text mining

Tao He

11/26/2021

Introduction

I choose ***Heidi*** which was written by Johanna Spyri to do the text mining, because I really love it since I was a little girl. I decided to find the sentiment trend in this book.

Task 1

Data

The book is downloaded from the gutenbergr package, which helps download and process public domain works from the Project Gutenberg collection, including downloading books.

After loading ***Heidi***, I change the whole text to a data frames of individual words, which to manipulate, summarize, and visualize the characteristics of text easily, including putting it into different chapters, marking the text as individual words, removing all punctuation and capital letters and adding the line numbers. Then, I listed the initial six rows of the data set, and we can see the book id in the "gutenbergr" package, chapters, words and line numbers in the Heidi.

gutenberg_id	linenumber	chapter	word
1,448	1	0	heidi
1,448	3	0	by
1,448	3	0	johanna
1,448	3	0	spyri
1,448	7	0	contents
1,448	9	0	i

Chapter 0 is the text before the Chapter 1, which always include author name, book name, public data, content and so on.

Then, when we looked at that the word that appears most in the books is "heidi", except some stop-words, like "the", "and", "to", "her", "she", "of" and so on. Also, the second word is "peter", who is Heidi's best friend in the mountain. It comes as no surprise since the main character of this book is Heidi, and the story is also around her and the people around her.

word	n
heidi	924
peter	335

word	n
child	291
grandfather	281
clara	270
grandmother	242
day	239
time	195
mountain	181
uncle	151

Now, we take a quick look at the words whose the frequency of words over 150 times, except those stop words.

