text mining

Tong Sun

11/29/2021

Task one and two

Download and Tidy

Jane Eyre is divided into 38 chapters. It was originally published in three volumes in the 19th century, comprising chapters 1 to 15, 16 to 27, and 28 to 38. The novel is a first-person narrative from the perspective of the title character. It has five distinct stages: Jane's childhood at Gateshead Hall, where she is emotionally and physically abused by her aunt and cousins; her education at Lowood School, where she gains friends and role models but suffers privations and oppression; her time as governess at Thornfield Hall, where she falls in love with her mysterious employer, Edward Fairfax Rochester; her time in the Moor House, during which her earnest but cold clergyman cousin, St. John Rivers, proposes to her; and ultimately her reunion with, and marriage to, her beloved Rochester. Throughout these sections, it provides perspectives on a number of important social issues and ideas, many of which are critical of the status quo.

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

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

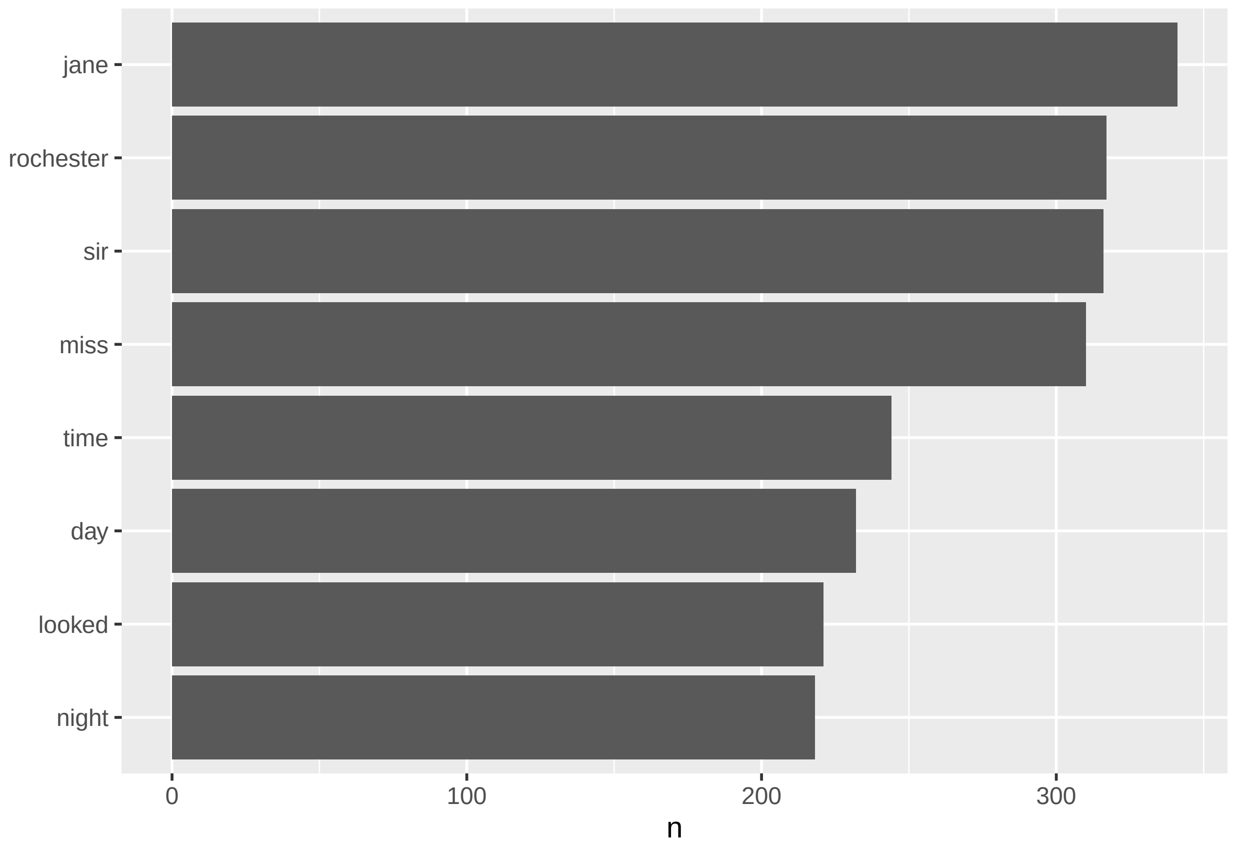
```
## # A tibble: 188,399 x 4
##    gutenberg_id linenumber chapter word
##           <int>      <int>   <int> <chr>
##  1         1260          1       0 jane
##  2         1260          1       0 eyre
##  3         1260          2       0 an
##  4         1260          2       0 autobiography
##  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
```

```
## [1] 39
```

```
## # A tibble: 39 x 2
##    chapter     n
##      <int> <int>
##  1       0   959
##  2       1  1954
##  3       2  2771
##  4       3  3230
##  5       4  5866
##  6       5  5046
##  7       6  2923
##  8       7  3604
##  9       8  3015
## 10       9  3276
## # ... with 29 more rows
```

