task 1 and 2

Jinyu

12/07/2021

Task 1: Pick a Book

The book I chose is written by Charlse Dickens, and the book name is "Hard Times" and shortened as "A Christmas Carol" because we are gonna spend our final time of the year at Christmas.

TASK 2: Bag of Word Analysis

In this part, I will show the sentiment analysis by using AFINN, Bing and NRC respectively. I am going to plot several barplots to compare these 3 methods And show the differences of them.

The book, A Christmas Carol, is in general a book with more negative sentiments than positive sentiments.

To briefly summarize the book, at the very beginning, the book describes the background of Ireland in the 1840s, where people were suffering from hungers and coldness. The leading character of the novel is Scrooge, who is a scrooge literally. He loves money and does not spare his mercy to others. On Christmas eve, so many ghosts visited Scrooge's house and made him see his death. After the night, Scrooge realized that money would be gone one day and changed himself and

In conclusion, according to the plotline, the sentiment of the book should be negetive at first and positive at the very last, which I find Bing lexicon and Afinn lexicon both work well. It's hard to tell which one works the better. The only difference bewteen these 2 I find is that the sentiment by Bing lexicon is more negetive, which I think may fit the book better.

Consequently, in the following part, I will mainly do the sentiment analysis by using Bing lexicon.

The difference in the sentiment analysis with different lexicons

I now use 3 lexicons to get the sentiment analysis for every word in the book, and try to tell the difference among the 3 methods

The first one is Afinn lexicon(outcome is 1945rows X 5columns):

Table 1: Word-level analysis – Afinn lexicon

gutenberg_id	linenumber	chapter	word	value
46	1	1	ghost	-1
46	3	1	dead	-3
46	3	1	no	-1
46	3	1	doubt	-1
46	7	1	good	3
46	8	1	dead	-3

Table 2: Value column – Afinn lexicon

sentiment
-1
-3
3
-2
2
4
1
-4
5
-5

The second one is Bing lexicon(outcome is 1926rows X 6columns):

Table 3: Word-level analysis – Bing lexicon

gutenberg_id	linenumber	chapter	word	sentiment	method
46	3	1	dead	negative	Bing et al.
46	3	1	doubt	negative	Bing et al.
46	6	1	mourner	negative	Bing et al.
46	7	1	good	positive	Bing et al.
46	8	1	dead	negative	Bing et al.
46	12	1	dead	negative	Bing et al.

Table 4: Sentiment column – Bing lexicon

sentiment
negative
positive

The third one is nrc lexicon(outcome is 6203rows X 6columns):

Table 5: Word-level analysis – NRC lexicon

gutenberg_id	linenumber	chapter	word	sentiment	method
46	1	1	ghost	fear	NRC
46	3	1	doubt	fear	NRC
46	3	1	doubt	negative	NRC
46	3	1	doubt	sadness	NRC
46	3	1	doubt	trust	NRC
46	4	1	burial	anger	NRC

Table 6: Sentiment column – NRC lexicon

sentiment
fear
negative
sadness
trust
anger
anticipation
joy
positive
surprise
disgust

According to the tables from 1 to 6, we can tell that:

In afinn lexicon, the value column represents the sentiment of each word ranged from -5 to 5, where negative values refer to the negative sentiment and positive values represent the positive sentiment.

In the Bing lexicon, the sentiment column represents the sentiment of each word with 2 kinds of outputs- “negative” and “positive”, which is easy to tell what they represent respectively.

In the NRC lexicon, the sentiment column also represents the sentiment of each word with not only “negative” and “positive”, but also other outputs like “fear”, “sadness”, “anger”,“anticipation”, “disgust”, “joy”, “trust”, and “surprise”, so the sentiment description in NRC lexicon is much more detailed.

And because of the different numbers of words in lexicons, the outcomes of inner join different. The results inner joined by nrc lexicon is larger because there are more words in the nrc lexicon.

In my opinion, the values in afinn lexicon and the sentiment in Bing lexicon(which can be transformed into dummy variables) are machine-readable and easy to process. But for the nrc lexicon, for now I can only take the “positive” and “negative” into account to make it machine-readable and easy to process because of the lack of assessments of other sentiments.

Figure 1

Figure 1 is the sentiment progress for every 80 lines in this book with different books. We can tell the sentiment changes through the progression of plotline. And different lexicons show a little different result.

From my perspective, the sentiment analysis by Bing lexicon better explains the sentiment of the book as the progression of the plots.

Now I will pick up the Bing lexicon as the main lexicon to do the word count analysis and other analysis.

Figure 2

Figure 2 presents the top 10 negative and positive word count. In the negative barplot chart, the word “poor” is the most common word followed by word “cold” and “dark”. In the positive part, The most frequent word is “good” with “like” and “great” following after.

