Analyzing Text Data

Angel Salazar

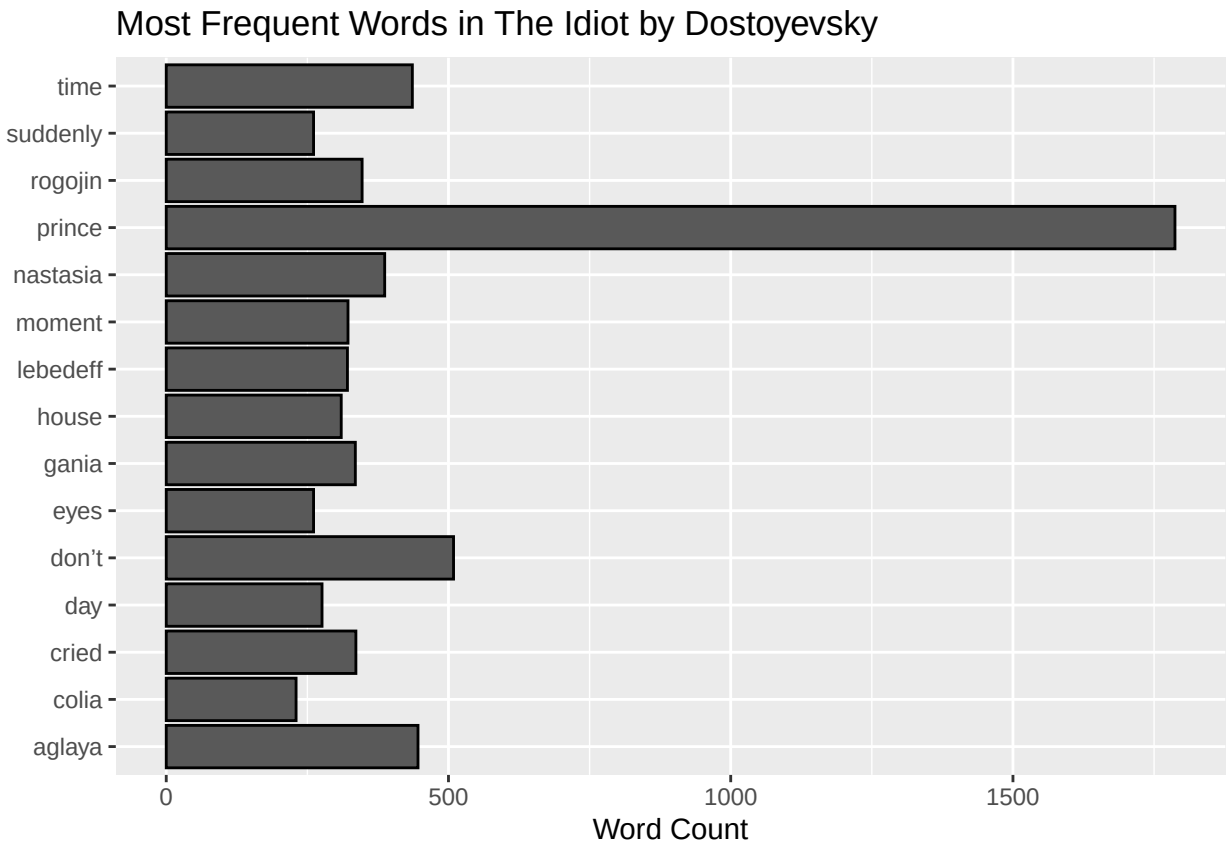
12/5/2021

```
## Searching for "The Idiot" by Dostoyevsky -- has id 2638
gutenberg_metadata %>%
  filter(title == "The Idiot" )
```

```
## # A tibble: 2 x 8
##   gutenberg_id title     author gutenberg_autho~ language gutenberg_books~ rights
##          <int> <chr>     <chr>             <int> <chr>    <chr>            <chr>
## 1         2638 The Idiot Dosto~              314 en       Best Books Ever~ Publi~
## 2        18881 The Idiot Bangs~              979 en       Humor            Publi~
## # ... with 1 more variable: has_text <lgl>
```

