MA615 Assignment4_Text Analysis

ZhangBiyao

12/9/2021

Task 1

Intriduction

The book I choose is The Cash Boy,a juvenile fiction. Here is the link https://www.gutenberg.org/ebooks/296. I used the gutenbergr package which provides access to the public domain works from Project Gutenberg collection to download this book.

The story is about a little poor boy, which lost everyone he loved except his little sister. Because the situation with no parents and income, the boy moved to New York for himself, so he could make some money for him and his sister.

Task 2

Preface

I used tidytext package which provides access to "AFINN", "bing" and "nrc", three general sentiment lexicons.

```
## # A tibble: 13,875 x 2
##    word        sentiment
##    <chr>       <chr>
##  1 abacus      trust
##  2 abandon     fear
##  3 abandon     negative
##  4 abandon     sadness
##  5 abandoned   anger
##  6 abandoned   fear
##  7 abandoned   negative
##  8 abandoned   sadness
##  9 abandonment anger
## 10 abandonment fear
## # ... with 13,865 more rows
```

```
## # A tibble: 6,786 x 2
##    word        sentiment
##    <chr>       <chr>
##  1 2-faces     negative
##  2 abnormal    negative
```

```
##  3 abolish     negative
##  4 abominable  negative
##  5 abominably  negative
##  6 abominate   negative
##  7 abomination negative
##  8 abort       negative
##  9 aborted     negative
## 10 aborts      negative
## # ... with 6,776 more rows
```

```
## # A tibble: 2,477 x 2
##    word       value
##    <chr>      <dbl>
##  1 abandon       -2
##  2 abandoned     -2
##  3 abandons      -2
##  4 abducted      -2
##  5 abduction     -2
##  6 abductions    -2
##  7 abhor         -3
##  8 abhorred      -3
##  9 abhorrent     -3
## 10 abhors        -3
## # ... with 2,467 more rows
```

Word frequencies

I look at word frequencies and the most common words in The Cash Boy.

```
## # A tibble: 2,486 x 2
##    word        n
##    <chr>   <int>
##  1 frank     291
##  2 boy       118
##  3 john      118
##  4 sir       106
##  5 wharton    91
##  6 wade       80
##  7 grace      79
##  8 bradley    68
##  9 house      58
## 10 uncle      53
## # ... with 2,476 more rows
```

According to the above results, we can conclude that "frank" (the cash boy's name)is the most common word. After looking at word frequencies, most of the words are name. Frank's experiences are unfortunate, I want to find some negative words, such as fear, and do the sentiment analysis about fear.

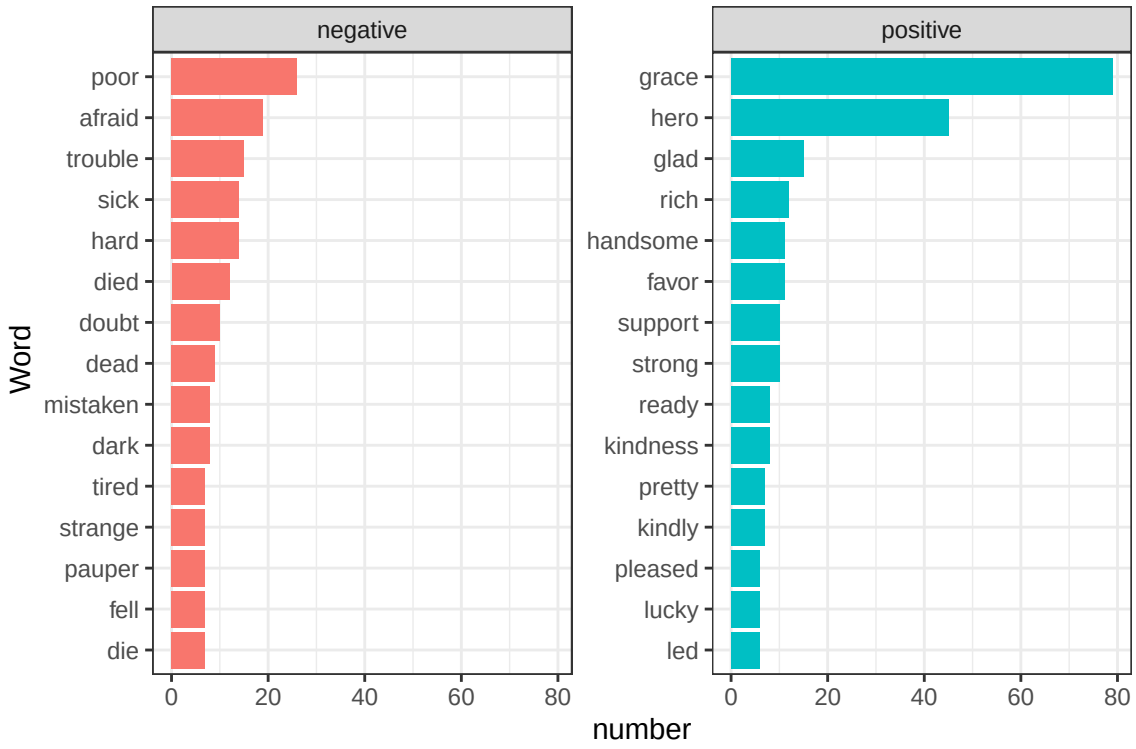
I also used wordcould package to visualize these words. Word cloud plot easily show the frequency of different words.