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

```
## # A tibble: 12,309 x 2
##    word           n
##    <chr>     <int>
##  1 jane         341
##  2 rochester    317
##  3 sir          316
##  4 miss         310
##  5 time         244
##  6 day          232
##  7 looked       221
##  8 night        218
##  9 eyes         187
## 10 john         184
## # ... with 12,299 more rows
```



Sentiment analysis with tidy data

In the previous part, I explored what changed the original book into the tidy text format and showed how this format can be used to approach questions about word frequency. Next I would like to address the topic of sentiment analysis. When human readers approach a text, we use our understanding of the emotional intent of words to infer whether a section of text is positive or negative. So here I use the tools of text mining to appraoch the emotional content of text.

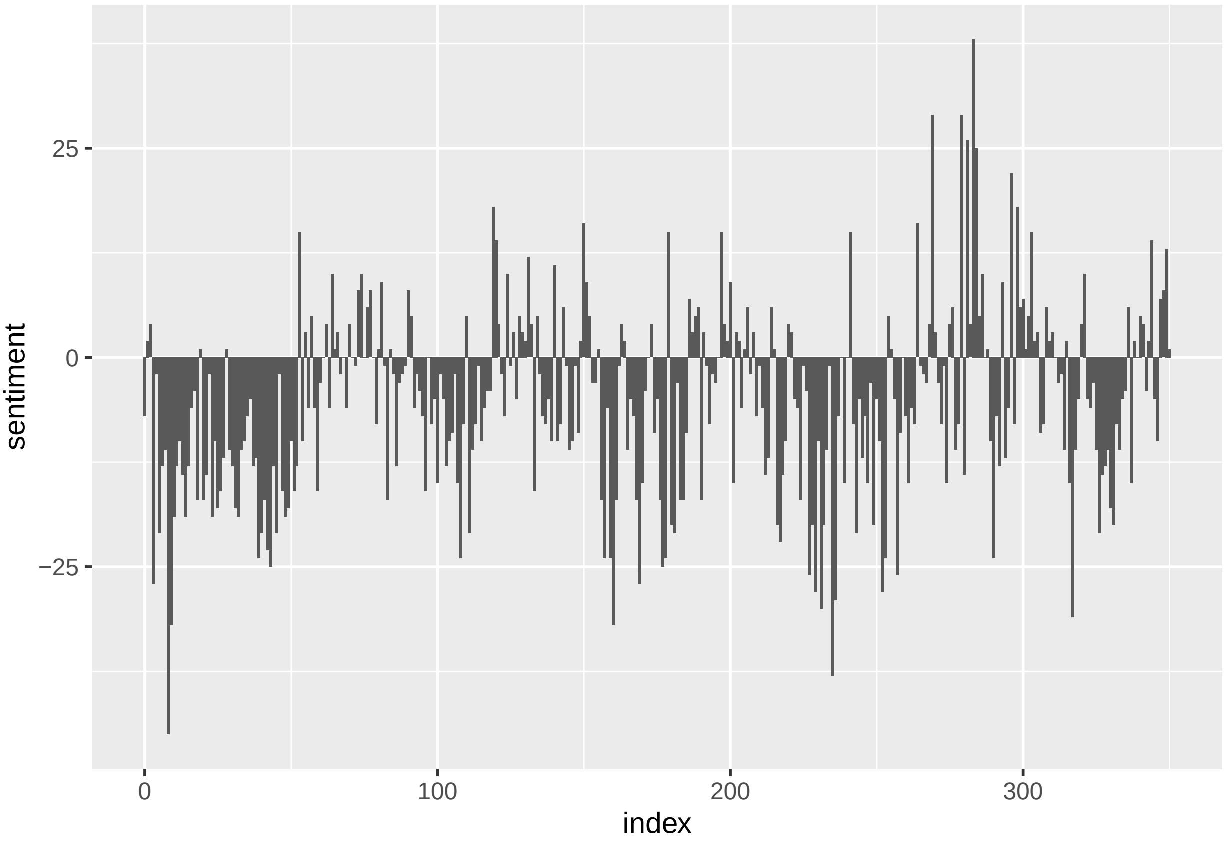
```
## Joining, by = "word"
```

```
## # A tibble: 404 x 2
##    word        n
##    <chr>   <int>
##  1 love      151
##  2 found     126
##  3 god        96
##  4 child      88
##  5 hope       75
##  6 feeling    67
##  7 happy      54
##  8 marry      49
##  9 smile      47
## 10 mother     45
## # ... with 394 more rows
```