In this assignment, we would like to analyze words in The Idiot by Dostoyevsky. One of the first things we can do is create a column graph of the top 15 words and their frequencies.

```
## want to portray most frequent words used in the book
el_idiota_words_count %>%
  slice_max(order_by = n, n = 15) %>%
  ggplot(aes(x = word, y = n)) +
  geom_col(color = "black") +
  scale_x_reordered() +
  labs (title = "Most Frequent Words in The Idiot by Dostoyevsky",
        x = NULL,
        y = "Word Count") +
  coord_flip()
```



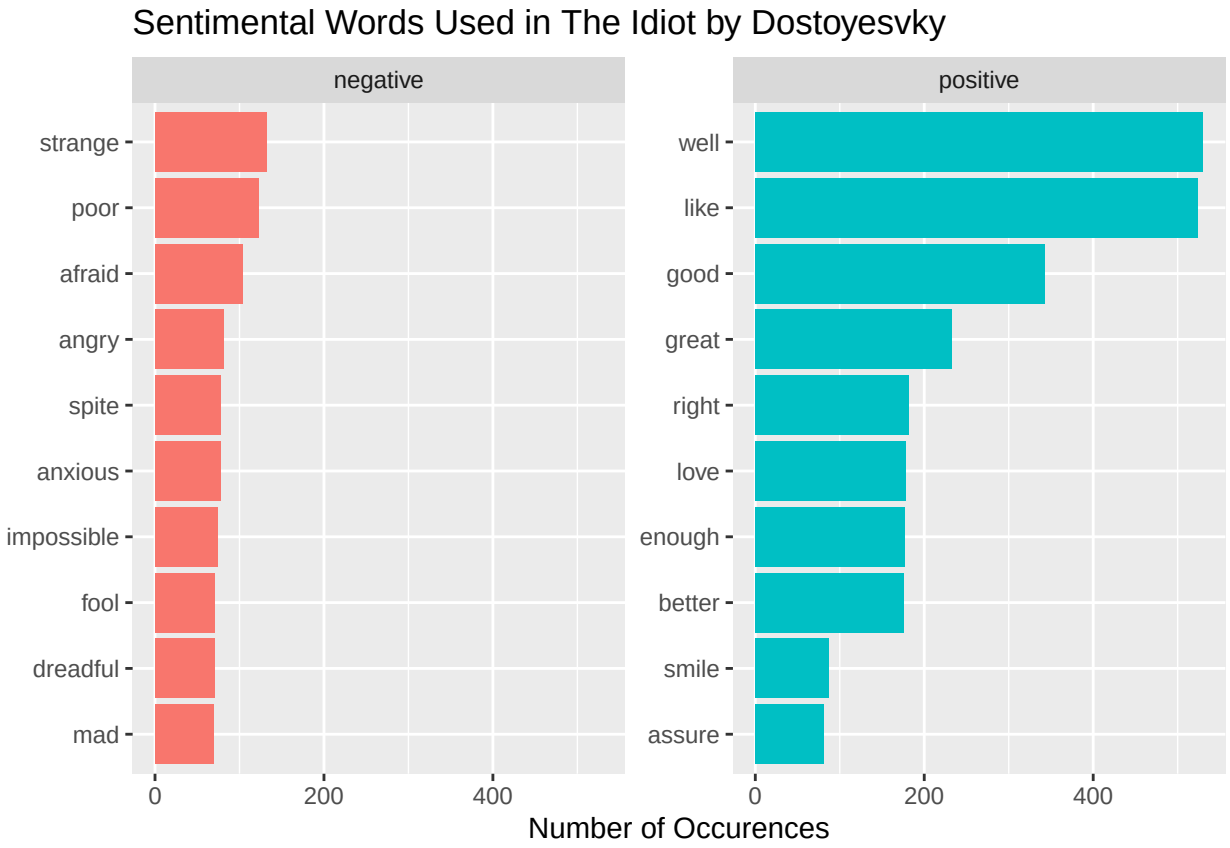

```
## generate df with sentiment derived from the Bing dictionary
(el_idiota_bing <- el_idiota_words %>%
   inner_join(get_sentiments("bing")))
```