Figure 3

Figure 3 displays word cloud where we can get the frequency for top 100 words with the size representing the word count. As is shown in the figure, “Scrooge”, the leading character in this book, is the most common

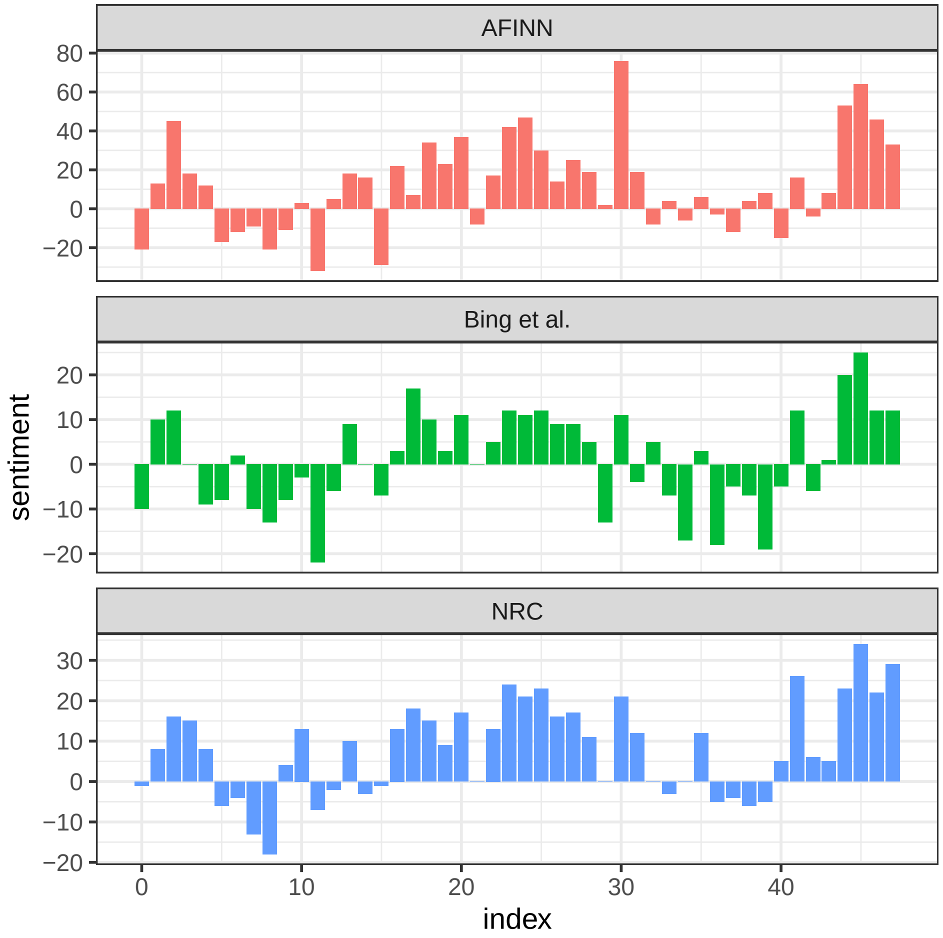

Figure 1: sentiment plot for A Christmas Carol

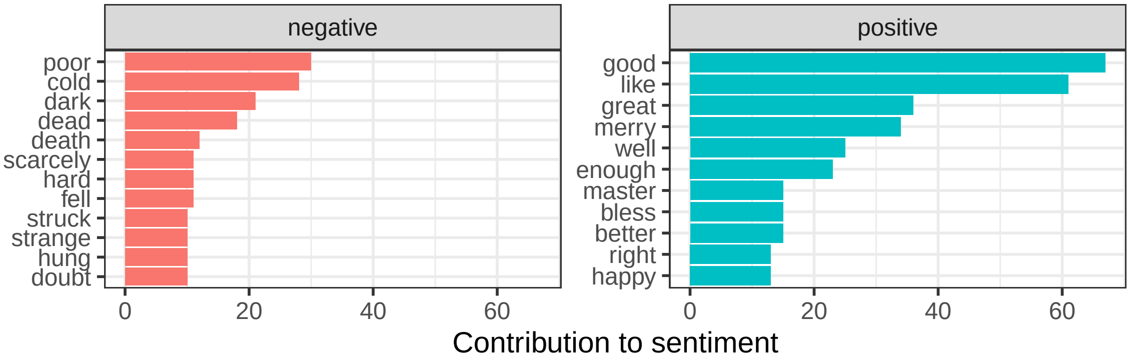

Figure 2: negative positive words count

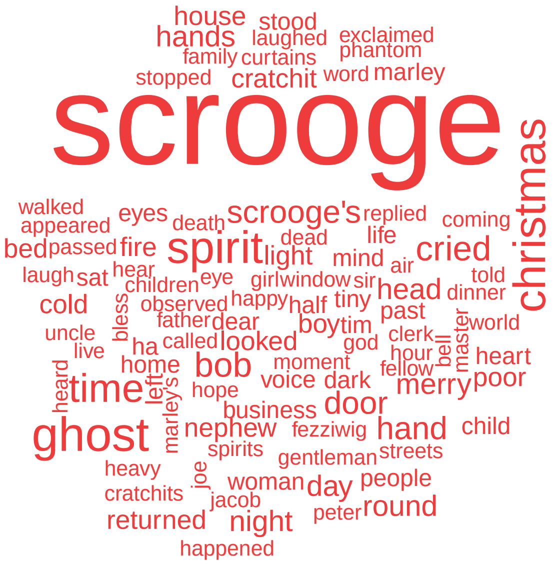

Figure 3: word cloud

word by bing lexicon, which makes sense in this case. And the "Christmas", "ghost" and "spirit" follow after, which represent the period of time, objects talking about in the main plotline. the It is reasonable because they are the main characters in that fiction book. And we can also find "Bob", another main character in this fiction, which I will do the character analysis along with Scrooge.

Figure 4

Figure 4 shows the word frequency of "positive" and "negative" words, where the size represents the word count. This part is also shown in Figure 2, but now we show them in a different way.

Summarization

To summarize, from Figure 1, we can tell the sentiment changes through the progression of plotline. And different lexicons show a little different result. But here, as I said at the beginning, I choose Bing lexicon as the best lexicon to explain.