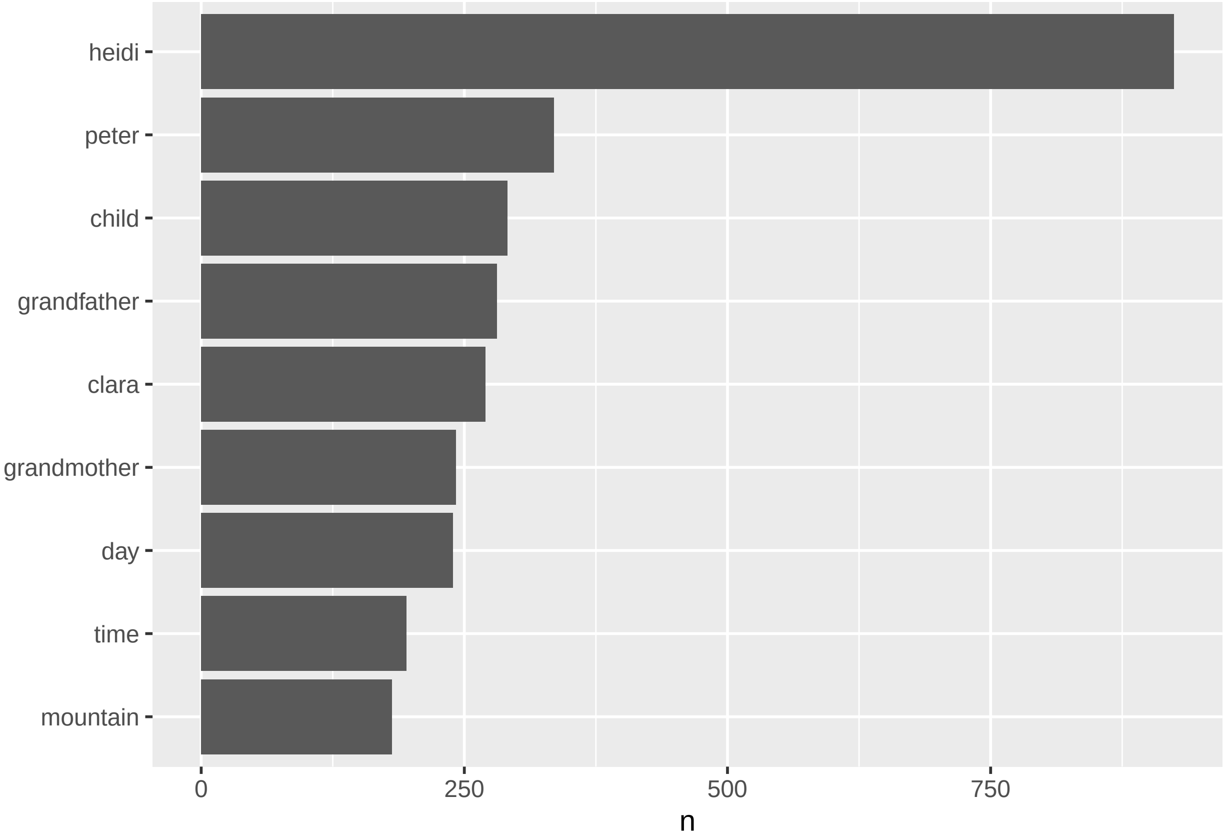


Figure 1: Most words in Heidi

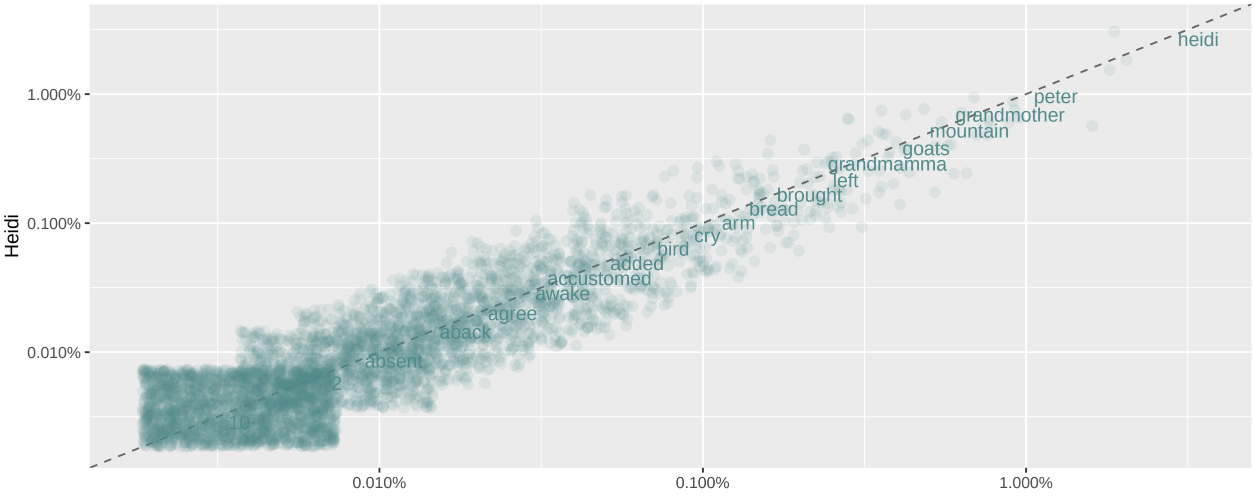


Figure 2: Frequencies words across all the novels

In figure 2, words that are close to the line in these plots have similar frequencies across all the novels. For example, words such as “heidi”, “peter”, “grandmother” are fairly common and used with similar frequencies across most of the books. Words that are far from the line are words that are found more in one set of texts than another.

Task 2

Sentiment Display

Now, we start the sentiment analysis. We use the sentiment dictionary from tidyverse package, which contains dictionaries for different sentiment categories, such as ***afinn***, ***bing*** and ***nrc***, to perform the analysis.