I found that mostly positive, happy words about love, hope and happy here.

In addition, I want to examine how sentiment changes throughout each novel. First, I find a sentiment score for each word using the Bing lexicon and 'inner_join()'. Next I count up how many positive and negative words there are in defined sections of the book and define an "index" here to keep track of where we are in the narrative; this index (using integer division) counts up sections of 60 lines of text. In addition, I use 'pivot_wider()' so that we can have negative and positive sentiment in separate columns and finally calculate a net sentiment (positive - negative).

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

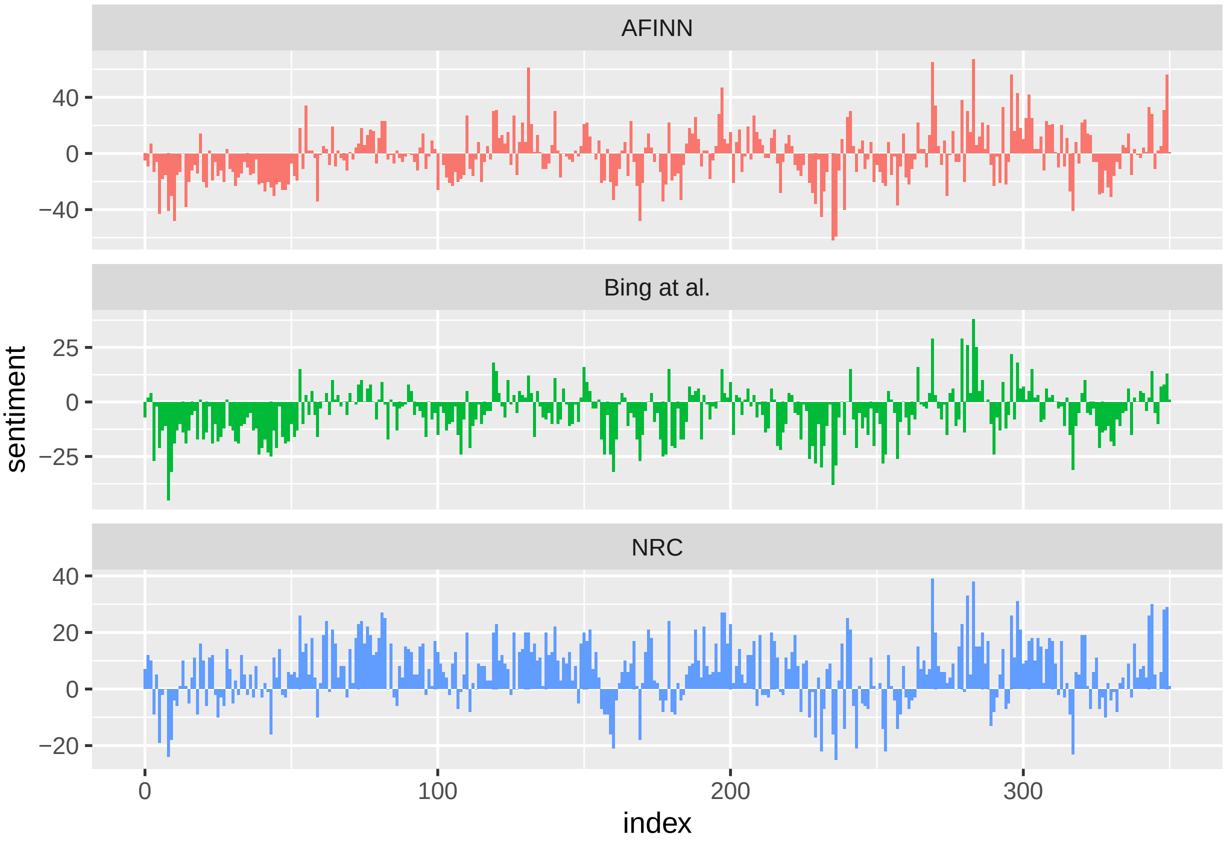


We can see how the plot of each novel changes toward more positive or negative sentiment over the trajectory of the story. From the plot above, I find that there are more negative sentiments during the early index and more positive sentiments during the late index. I think this fits the story. During Eyre's childhood, she was abuted by her aunt,Sarah Reed, she did not have a happy childhood in her early years. Therefore the sentiment is negative. But when she met Rochester as written in the late plots, she was married with him and had a sweet life. So there are more positive ones here.

Compare the three sentiment dictionaries

- AFINN: the AFINN lexicon measures sentiment with a numeric score between -5 and 5;
- Bing: the Bing lexicon categorizes words in a binary fashion into positive and negative categories;
- nrc: the nrc lexicon categorizes words in a binary fashion ("yes"/"no") into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

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



The three different lexicons for calculating sentiment give results that are different in an absolute sense but have similar relative trajectories through the novel. We can see similar dips and peaks in sentiment at about the same places in the novel, but the absolute values are significantly different. The AFINN lexicon gives the largest absolute values, with high positive values. The Bing lexicon has lower absolute values. The NRC lexicon are shifted higher realtive to the other two, labeling the text more positively.

```
## # A tibble: 2 x 2
##   sentiment     n
##   <chr>     <int>
## 1 negative   3318
## 2 positive   2308
```

