text mining

Tong Sun

12/1/2021

Firstly, I would like to tidy the data.Annotate a 'linenumber' quantity to keep track of lines in the original format and a 'chapter'(with a regex) to find where all the chapters are.

```
## Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest
```

```
## Using mirror http://aleph.gutenberg.org
```

```
## # A tibble: 188,399 x 4
##    gutenberg_id linenumber chapter word
##           <int>      <int>   <int> <chr>
##  1         1260          1       0 jane
##  2         1260          1       0 eyre
##  3         1260          2       0 an
##  4         1260          2       0 autobiography
##  5         1260          4       0 by
##  6         1260          4       0 charlotte
##  7         1260          4       0 brontë
##  8         1260          6       0 _illustrated
##  9         1260          6       0 by
## 10         1260          6       0 f
## # ... with 188,389 more rows
```

Task three

Truenumbers (TN) is a system for building data resources that are hosted on a server and accessed through clients that interact with with server through the Truenumbers API. The system uses natural language descriptions of data items and includes a tagging function for augmenting data items, defining subsets, and tracking process metadata. Following I would like to use TN to do analysis on Eyre book.

In this part, I used the tnum package to explore and tag text from the book Jane Eyre. I explored the frequency of words and characters throughout each section of the book, and created visualizations to show these frequencies. The process is described in more detail below. ## Download packages and the book

Assign the return from 'tnum.query' function to a variable so that we can examine the list items in the environment. And then convert the TN list to a data frame (use the 'tnum.objectstoDF()' function) so that each TN is a row in the data frame.

```
## Returned 1 thru 500 of 1637 results
```

```
## Returned 1 thru 10 of 1637 results
```

```
## Returned 1 thru 6 of 6 results
```

```
## Returned 1 thru 2 of 2 results
```

```
## Returned 1 thru 6 of 6 results
```

Use tnum in text analysis

In this part, I want to use TNs for text analysis. As I did before, loading the libraries I need, authorizing the server if needed and setting the number space to "test2"

```
## Returned 1 thru 1637 of 1637 results
```

```
## Returned 4 thru 21 of 1637 results
```

```
## Returned 1 thru 24 of 24 results
## Returned 1 thru 24 of 24 results
```

```
## Joining, by = "subject"
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
## Returned 1 thru 10 of 74 results
```

```
## Returned 1 thru 3 of 4887 results
```

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

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

```
## [1] "\""Jane, you are under a mistake: what is the matter with you?\""
## [2] "\"Why do you tremble so violently?\""
## [3] "\"Why do you tremble so violently?? Would you like to drink some water?\""
```

```
## Returned 1 thru 9 of 9 results
```