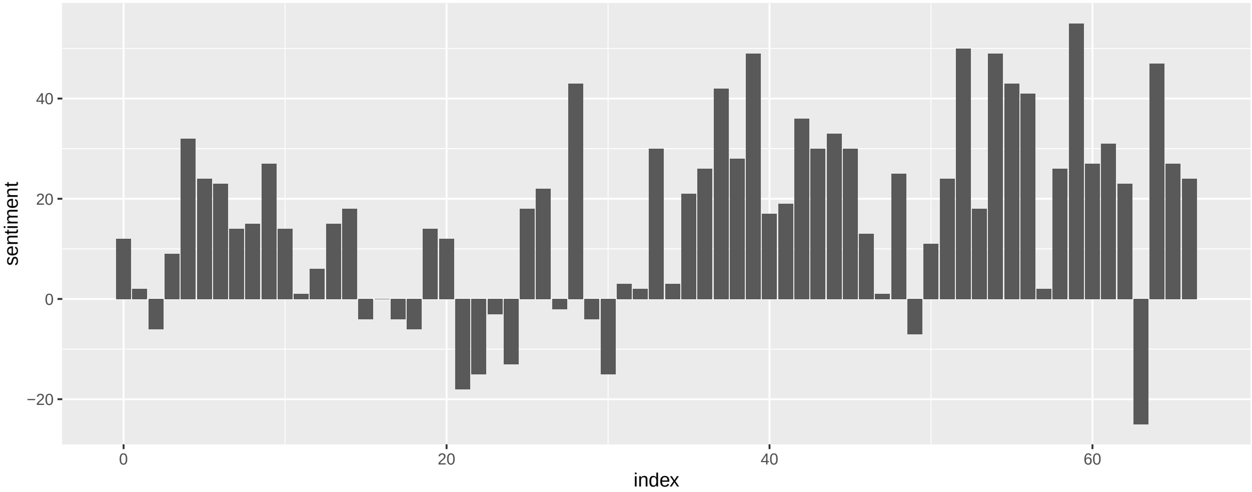
The sentiments datasets

Now that the text is in a tidy format with one word per row, we are ready to do the sentiment analysis. First, let’s use the NRC lexicon to see how many “joy” words in ***Heidi***.

word	n
child	291
good	188
god	82
sun	75
beautiful	65
happy	65

We see mostly positive, happy words about hope, friendship, and love here, like “sun” and “beautiful”. We also notice that **“child”** is also labeled as **“joy” words** in ***nrc*** sentiment dictionary, which makes sense since Children are symbolic of innocence, vitality, and are synonymous with beauty.

Then, we use the ***bing*** sentiment dictionary to find the change sentiment scores across the plot trajectory by using every 150 lines.



We can see how the plot of text changes toward more positive or negative sentiment over the trajectory of the story by using the "bing" directory in this figure . It is clear to observe that there are far more positive words than negative words in this book, probably because although Heidi is a very unlucky girl who grew up with both parents dead, living in poverty and raised by her aunt, she naturally loves life, loves nature, helps people, is full of love and care for others, and the people around her life gain joy because of her. It is under the infection of her innocent feelings that her grandfather, who is full of vicissitudes and depression, becomes cheerful.

Comparing the three sentiment dictionaries

Since the definitions of positive and negative emotions are different in different emotion dictionaries, and I try to use different sentiment dictionaries to represent the emotion words in the book.

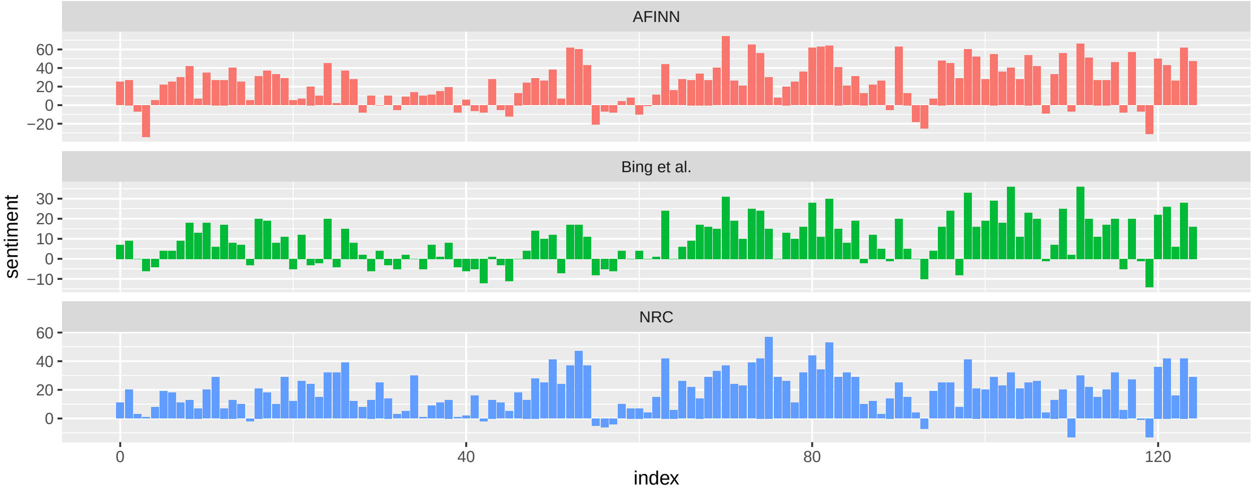


Figure 3: Comparing three sentiment lexicons with Heidi.

