This is one of my favorite books out of the ones I had to read in high school.
I downloaded the txt file from the Gutenberg Project. Excluding the foreword and the ending citation of Gutenberg project, I turned it into a tibble.



Generally, the visualization of the sentiment analysis matches the plotline of the book. The book is split into 6 parts with 1 epilogue. In the first part, the protagonist Raskolnikov commits murder; the next 5 parts describe in detail the moral anguish that Raskolnikov faces and the slow hunt as the detective gets closer to figuring him out. It makes sense that the sentiment through most of the book is negative. I was surprised to see that there wasn't much a huge positive sentiment difference at the very end of the book. In the epilogue, Raskolnikov confesses to his crime, and though he ends up in prison, he begins to accept unconditional love from Sonya and turns towards religion to start his journey towards redemption and moral restitution. That's why I anticipated seeing such a large positive spike at the end, since it is the first time in the book where he

is open to concepts such as "friendliness" and "love" and "redemption." However, it makes sense that there isn't a huge peak because he is still in prison, so the words probably pick up on that as a negative sentiment.

The different lexicons use different methods of categorizing sentiment. For example, NRC assigns factor(s) of sentiment, like trust, fear, or sadness, while Bing labels "negative" or "positive." Afinn assigns smaller scale values to different words on a numeric scale (like -2 for negative sentiment). All of the lexicons seem to pick up on the general trend of the sentiment throughout the novel. The only differences that I see are a few select indices that differ in their sign (e.g. index 130).

Task 3: sentence-level analysis

I've used tnum to load the book into test3 numberspace (not test2, it was running too slowly).

```
q1[[1]]
```

```
## [1] "\"CRIME AND PUNISHMENT  By Fyodor Dostoevsky CRIME AND PUNISHMENT     PART I    CHAPTER I  On a
## attr(,"tags")
## list()
## attr(,"class")
## [1] "tnum"
## attr(,"subject")
## [1] "dostoevsky/crime_and_punishment/section:0000/paragraph:0009/sentence:0001"
## attr(,"property")
## [1] "text"
## attr(,"guid")
## [1] "d257a079-a91c-4246-91c5-4eb3a1cdb613"
## attr(,"date")
## [1] "2021-12-09"
```

```
head(d1)
```

```
##                                                                  subject
## 1 dostoevsky/crime_and_punishment/section:0000/paragraph:0009/sentence:0001
## 2 dostoevsky/crime_and_punishment/section:0000/paragraph:0009/sentence:0002
## 3 dostoevsky/crime_and_punishment/section:0000/paragraph:0009/sentence:0003
## 4 dostoevsky/crime_and_punishment/section:0000/paragraph:0010/sentence:0001
## 5 dostoevsky/crime_and_punishment/section:0000/paragraph:0010/sentence:0002
## 6 dostoevsky/crime_and_punishment/section:0000/paragraph:0010/sentence:0003
##   property
## 1     text
## 2     text
## 3     text
## 4     text
## 5     text
## 6     text
##
## 1    "CRIME AND PUNISHMENT  By Fyodor Dostoevsky CRIME AND PUNISHMENT     PART I    CHAPTER I  On an
## 2
## 3
## 4
## 5                                                                                                "H
## 6 "The landlady who provided him with garret, dinners, and attendance, lived on the floor below, and
```

```
##   numeric.value error unit tags       date                                  guid
## 1            NA    NA   NA      2021-12-09 d257a079-a91c-4246-91c5-4eb3a1cdb613
## 2            NA    NA   NA      2021-12-09 7da10fdc-8178-4314-ba4b-26fd864a8555
## 3            NA    NA   NA      2021-12-09 44dc351c-9156-4872-81ec-ff322d1a147c
## 4            NA    NA   NA      2021-12-09 f756acc2-42a8-4072-a432-53a56f72991c
## 5            NA    NA   NA      2021-12-09 54376d3b-1d82-4fe1-9cc2-ba4d3cc18f10
## 6            NA    NA   NA      2021-12-09 a179d24a-5065-42d0-b8a2-8a128aa5eac5
```

Looking at the ingested data, I created a word cloud.