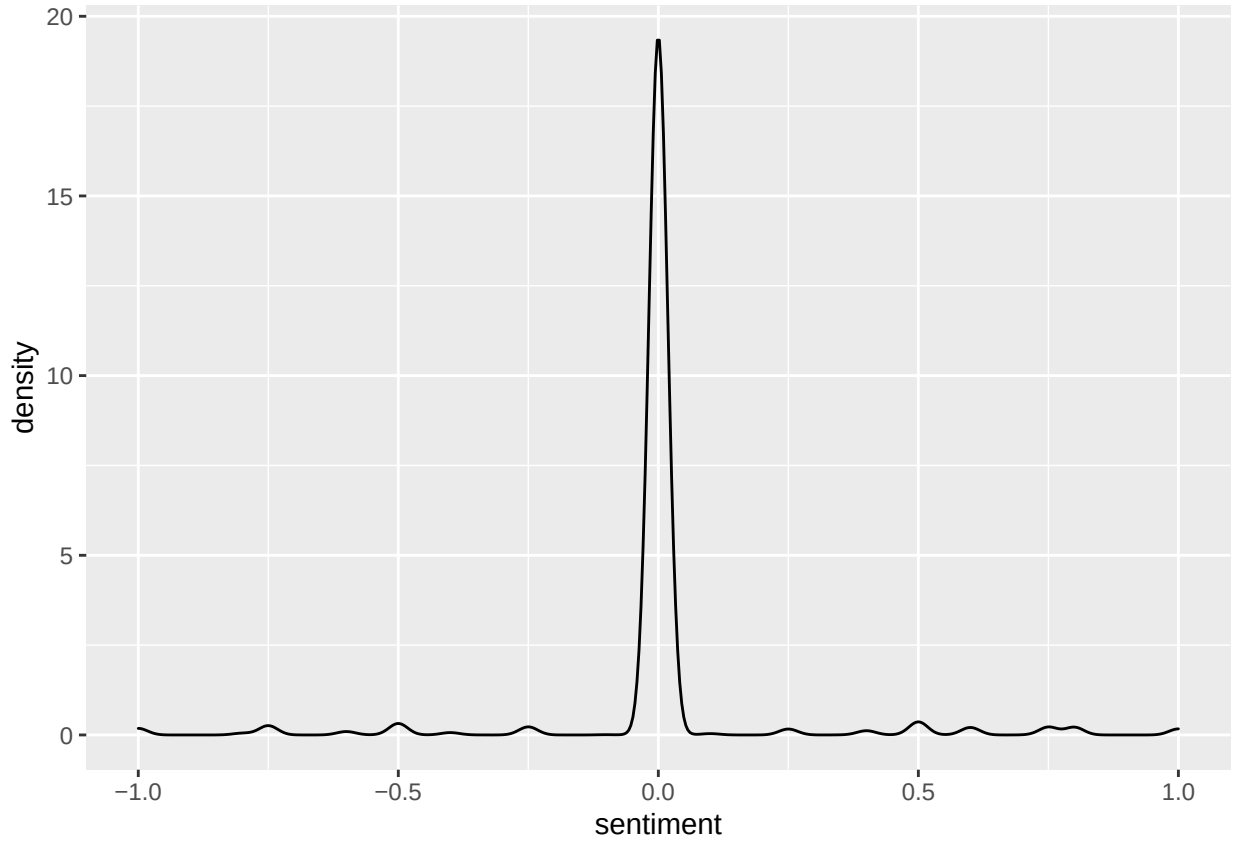
```
## [1] NA
## [2] NA
## [3] NA
## [4] NA
## [5] NA
## [6] NA
## [7] "\""Jane, you are under a mistake: what is the matter with you?\""
## [8] "\"Why do you tremble so violently?\""
## [9] "\"Why do you tremble so violently?? Would you like to drink some water?\""
```

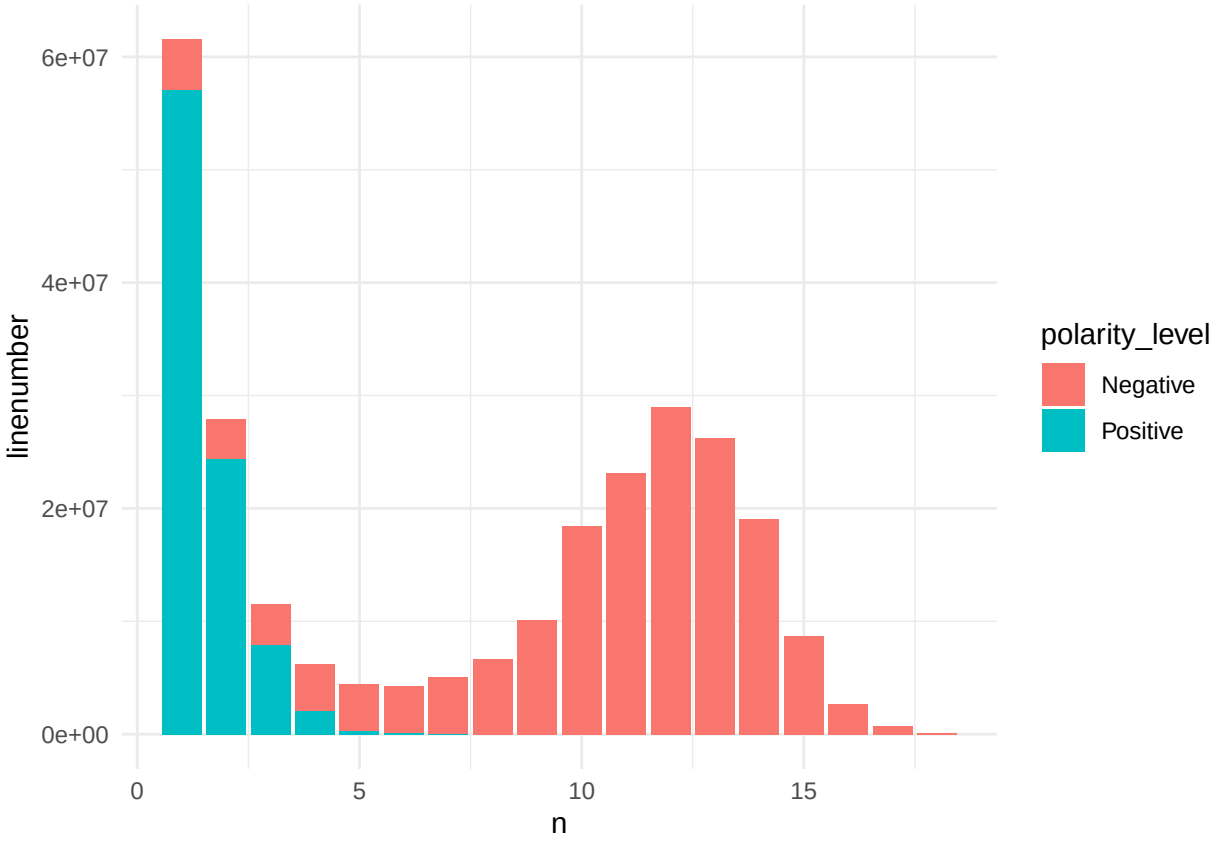
```
## [1] 9
```

```
## Returned 1 thru 9 of 9 results
```