At the very beginning, the sentiment of the book is negative, which corresponding to the content of the book which tells the background of the story and it's a poor situation in Ireland. Later on, the book tells that Scrooge's living situation and he is rich, and in Figure 1 the sentiment is positive. And, the sentiment turns to negative for a while, goes back to positive and this process repeats twice. And that makes sense for the book because Scrooge experiences something bad at Christmas Eve and then changes himself into a good man at last.

From Figure 2 and Figure 4, we can capture the top 100 most frequent sentimental words exist in this book. The most frequent negative one is "poor" and it can be translated into a poor situation in Ireland in 1840s,

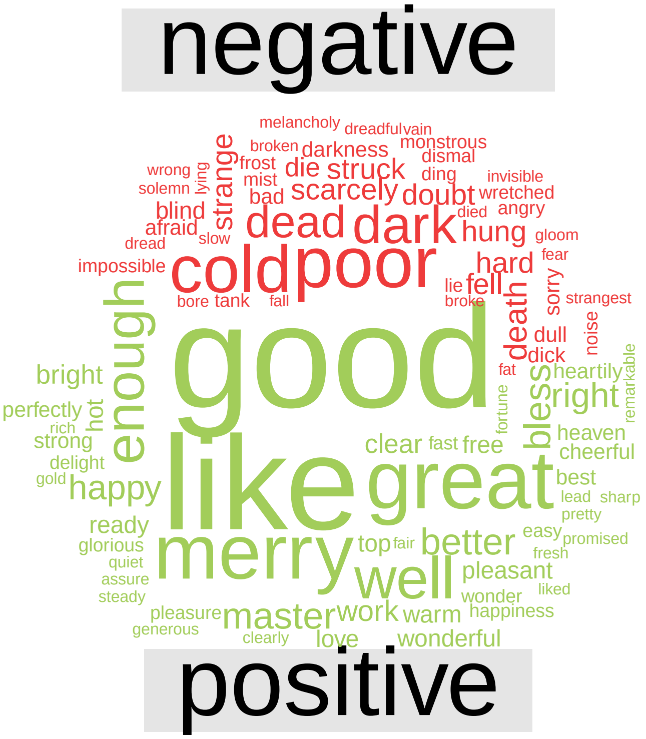

Figure 4: Word Cloud with Sentiment Analysis

which is the background of the story. The most frequent positive one is "good", which is general in many books and here it can also represent the personality of Scrooge at last.

In Figure 3 part, I've already explained something below the figure, and the count of word in this book gives us the information about who is the leading character–Scrooge.

Extra Credit

In addition to the 3 lexicons I used, I noticed that there is another lexicon called "Loughran-McDonald". Now, I am going to use this lexicon to make some similar plot and give an general idea of the plotline of the book from the start to the end.

According to the plot we found that the sentiment in some part are zero, which doesn't happen in other lexicons. In the webpage https://sraf.nd.edu/textual-analysis/contributed-materials/, I notice that this lexicon is specially for accounting and financial documents.