The three different dictionaries used to calculate emotions give results that differ in absolute terms, but have similar relative trajectories in the novel. We see similar emotional lows and peaks in almost the same places in the novel, but there are significant differences in absolute values. The ***"NRC"*** dictionary gives maximum absolute values with high positive values. ***"Bing et al.'s"*** dictionary and ***"AFINN"*** have lower absolute values and seem to mark more consecutive negative text blocks. Emotion seems to find longer similar texts, but all three broadly agree on the general trend of emotion through the book.

Then, I decided to take a quick look at how many positive and negative words are in ***Heidi*** by using those lexicons.

In Bing,

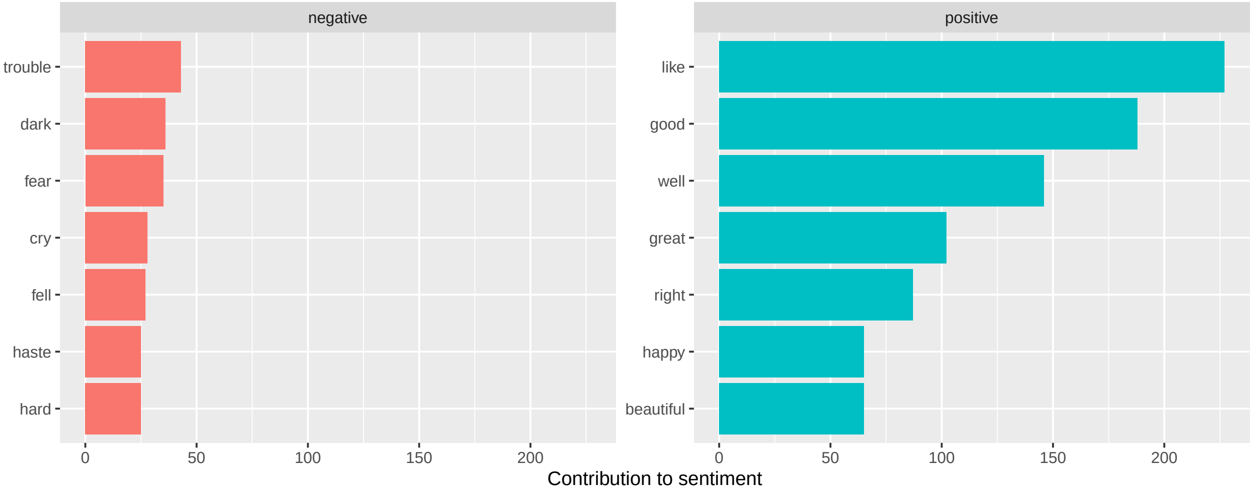
sentiment	n
negative	1,874
positive	2,979

In NRC,

sentiment	n
negative	1,904
positive	4,269

It is clear that compared with “bing” lexicon, by “NRC” ***Heidi*** has more positive words, however, it has more negative words as well. At the same time, we can see that the number of positive words is far more than the number of negative words, by nearly 3 times.

Most common positive and negative words



Also, we found that “like” was used as the word that contributed the most to the positive sentiment, but it should be noted that like is positive when used as a verb, but does not mean anything when used as a preposition, so it is more numerous than any other word because it is counted when used as both a verb and a preposition.

Wordclouds

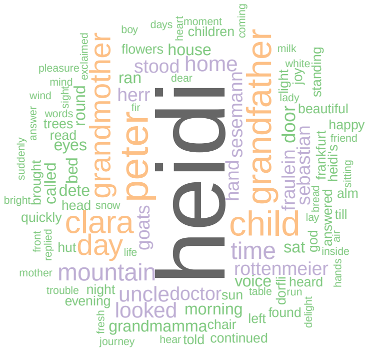


Figure 4: The most common words in Heidi.

Then, when we looked at that the word that appears most in the books is “heidi”, except some stop-words, like “the”, “and”, “to”, “her”, “she”, “of” and so on. It comes as no surprise since the main character of this book is Heidi, and the story is also around her and the people around her.