Sentimentr

Sentimentr is designed to quickly calculate text polarity sentiment in the English language at the sentence level and optionally aggregate by rows or grouping variable(s).





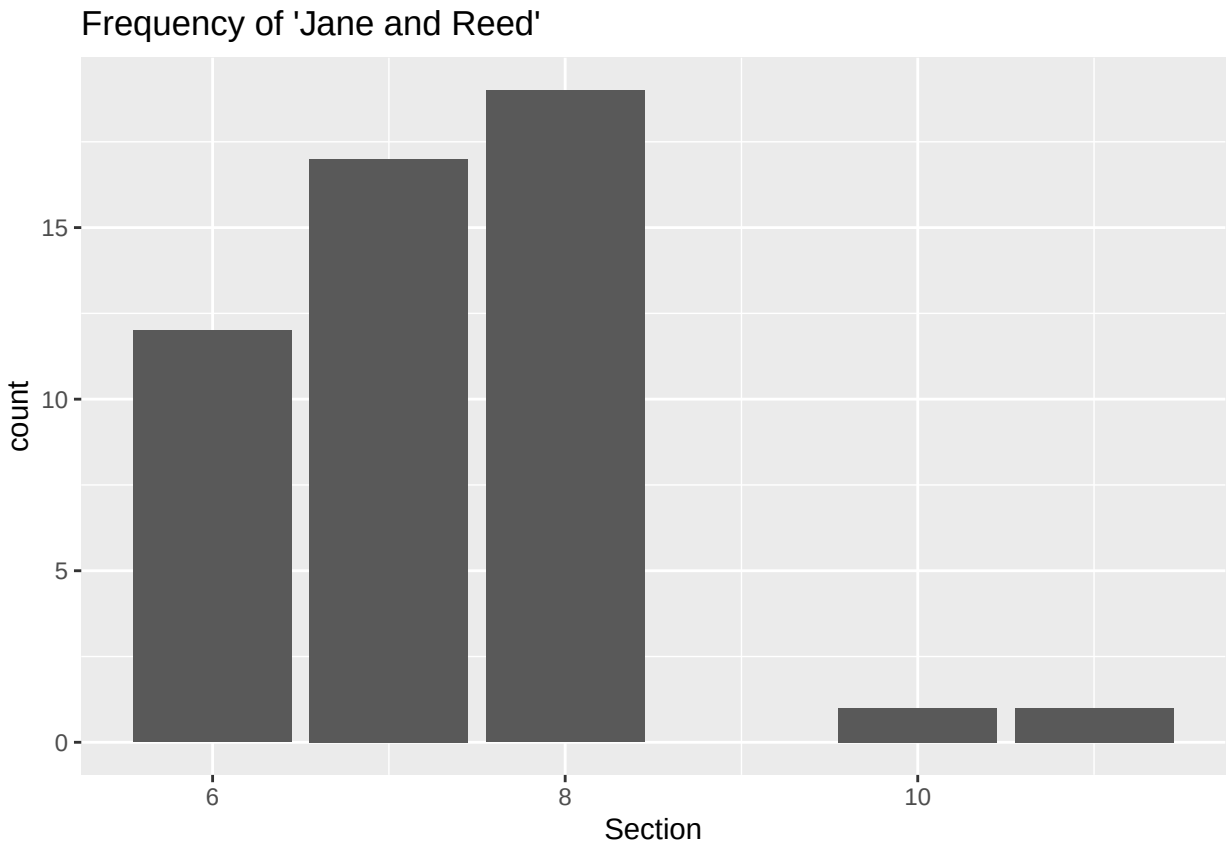
From the density plot, we can see most of the sentiments are zero. For the plot with 'polarity_level', each linenumber has different number of positive and negative sentiments.

