Text Analysis of Wuthering Heights (Task1&2)

Jessie Xu

2021/12/8

Introduction

Wuthering Heights is an 1847 novel by Emily Bronte, initially published under the pseudonym Ellis Bell. It concerns two families of the landed gentry living on the West Yorkshire moors, the Earnshaws and the Lintons, and their turbulent relationships with Earnshaw's adopted son, Heathcliff. The novel was influenced by Romanticism and Gothic fiction.

Tidy Text

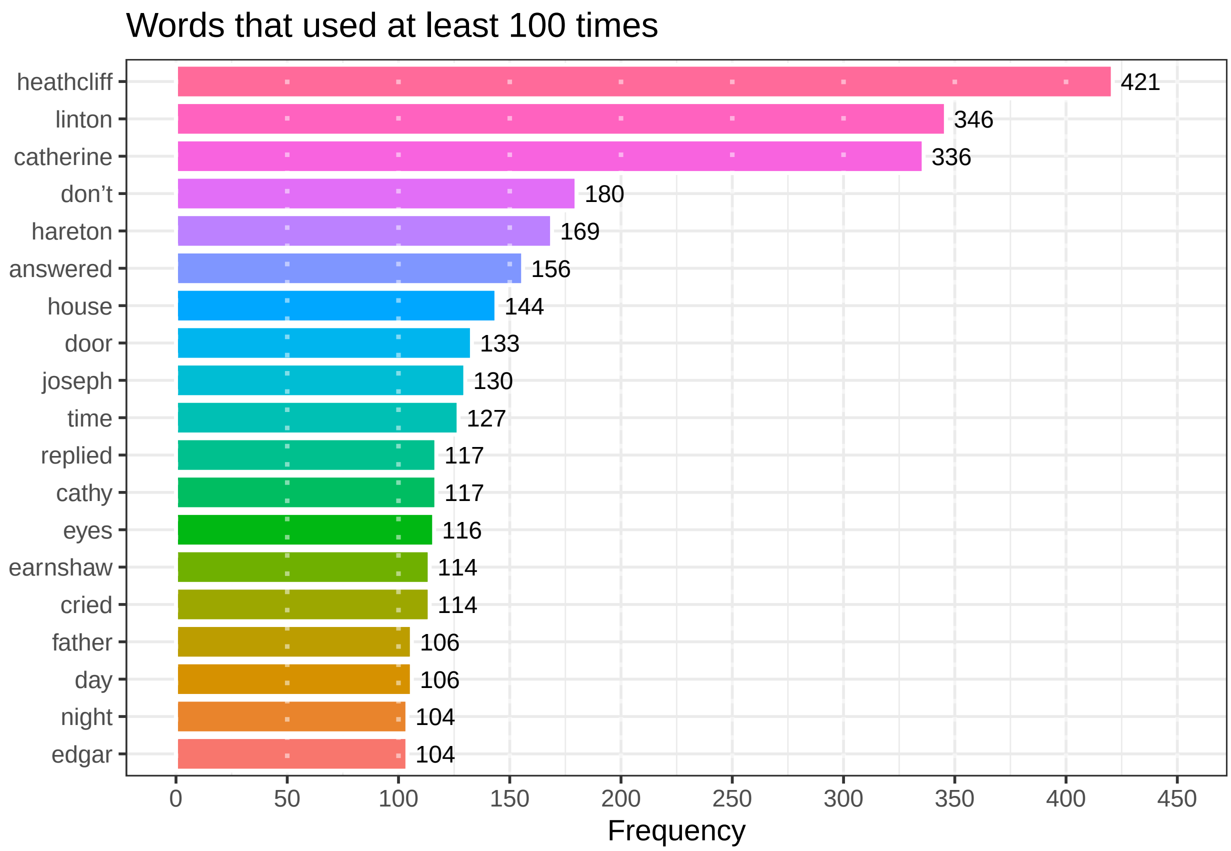
word	**n**
heathcliff	421
linton	346
catherine	336
i'm	192
i'll	189
master	185
don't	180
hareton	169
answered	156
till	151

Firstly, I try to figure out after excluding the stop words what are the most frequent words in Wuthering Heights (show as above), but there is an anomaly in this basic analysis. The words "i'm", "i'll", "till", "he's", "i've", "it's", and "you'll" have no practical meaning, and I am not interested in analyzing them. Also in the later sentiment analysis part, the word "miss" is coded as negative but it is used as a title for young, unmarried women in the book. And the word "master" is coded as positive but it's used as a title for the host of Wuthering Heights.