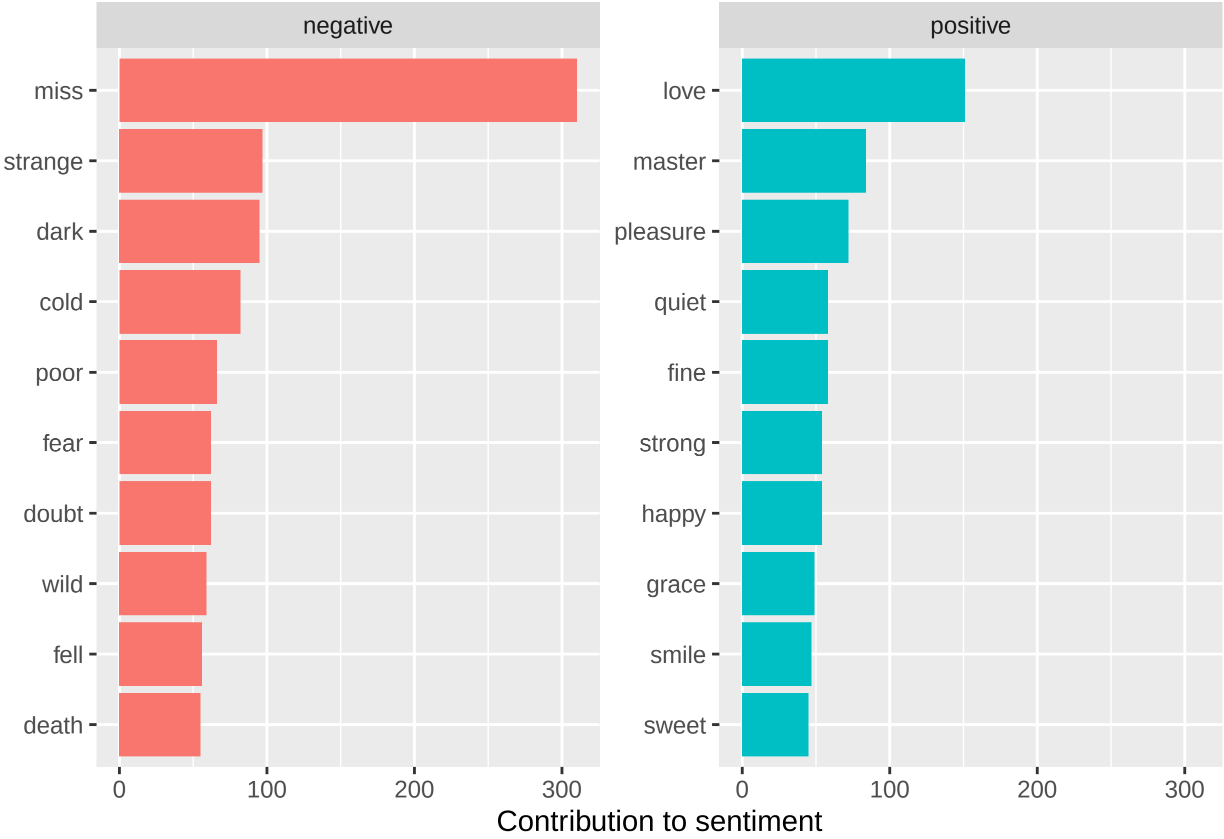
```
## # A tibble: 2 x 2
##   sentiment     n
##   <chr>     <int>
## 1 negative   4781
## 2 positive   2005
```

Both lexicons have more negative than positive words, but the ratio of negative to positive words is higher in the Bing lexicon than the NRC lexicon. This will contribute to the effect we see in the plot above, as will any systematic difference in word matches, e.g. if the negative words in the NRC lexicon do not match the words that Jane Austen uses very well. Whatever the source of these differences, we see similar relative trajectories across the narrative arc, with similar changes in slope, but marked differences in absolute sentiment from lexicon to lexicon. This is all important context to keep in mind when choosing a sentiment lexicon for analysis.

Most common positive and negative words – One advantage of having the data frame with both sentiment and word is that we can analyze word counts that contribute to each sentiment