Extra attempts: tag

From the content, I would like to analysis the relationship between Jane and her uncle, Reed. Firstly I create a tag to 'Jane|Reed' in order to find where they comes out together in this book.

```
## Returned 1 thru 50 of 141 results
```

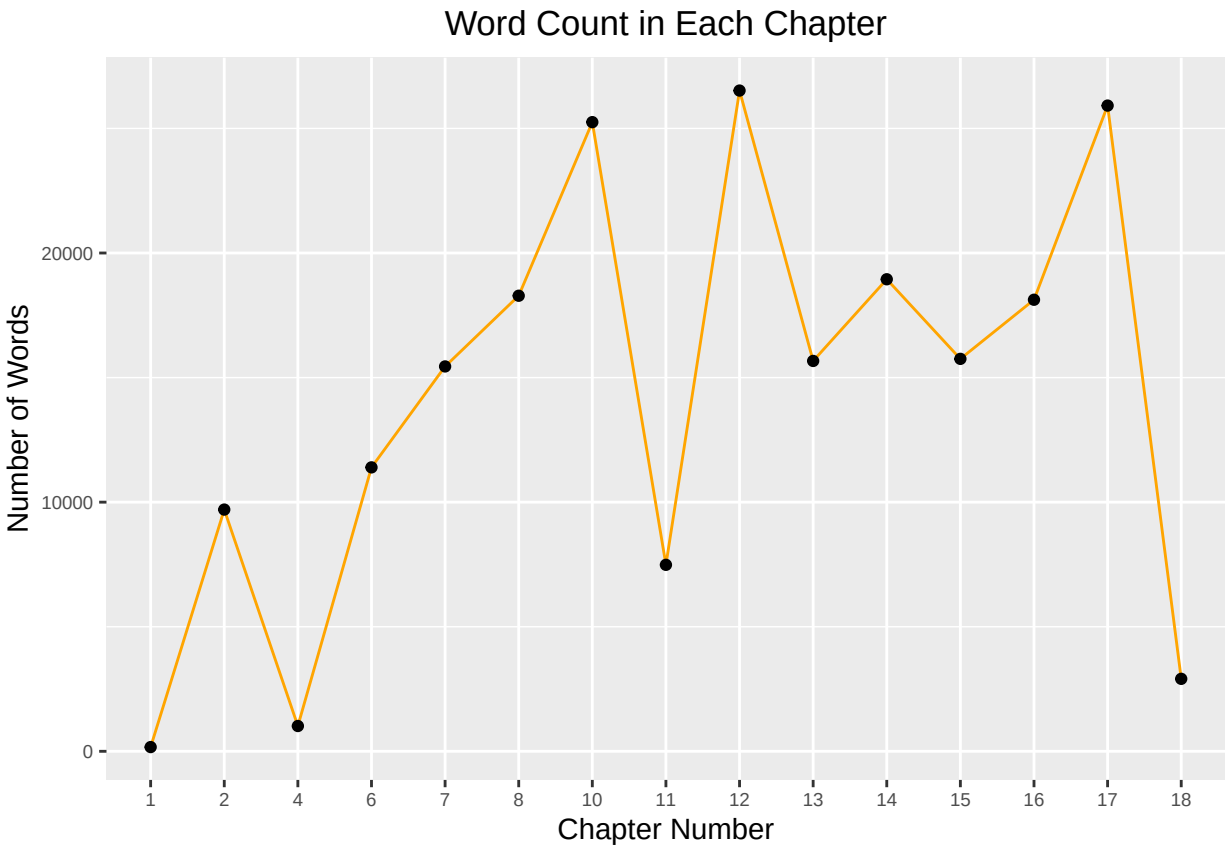
```
## list(modifiedCount = 141, tagged = 141, removed = 0)
```