So, based on the idea of stop words, I easily add the words mentioned above to a custom stop-words list using bind_rows(), and exclude all of them when I do something analysis in word-level So the most frequent words with actual meaning in Wuthering Heights is:

word	**n**
heathcliff	421
linton	346
catherine	336
don't	180
hareton	169
answered	156
house	144
door	133
joseph	130
time	127

Then I visualize the frequency of the most commonly used words.



Sentiment Analysis: Words level

There is a contemporary review of Wuthering Heights:

The American Whig Review wrote:Respecting a book so original as this, and written with so much power of imagination, it is natural that there should be many opinions. Indeed, its power is so predominant that it is not easy after a hasty reading to analyze one's impressions so as to speak of its merits and demerits with confidence. We have been taken and carried through a new region, a melancholy waste, with here and there patches of beauty; have been brought in contact with fierce passions, with extremes of love and hate, and with sorrow that none but those who have suffered can understand.

(https://en.wikipedia.org/wiki/Wuthering_Heights#Contemporary_reviews)

As you can see, Wuthering Heights contains extreme emotions and is a story of love, revenge, and forgiveness. So I assume that the tone of this book is negative. Then I use 4 sentiment lexicons to analyze the sentiment of this book at a sentence level. I choose the fourth lexicon by checking "?get_sentiments" in R Documentation. As the **Arguments** of **get_sentiments** said: "lexicon: The sentiment lexicon to retrieve; either"afinn", "bing", "nrc", or "loughran"". So I used"loughran" as my additional lexicon.

Sentiment Lexicon	Definition
AFINN	Assigns words with a score between -5 and 5
bing	Categorizes words as positive or negative
nrc	Uses binary yes/no score in categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust
loughran	This lexicon labels words with six possible sentiments important in financial contexts: "negative", "positive", "litigious", "uncertainty", "constraining", or "superfluous"

1. an example using "nrc" lexicon

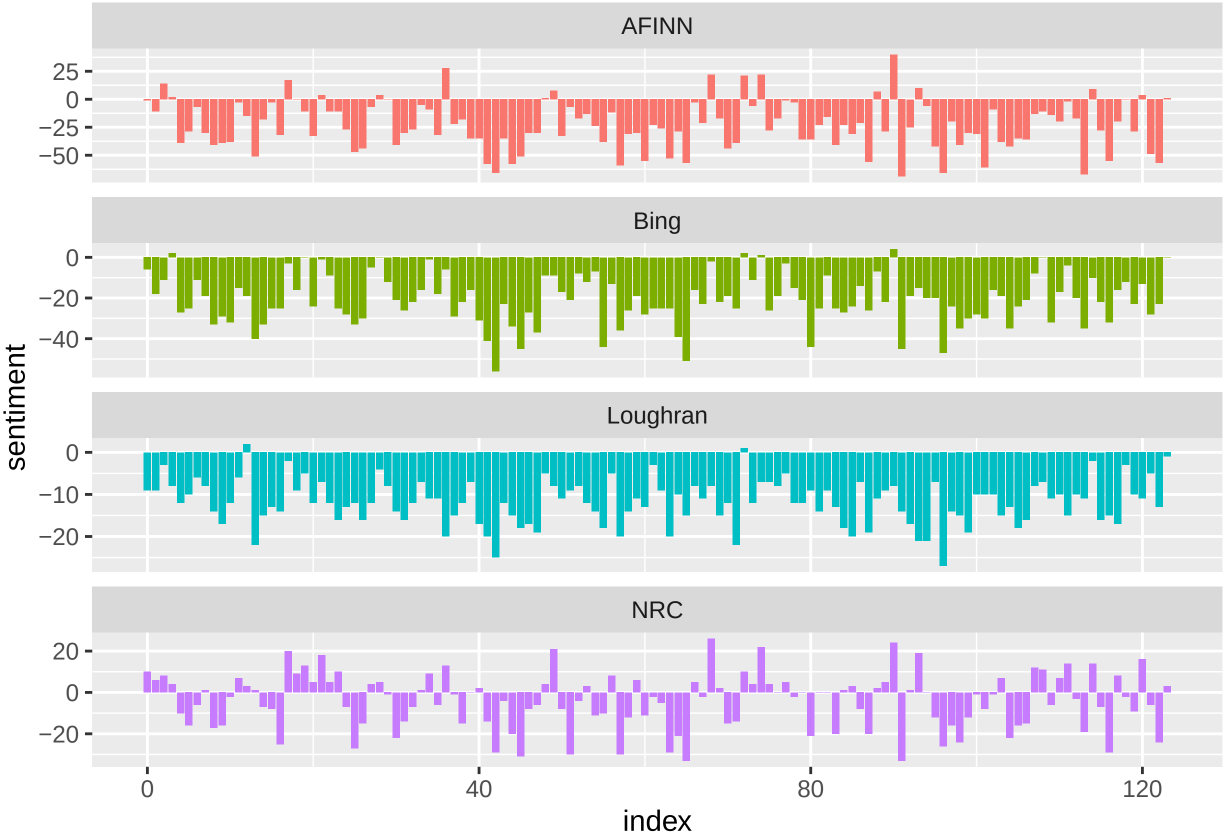
I select out the most frequent words filtered by sentiment "joy" from "nrc" lexicon