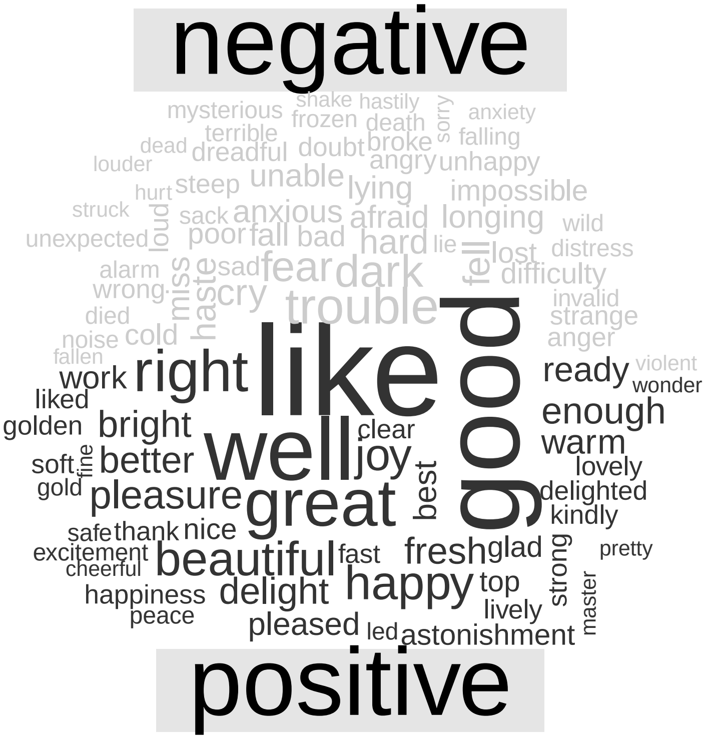


Figure 5: Most common positive and negative words in Heidi.

The size of a word’s text in this figure is in proportion to its frequency within its sentiment. We can use this visualization to see the most important positive and negative words, but the sizes of the words are not comparable across sentiments.

Task 2 Extra Credit

Moreover, I find another lexicon in tidyverse, called “loughran”, to figure out the change sentiment scores across the plot trajectory by using every 150 lines.

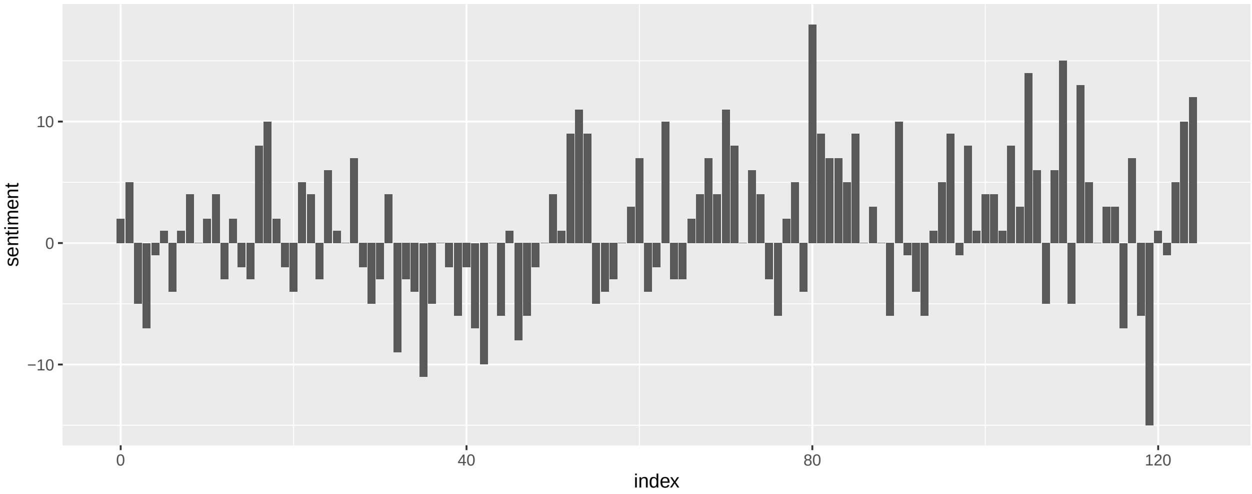


Figure 6: Sentiment through the narratives of Heidi by using “bing” lexicon.