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

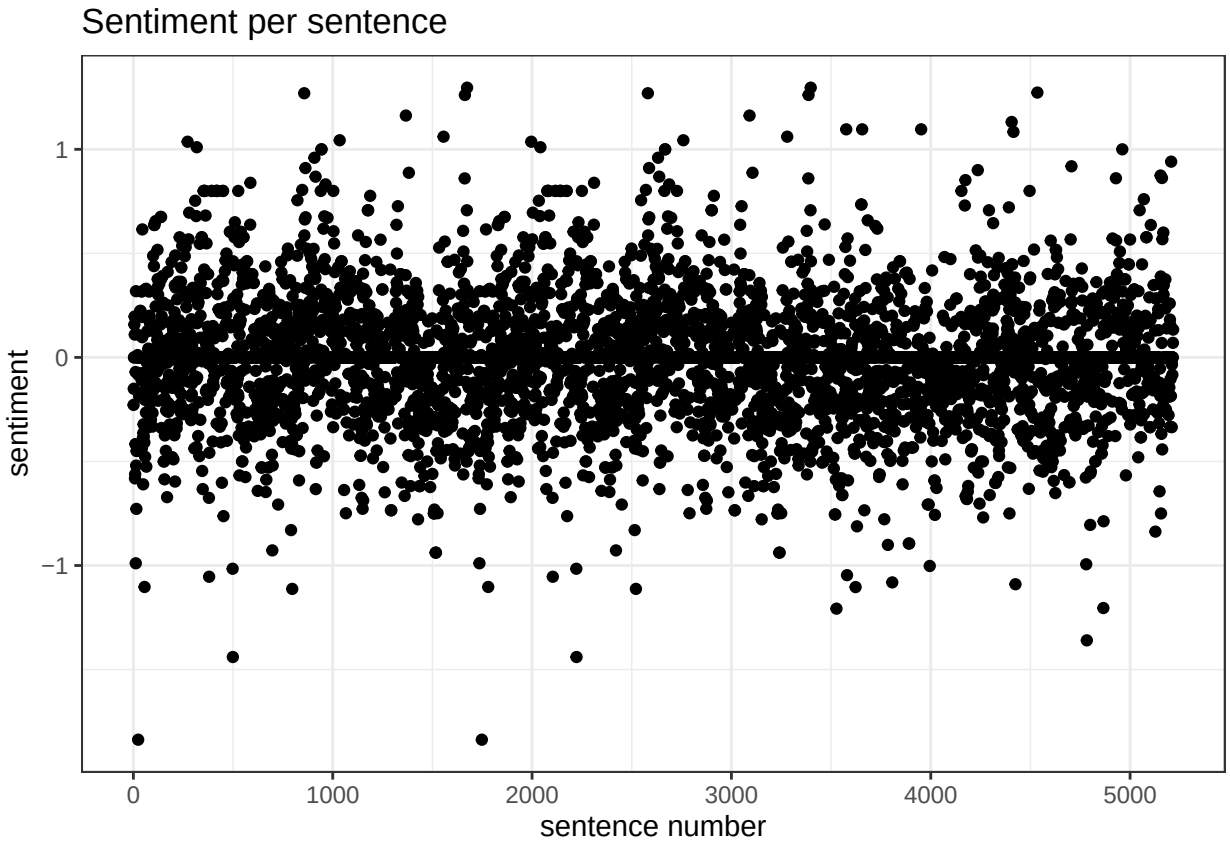
Now looking at the most negative words used, using Bing as the lexicon, we have:

```
## # A tibble: 6 x 2
##   word         n
##   <chr>    <int>
## 1 strange    127
## 2 afraid      90
## 3 drunk       76
## 4 nonsense    72
## 5 pale        71
## 6 lying       68
```

This is really interesting! I do remember in this book that "strange" is used a lot, especially because Raskolnikov likes to mumble "strange... strange" to himself. "Afraid," "pale," and "lying" all also make

sense because Raskolnikov committed murder and he is afraid of being caught, so these are all words that describe him being on edge around the detectives.

Now I can investigate sentiment polarity on a sentence level using truenumbers. I used -0.2 and 0.2 for cutoffs of negative and positive sentiments respectively, with "neutral" being the category between them. I plotted the sentiment of the sentences, but it is hard to see a general trend. I provide a table with the summary data for reference.




```
##    element_id  sentence_id     word_count       sentiment
##  Min.   :1    Min.   :   1   Min.   :  1.00   Min.   :-1.8371
##  1st Qu.:1    1st Qu.:1433   1st Qu.:  8.00   1st Qu.:-0.4750
##  Median :1    Median :2646   Median : 14.00   Median :-0.3386
##  Mean   :1    Mean   :2616   Mean   : 18.51   Mean   :-0.4005
##  3rd Qu.:1    3rd Qu.:3859   3rd Qu.: 25.00   3rd Qu.:-0.2652
##  Max.   :1    Max.   :5209   Max.   :108.00   Max.   :-0.2004
##  polarity_level
##  Length:980
##  Class :character
##  Mode  :character
##
##
##
```

```
##    element_id  sentence_id     word_count       sentiment
##  Min.   :1    Min.   :   1   Min.   :  1.00   Min.   :-1.8371
##  1st Qu.:1    1st Qu.:1433   1st Qu.:  8.00   1st Qu.:-0.4750
##  Median :1    Median :2646   Median : 14.00   Median :-0.3386
```

```
##  Mean   :1    Mean   :2616   Mean   : 18.51   Mean   :-0.4005
##  3rd Qu.:1    3rd Qu.:3859   3rd Qu.: 25.00   3rd Qu.:-0.2652
##  Max.   :1    Max.   :5209   Max.   :108.00   Max.   :-0.2004
##  polarity_level
##  Length:980
##  Class :character
##  Mode  :character
##
##
##
```

Analysis

The book is based on the idea of duality: crime and punishment. According to the foreward in the translated book, Dostoevsky apparently was aware of the artistic form of writing, and he purposefully made the book "symmetric" in its negative and positive connotations, and with the character development of the main character Raskolnikov. His name literally means "schism," and the book demonstrates the transition from one personality (logical) to another (emotional). The sentiment analysis I have done here on both a word and sentence level speak to what Dostoevsky claimed to do. I thought this was an interesting word sentiment analysis to perform on one of my favorite books.