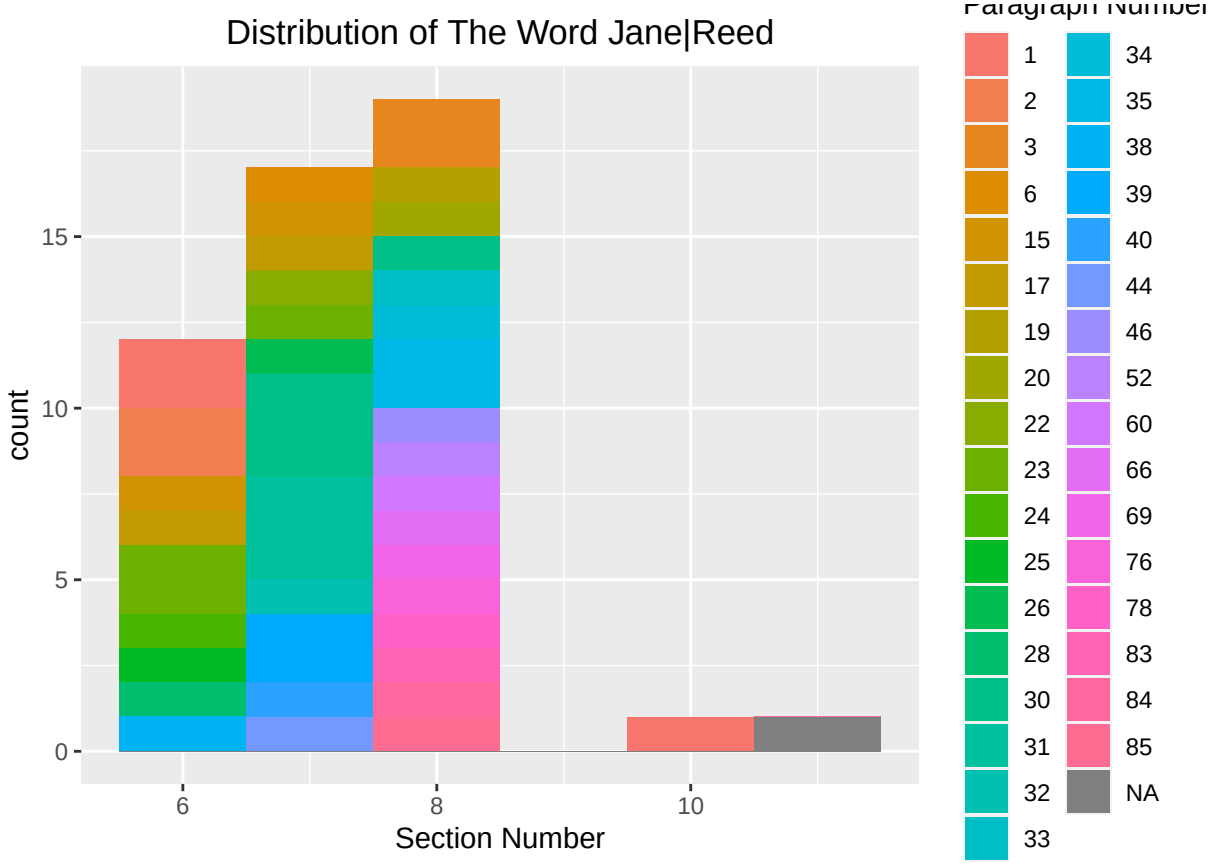


From the plot above, it shows that 'Jane and Reed' comes out together in several sections. The most important part is in the earlier part of the book, which makes sense. This fits the story that the book tells – her uncle, Reed, raises Jane when she was a child.

Show virsually

```
## Returned 1 thru 1500 of 1613 results
```


Word Count in Each Chapter
Number of Words
20000
10000
0
1
2
4
6
7
8
10
11
12
13
14
15
16
17
18
Chapter Number



This part I would like to show the relationship between Jane and Reed, but I cannot use the function such as 'tnum.plotGraph()' in the file you showed us in class. There is an error saying these functions do not exist. What I can do is drawing a plot showing the distribution of the Word 'Jane|Reed' occurances. I will work on the part that does not go well in the future.