sentiment analysis

To analyze the sentiment of The Cash Boy, I consider the text as a combination of individual words and the sentiment content as the sum of individual words.

Firstly, I used "bing" to count the number of positive words and negative words.

```
## # A tibble: 547 x 3
##    word    sentiment     n
##    <chr>   <chr>     <int>
##  1 grace   positive     79
##  2 hero    positive     45
##  3 poor    negative     26
##  4 afraid  negative     19
##  5 glad    positive     15
##  6 trouble negative     15
##  7 hard    negative     14
##  8 sick    negative     14
##  9 died    negative     12
## 10 rich    positive     12
## # ... with 537 more rows
```



Secondly, I used "nrc" to select words which are negative.

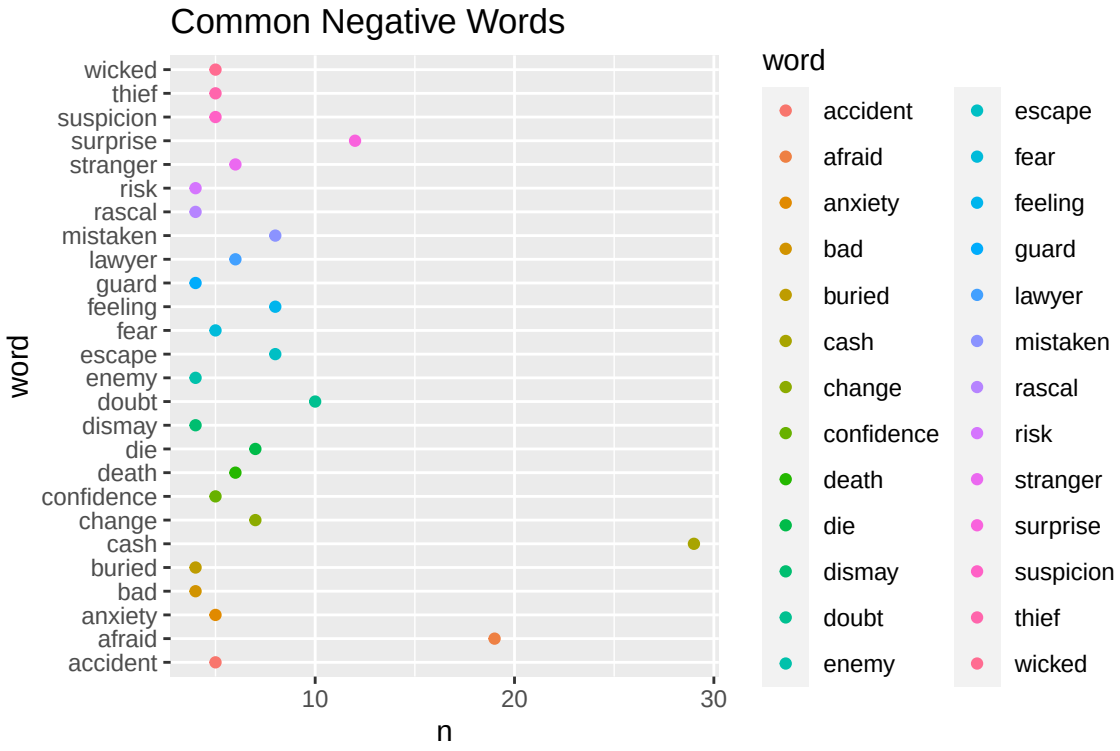
```
## # A tibble: 152 x 2
##    word          n
##    <chr>     <int>
##  1 cash         29
##  2 afraid       19
##  3 surprise     12
##  4 doubt        10
##  5 escape        8
##  6 feeling       8
##  7 mistaken      8
##  8 change        7
##  9 die           7
## 10 death         6
## # ... with 142 more rows
```

To tag positive and negative words in another way, I send datanto comparison.cloud() which can all be done with joins, poping, and dplyr.