word	**n**
love	90
child	51
found	48
hope	41
companion	38
garden	37
god	36
glad	32
friend	31
mother	29

2. Comparing the three sentiment dictionaries: visualize sentiments using 4 lexicons

By checking the line number of each chapter, I decided to use 100 lines as my index length. And to make the result comparable, I only use the sentiment "negative" and "positive" classifications from the "nrc", "bing", and "loughran" lexicon.

word	n
love	90
child	51
found	48
hope	41
companion	38
garden	37
god	36
glad	32
friend	31
mother	29



The four different lexicons for calculating sentiment give results that are different in an absolute sense but have similar relative trajectories throughout the novel. We see similar dips and peaks in sentiment at about the same places in the novel, but the absolute values are significantly different.

As the book "Text Mining with R" said, "The AFINN lexicon gives the largest absolute values, with high positive values. The lexicon from Bing has lower absolute values and seems to label larger blocks of contiguous positive or negative text. The NRC results are shifted higher relative to the other two, labeling the text more positively, but detecting similar relative changes in the text. We find similar differences between the methods when looking at other novels; the NRC sentiment is high, the AFINN sentiment has more variance, the Bing sentiment appears to find long stretches of similar text, but all three agree roughly on the overall trends in the sentiment through a narrative arc.”

Matching with the plotline:

I notice that when the index is about 90, there is an obvious peak followed by a significant dip, which aroused my interest to detect what is the specific plot in this area.