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

## # A tibble: 16,017 x 3
##    gutenberg_id word          sentiment
##           <int> <chr>         <chr>
##  1         2638 idiot         negative
##  2         2638 great         positive
##  3         2638 difficulty    negative
##  4         2638 succeeded     positive
##  5         2638 breaking      negative
##  6         2638 impossible    negative
##  7         2638 best          positive
##  8         2638 insignificant negative
##  9         2638 weary         negative
## 10         2638 poorly        negative
## # ... with 16,007 more rows
```

Next, it would also be a good idea to get a sense of sentimental words being used in the novel. Thus, we will work to visualize negative and positive words utilized in the novel using the Bing Dictionary.

```
## visualize the most frequent positive/negative words in the entire book
## using the Bing dictionary
el_idiota_bing %>%
  group_by(sentiment) %>%
  count(word) %>%
  group_by(sentiment) %>%
  slice_max(order_by = n, n = 10) %>%
  mutate(word = reorder_within(word, n, sentiment)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  scale_x_reordered() +
  facet_wrap(facets = vars(sentiment), scales = "free_y") +
  labs(title = "Sentimental Words Used in The Idiot by Dostoyesvky",
       x = NULL,
       y = "Number of Occurences") +
  coord_flip()
```



```
## generate df with sentiment derived from the AFINN sentiment dictionary
(el_idiota_afinn <- el_idiota_words %>%
   inner_join(get_sentiments("afinn")))
```

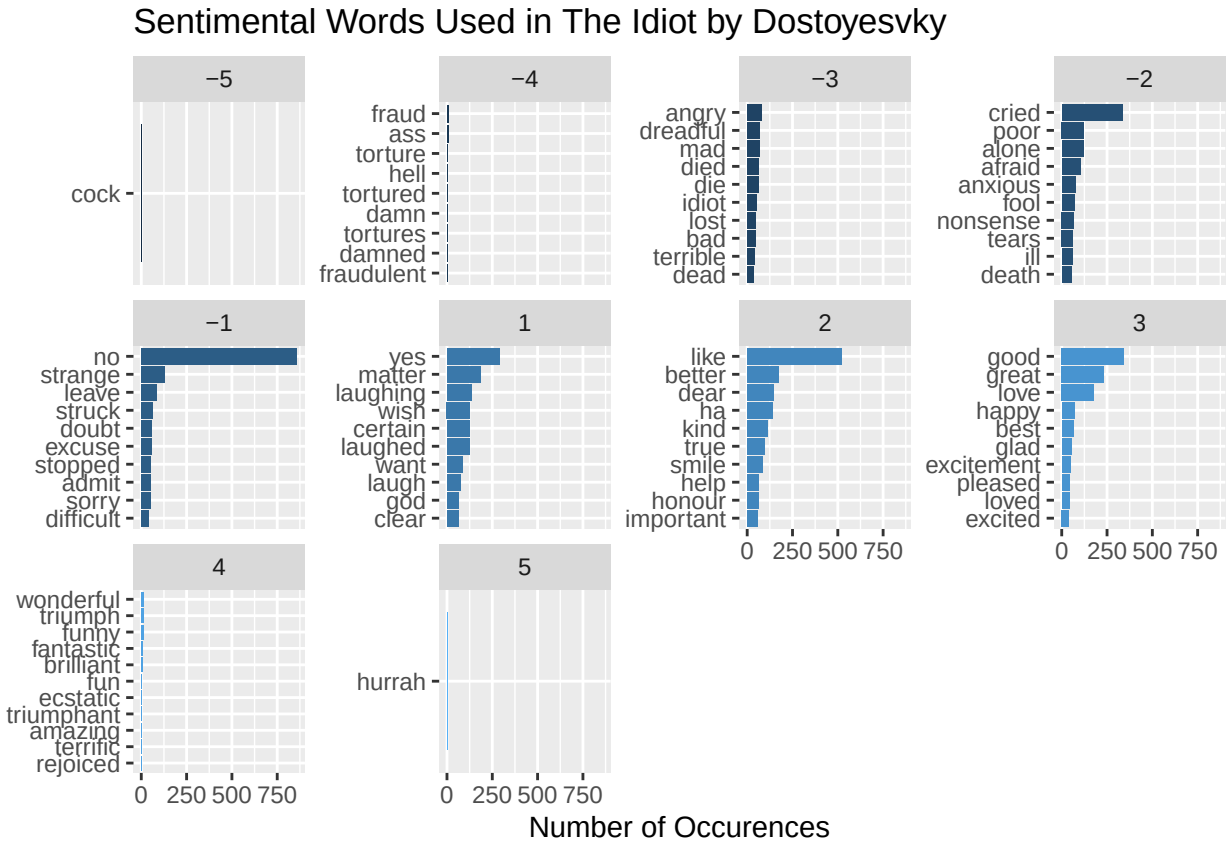
```
## Joining, by = "word"
```