According to the results of ff, there are indeed some words related to fear

Thirdly, I selected words with n>3 and visualize them.

```
## # A tibble: 26 x 2
##    word         n
##    <chr>    <int>
##  1 cash        29
##  2 afraid      19
##  3 surprise    12
##  4 doubt       10
##  5 escape       8
##  6 feeling      8
##  7 mistaken     8
##  8 change       7
##  9 die          7
## 10 death        6
## # ... with 16 more rows
```



According to the plot, it is obvious that "afraid" is the most common negative word related to fear.

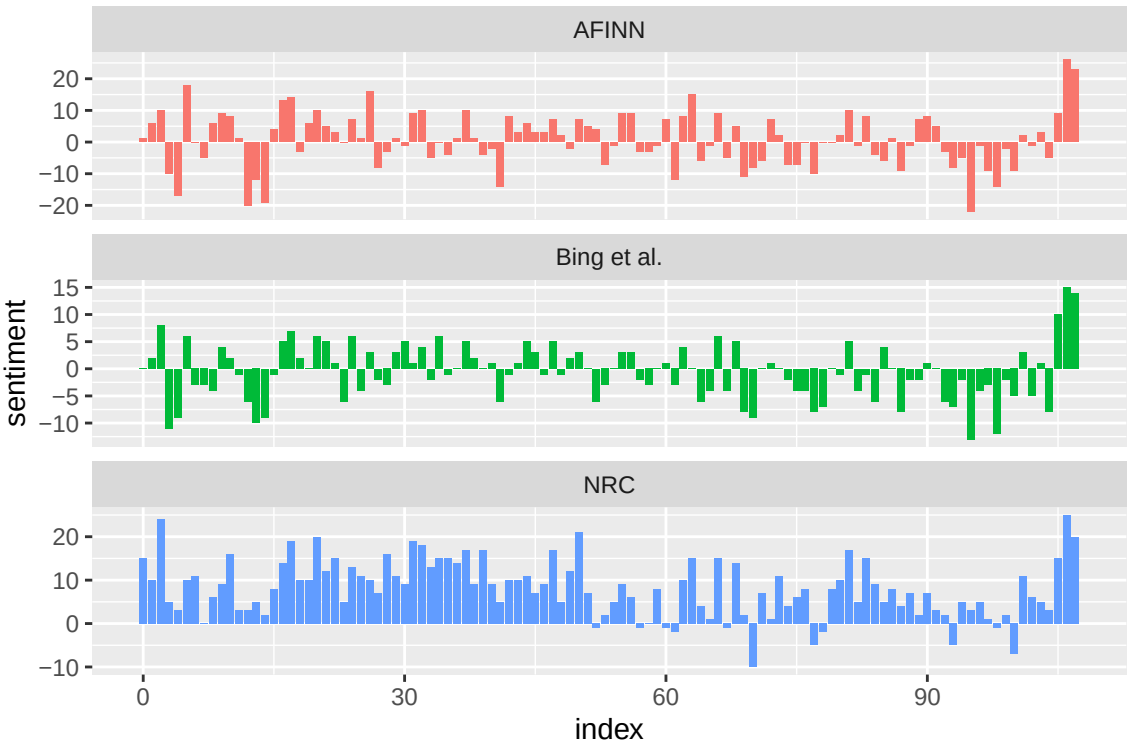
count up how many positive and negative words there are in defined section

I defined an index(integer division)to keep track of where we are in this fiction. This index counts up sections of 40 lines of the text. Besides, I also calculated a net sentiment(positive-negative).

```
## # A tibble: 108 x 4
##    index negative positive sentiment
##    <dbl>    <int>    <int>     <int>
##  1     0        5        5         0
##  2     1        0        2         2
##  3     2        1        9         8
##  4     3       13        2       -11
##  5     4       15        6        -9
##  6     5        5       11         6
##  7     6        5        2        -3
##  8     7       10        7        -3
##  9     8        7        3        -4
## 10     9        3        7         4
## # ... with 98 more rows
```

Comparative Analysis

I compared the results when different lexicons are used. It is obvious that these three lexicons differ in the kind of output they produce–signed real numbers, binary outcomes, multi-dimensional indicators. I visualized the sentiment score and examine how the sentiment changes across the fiction.