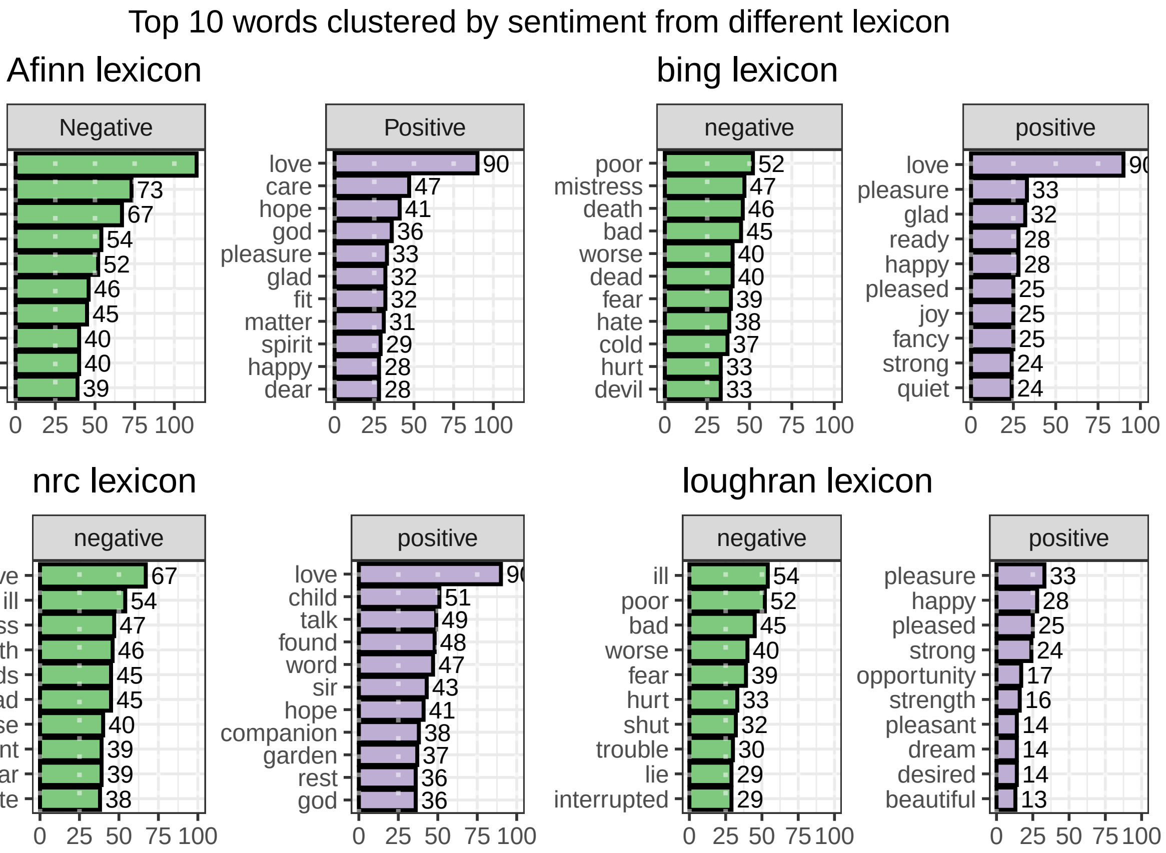
Line 9000-9100 is in CHAPTER XXIV. Catherine shares her story with Ellen when she went to Wuthering Heights secretly.

At first, it is a peaceful and lovely memory, as it said “She brought me some warm wine and gingerbread and appeared exceedingly good-natured, and Linton sat in the armchair, and I in the little rocking chair on the hearth-stone, and we laughed and talked so merrily”.

Then, Catherine told Ellen her private meeting was detected by Earnshaw, as it said “when Earnshaw burst the door open: having gathered venom with reflection. He advanced direct to us, seized Linton by the arm, and swung him off the seat”.

Suddenly the content is about “I gave him a cut with my whip, thinking perhaps he would murder me. He let go, thundering one of his horrid curses, and I galloped home more than half out of my senses.” which has strong negative sentiment.