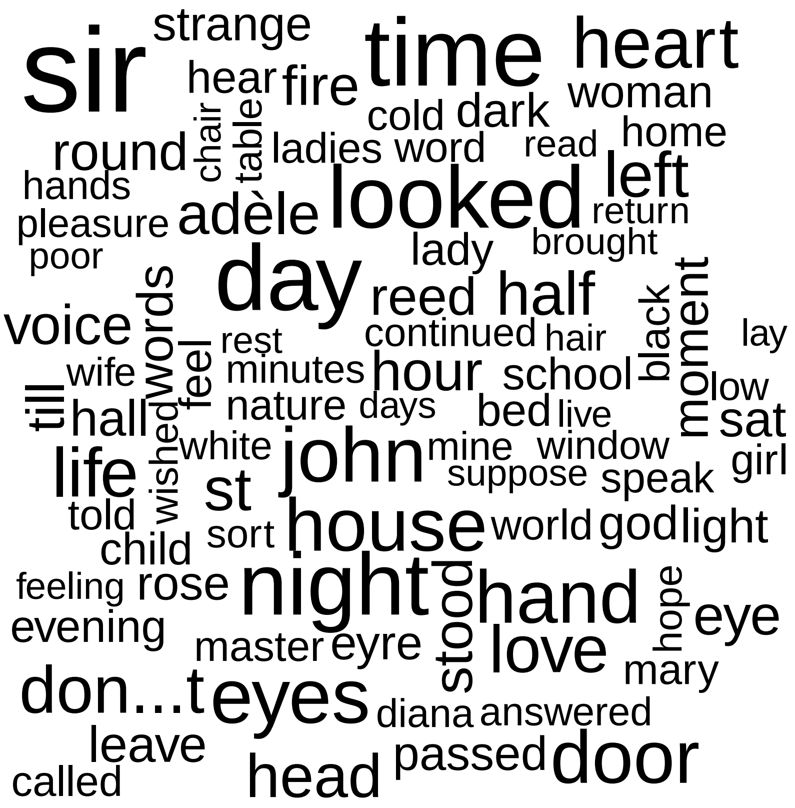
```
## Joining, by = "word"
```

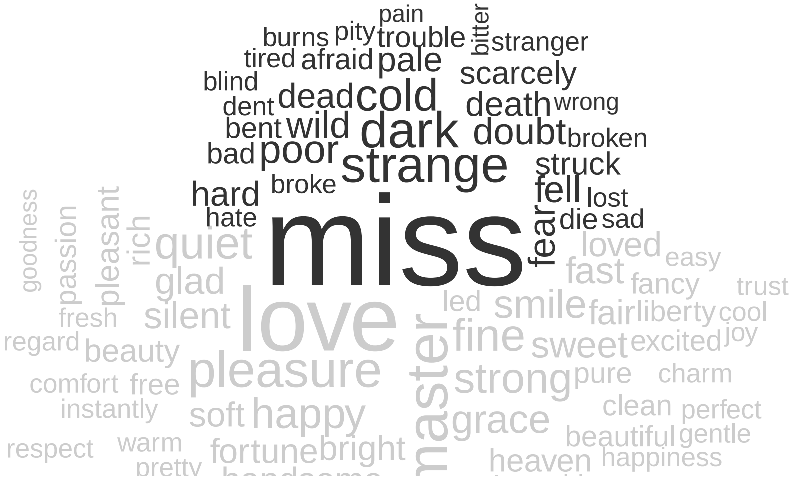
```
## # A tibble: 2,326 x 3
##    word     sentiment     n
##    <chr>    <chr>     <int>
##  1 miss     negative    310
##  2 love     positive    151
##  3 strange  negative     97
##  4 dark     negative     95
##  5 master   positive     84
##  6 cold     negative     82
##  7 pleasure positive     72
##  8 poor     negative     66
##  9 doubt    negative     62
## 10 fear     negative     62
## # ... with 2,316 more rows
```



The plot above shows words that contribute to positive and negative sentiment in Eyre.

Wordclouds


sir
strange
time
heart
hear
fire
woman
cold
dark
home
chair
table
round
ladies
word
read
hands
looked
left
pleasure
adèle
return
poor
lady
brought
day
words
reed
half
black
moment
voice
feel
rest
continued
hair
lay
wife
minutes
hour
school
till
hall
nature
days
bed
live
low
sat
wished
white
john
mine
window
life
suppose
speak
girl
st
told
house
world
god
light
sort
child
feeling
rose
night
stood
hand
hope
eye
evening
master
eyre
love
mary
don...t
eyes
diana
answered
leave
passed
door
called
head


pain
bitter
burns
pity
trouble
stranger
tired
afraid
pale
scarcely
blind
dead
cold
death
dent
wrong
bent
wild
dark
doubt
broken
bad
poor
strange
struck
broke
hard
fell
lost
hate
die
sad
fear
goodness
passion
pleasant
rich
quiet
miss
loved
easy
glad
fast
fancy
trust
led
smile
fair
liberty
cool
fresh
silent
love
master
regard
fine
sweet
excited
joy
beauty
comfort
free
pleasure
strong
pure
charm
instantly
soft
happy
grace
clean
perfect
respect
warm
fortune
bright
beautiful
gentle
pretty
heaven
happiness

The size of a word's text is in proportion to its frequency within its sentiment. We can use this visualization to see the most important positive and negative words. Here the most important positive word is 'positive' and the most important negative word is 'miss'.