The results of three different lexicons are different in an absolute sense. AFINN lexicon gives the largest absolute values. Bing et al. lexicon has lower absolute values. The results of AFINN and Bing roughly on the the same trends in sentiment. However, the result of NRC lexicon is very different from AFINN and Bing. Most of the sentiment scores in NRC are positive which contradicts the plot of the novel.

look at how many positive and negative words are in these lexicons

```
## # 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
```

Based on my understanding of the plot of this fiction, bing lexicon is better than nrc lexicon.

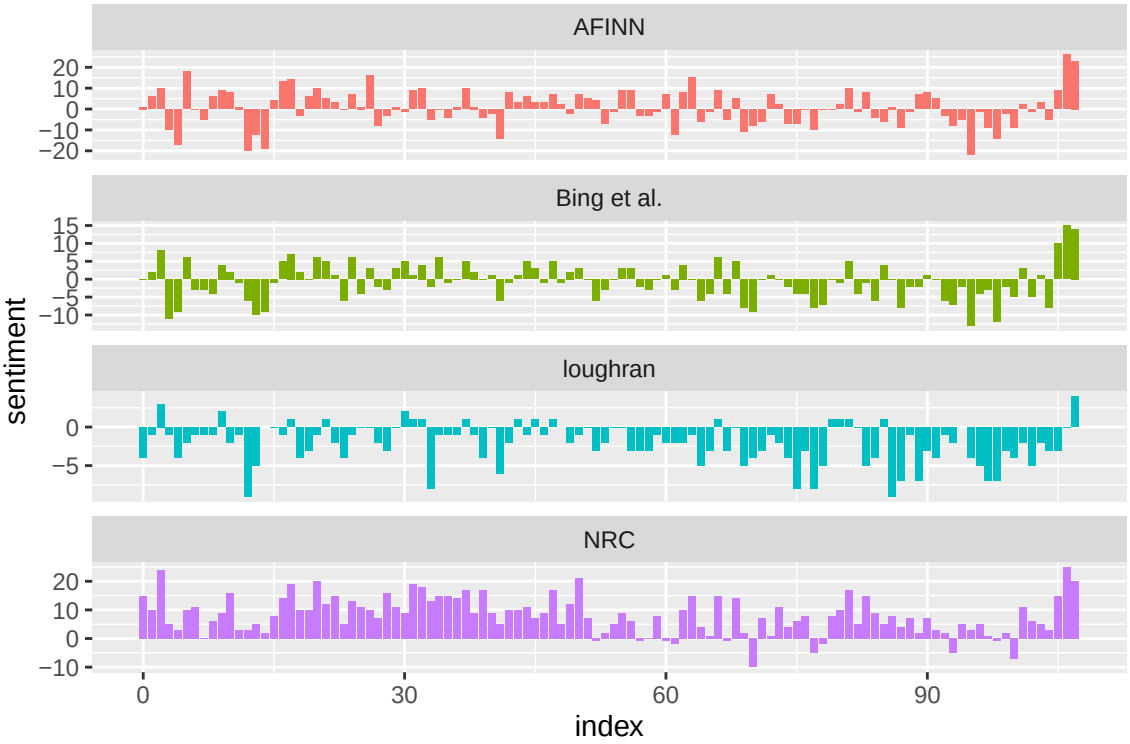
Conclusion

Based on the above comparative analysis, both bing lexicon and AFINN lexicon are suitable for sentiment analysis of The Cash Boy.

Extra credit

I tried loughran lexicon to make sentiment analysis, because its output is similar to bing lexicon and is in line with sad plots of The Cash Boy.

```
## # A tibble: 4,150 x 2
##    word         sentiment
##    <chr>        <chr>
##  1 abandon      negative
##  2 abandoned    negative
##  3 abandoning   negative
##  4 abandonment  negative
##  5 abandonments negative
##  6 abandons     negative
##  7 abdicated    negative
##  8 abdicates    negative
##  9 abdicating   negative
## 10 abdication   negative
## # ... with 4,140 more rows
```



Match Analysis

According to the above plots, the score of loughran has lower absolute value. As I mentioned in introduction part, the story is about a little poor boy, which lost everyone he loved except his little sister. At the begin, Frank lost parents, so he experienced sadness. Sentiment Scores change from positive to negative twice. After Frank arrived New York, he makes some money so that he can support himself and little sister. There are joys and sorrows in New York, so there are positive and negative scores. At the end of sotry, Frank walked out of sorrow, sentiment scores are positive. Therefore, bing lexicon and AFINN lexicon are suitable for sentiment analysis of The Cash Boy.

Reference

Julia Silge & David Robinson. (2016). Welcome to Text Mining with R [online]. Avalilable from: https://www.tidytextmining.com/index.html [accessed 9 December 2021].