3. Top 10 words clustered by sentiment from different lexicon:

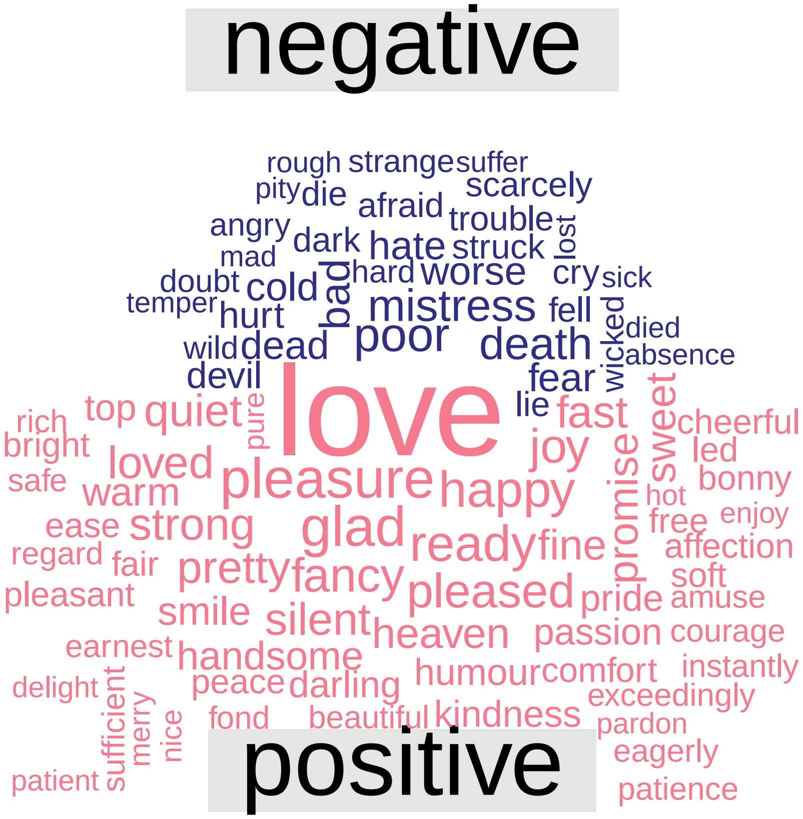


```
## NULL
```

Comparing the occurrences of the specific “positive, negative” vocabulary, we can tell that since different lexicon consists of different words (like evaluation criteria), it makes sense that the results of sentimental analysis above are different.

4. Word Cloud

Using "bing" lexicon to select out the most frequent words that belong to negative and positive sentiment in Wuthering Heights.



Inference:

https://www.tidytextmining.com/

http://yphuang.github.io/blog/2016/03/04/text-mining-tm-package/

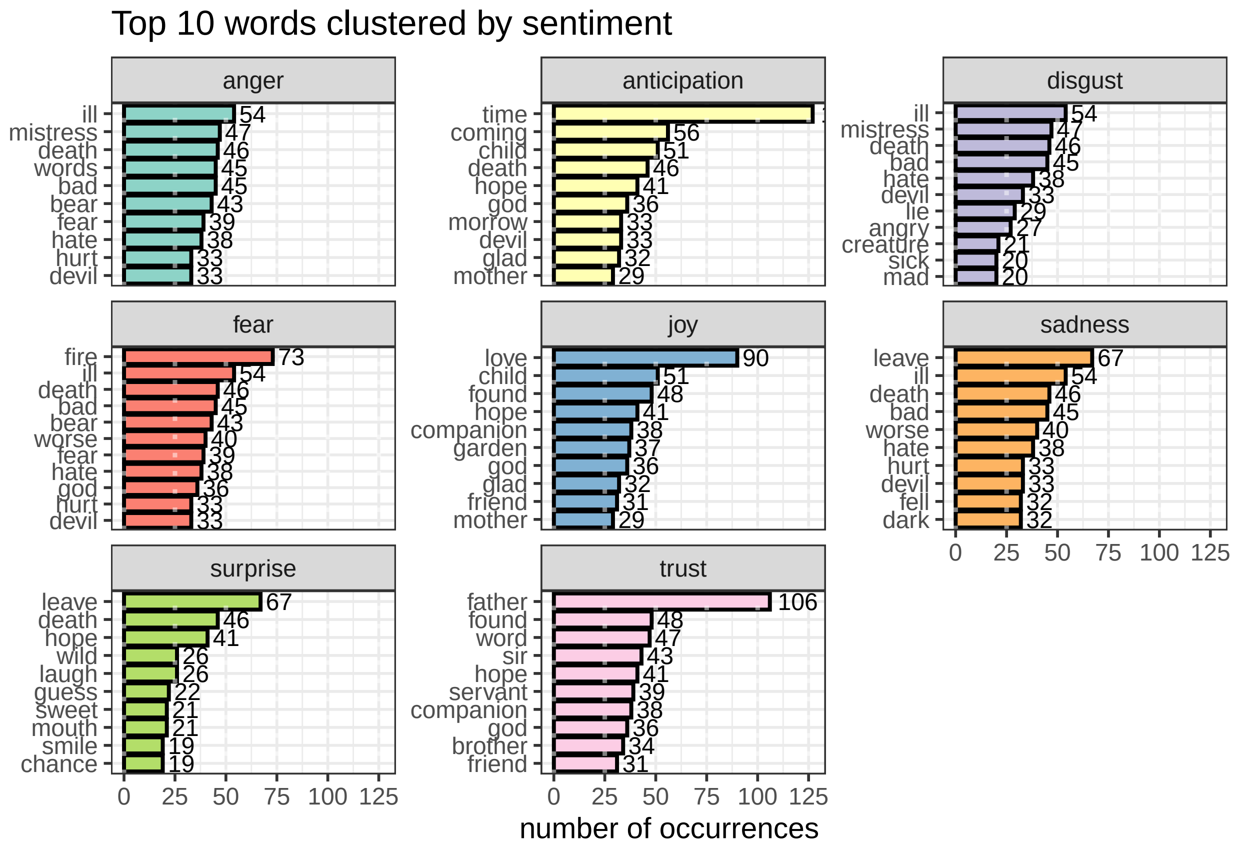
Appendix:

- About additional lexicons:

Actually, I don't think "loughran" is an ideal lexicon to analyze fiction because this lexicon is for financial contexts. As I search for other lexicons resources from GitHub, I noticed someone generate their personal additional lexicon, which aroused my interest. I want to try it for myself later.

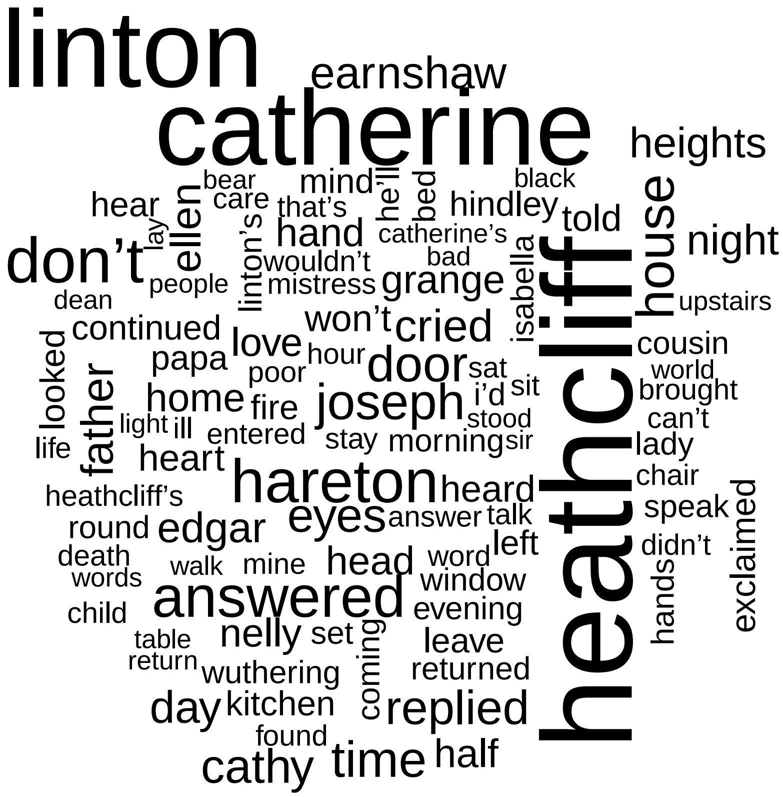
- Other EDA:

Visualize other sentiment from nrc lexicon: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.



Another word cloud:

Plot the most frequent words which have real meaning in the whole passage, instead of stop and custom words or personal pronouns.


linton
earnshaw
catherine
heights
bear
mind
he'll
bed
black
hear
care
that's
hindley
told
ellen
lay
hand
catherine's
night
don't
linton's
wouldn't
bad
isabella
house
heathcliff
dean
people
mistress
grange
upstairs
continued
won't
cried
looked
love
cousin
papa
hour
door
sat
world
father
poor
home
joseph
sit
brought
fire
i'd
light
ill
entered
stood
can't
life
stay
morning
sir
lady
heart
hareton
chair
heathcliff's
heard
round
edgar
eyes
answer
talk
speak
exclaimed
left
didn't
death
walk
mine
head
word
words
window
answered
hands
child
evening
table
nelly
set
coming
leave
return
wuthering
returned
day
kitchen
replied
found
cathy
time
half