```
## # A tibble: 15,417 x 3
##    gutenberg_id word          value
##           <int> <chr>         <dbl>
##  1         2638 idiot            -3
##  2         2638 great             3
##  3         2638 best              3
##  4         2638 insignificant    -2
##  5         2638 weary            -2
##  6         2638 remarkable        2
##  7         2638 anxious          -2
##  8         2638 remarkable        2
##  9         2638 strange          -1
## 10         2638 chance            2
## # ... with 15,407 more rows
```

Moving forward, let us now use a different text sentiment source, AFINN. Thus, now we are able to get a wider range of sentiment values related to words. This AFINN source analyze words within text and places values ranging from -5 to 5, thus allowing for a wider spread and nuance of words being used in the novel, The Idiot.

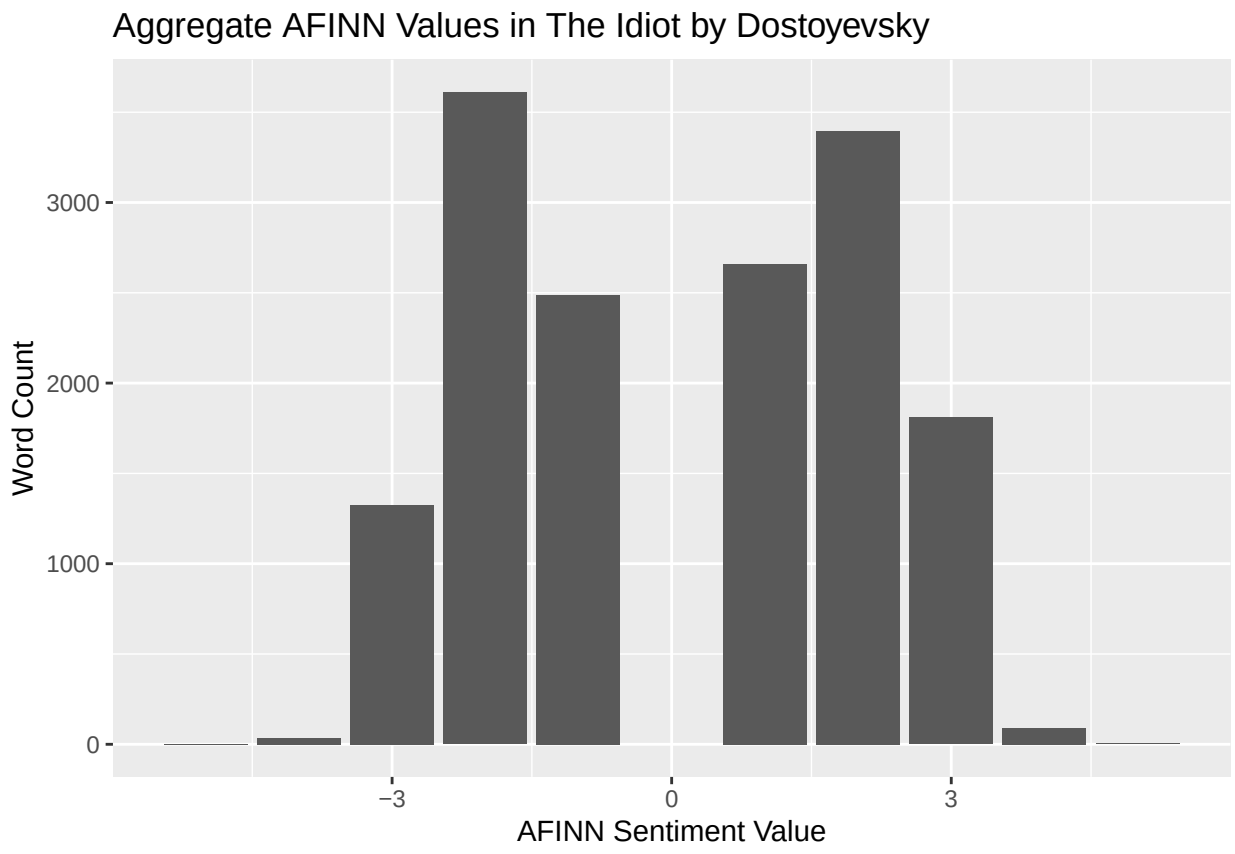
Let's now visualize words and their associated sentiment AFINN values, thus allowing us to further visualize word use in The Idiot.