By using “loughran” lexicons, we can see in the very first part, the number of negative words is more than the number of positive words, especially between 2400 and 4000 words. However, the later part of the book, the number of positive words is more than the number of negative words, which makes sense since this is a very warm book about family and friendship and the story has a happy ending.

sentiment	n
negative	881
positive	1,050

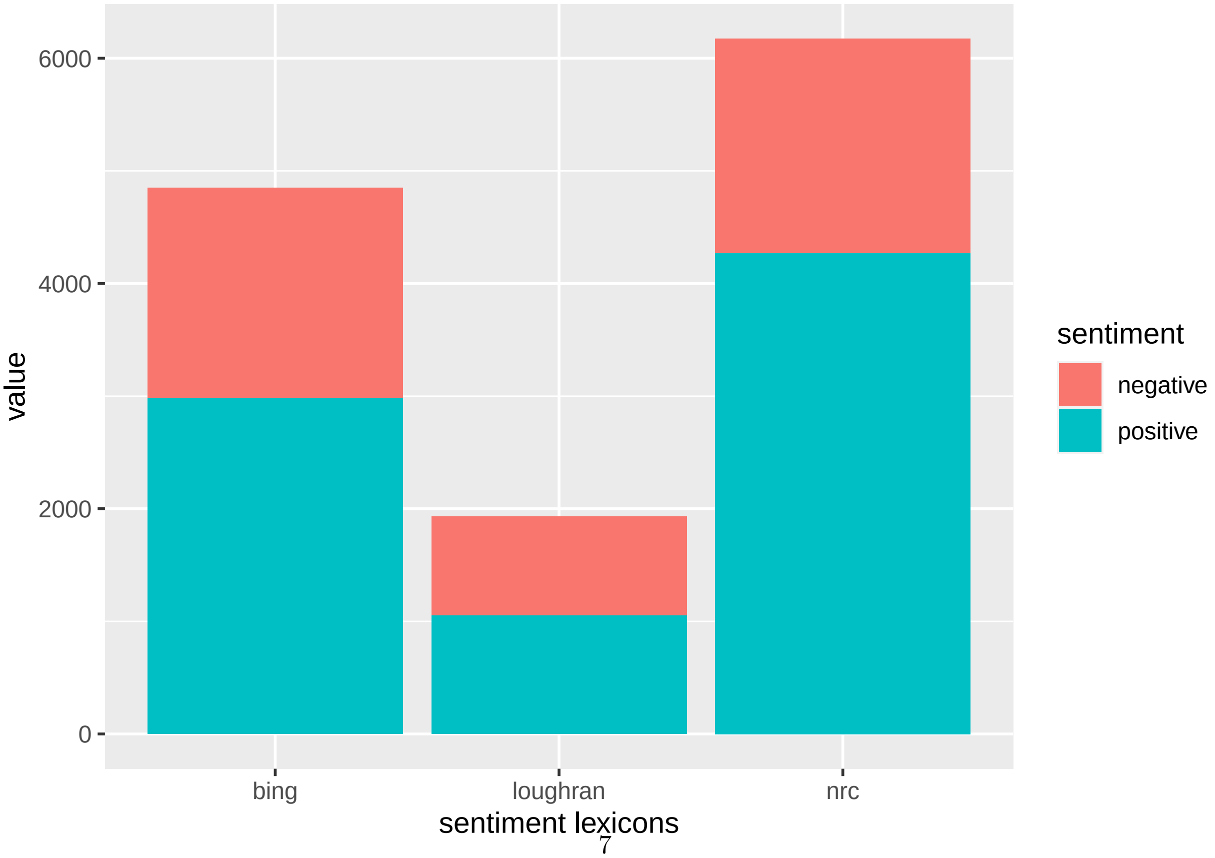


Figure 7: How many negative words and positive words in Heidi by different lexicons.

Obviously, “nrc” lexicon has the greatest number of negative words and “loughran” lexicon has the lowest number of negative words.

Conclusion

In short, after I did the sentiment analysis, I found that ***Heidi*** is a warm story with a happy ending although the number of negative words is more than the number of positive words at the very first beginning.

Reference

Data Source:
Johanna Spyri, Heidi, (September 1, 19989), from https://www.gutenberg.org/ebooks/1448

Works Cited:

David Robinson, gutenbergr: Search and download public domain texts from Project Gutenberg, (May 28, 2021), from https://cran.r-project.org/web/packages/gutenbergr/vignettes/intro.html

Julia Silge and David Robinson, Text Mining with R: A Tidy Approach, (June 8, 2017), or from https://www.tidytextmining.com/

Github[https://github.com/MA615-Yuli/MA615_assignment4_new/blob/main/TASK3.Rmd], https://github.com/MA615-Yuli/MA615_assignment4_new/blob/main/TASK3.Rmd

Haviland Wright, tnum_instructions and examples, https://learn.bu.edu/ultra/courses/_80585_1/cl/outline