Consequently, the conclusion can be drawn that this lexicon is not very suitable for fiction sentiment analysis, which can also tell in Figure 5 because the sentiment does not properly match the plotline in the book.

```
df_text <- tnum.query('Charlse_Dickens/A_Christmas_Carol_4# has text',max=60) %>% tnum.objectsToDf()

(df_text %>% select(subject:string.value)%>% head())
```

```
##                                                                 subject
## 1                              Charlse_Dickens/A_Christmas_Carol_4/heading:0001
## 2 charlse_dickens/a_christmas_carol_4/section:0001/paragraph:0001/sentence:0001
## 3 charlse_dickens/a_christmas_carol_4/section:0001/paragraph:0001/sentence:0002
```

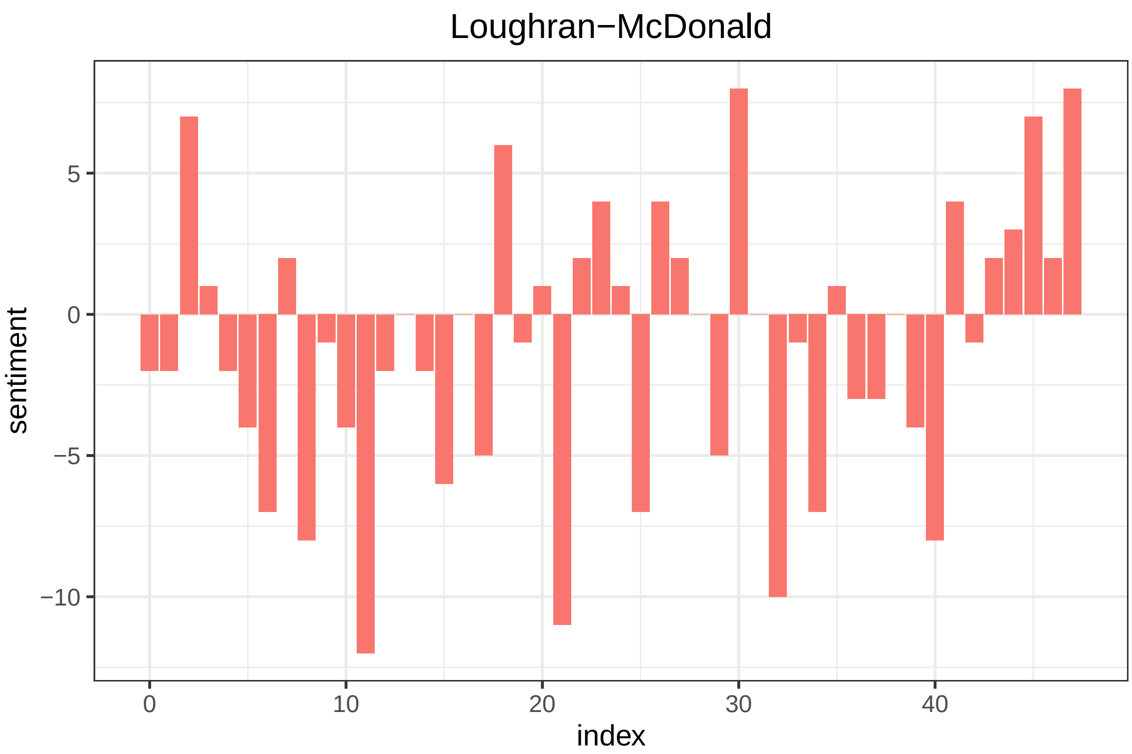

Figure 5: Sentiment Analysis by LM lexicon

```
## 4 charlse_dickens/a_christmas_carol_4/section:0001/paragraph:0001/sentence:0003
## 5 charlse_dickens/a_christmas_carol_4/section:0001/paragraph:0001/sentence:0004
## 6 charlse_dickens/a_christmas_carol_4/section:0001/paragraph:0001/sentence:0005
##   property
## 1     text
## 2     text
## 3     text
## 4     text
## 5     text
## 6     text
##                                                                                                    str
## 1                                                                              "<CHAPTER I:  MARLEY’
## 2                                                                           "MARLEY was dead: to beg
## 3                                                                    "There is no doubt whatever abou
## 4 "The register of his burial was signed by the clergyman, the clerk, the undertaker, and the chief
## 5    "Scrooge signed it: and Scrooge’s name was good upon ’Change, for anything he chose to put his
## 6                                                                    "Old Marley was as dead as a doo
```

Reference

1. A Christmas Carol in Prose; Being a Ghost Story of Christmas by Charles Dickens
2. Software Repository for Account and Finance
3. Text Mining with R
4. The ideas and supports from my dear classmate Yuli Jin.