```
## visualize the most frequent positive/negative words in the entire book
## using the afinn sentiment dictionary
el_idiota_afinn %>%
  group_by(value) %>%
  count(word) %>%
  group_by(value) %>%
  slice_max(order_by = n, n = 10) %>%
  mutate(word = reorder_within(word, n, value)) %>%
  ggplot(aes(word, n, fill = value)) +
  geom_col(show.legend = FALSE) +
  scale_x_reordered() +
  facet_wrap(facets = vars(value), scales = "free_y") +
  labs(title = "Sentimental Words Used in The Idiot by Dostoyesvky",
       x = NULL,
       y = "Number of Occurences") +
  coord_flip()
```



Being able to find nuances within the text is important as it allows us to really go in depth into word usage and its associated sentiment. But what if we would like to just get a general idea of word sentiment usage overall? We can aim to visualize this below in a bar graph. As can be seen, there seems to exist a pretty equal quantity of negative and positive word usage in The Idiot. It can be estimated that overall there is more negative word usage however, via the graph below.

```
## want to give a general visualization of sentiment scores overall
el_idiota_afinn %>%
  ggplot(aes(x = value)) +
  geom_bar() +
  labs(title = "Aggregate AFINN Values in The Idiot by Dostoyevsky",
       x = "AFINN Sentiment Value",
       y = "Word Count")
```



Lastly, let us create a word cloud of the top 50 words used in the Novel.

```
## visualize which words in the AFINN sentiment dictionary appear most
set.seed(420)

el_idiota_afinn %>%
  count(word) %>%
  slice_max(order_by = n, n = 50) %>%
  mutate(angle = 90 * sample(c(0, 1), n(), replace = TRUE, prob = c(70, 30))) %>%
  ggplot(aes(label = word, size = n, angle = angle)) +
  geom_text_wordcloud(rm_outside = TRUE) +
  scale_size_area(max_size = 15) +
  ggtitle("Most Frequent Tokens in The Idiot") +
  theme_minimal()
```

Most Frequent Tokens in The Idiot