Task 1 & Task 2

Yujia Wang

12/8/2021

Task 1: Pick a book

Alice's Adventures in Wonderland is a children literature that I loved to read when I was young. So I chose this book to look back at my childhood.

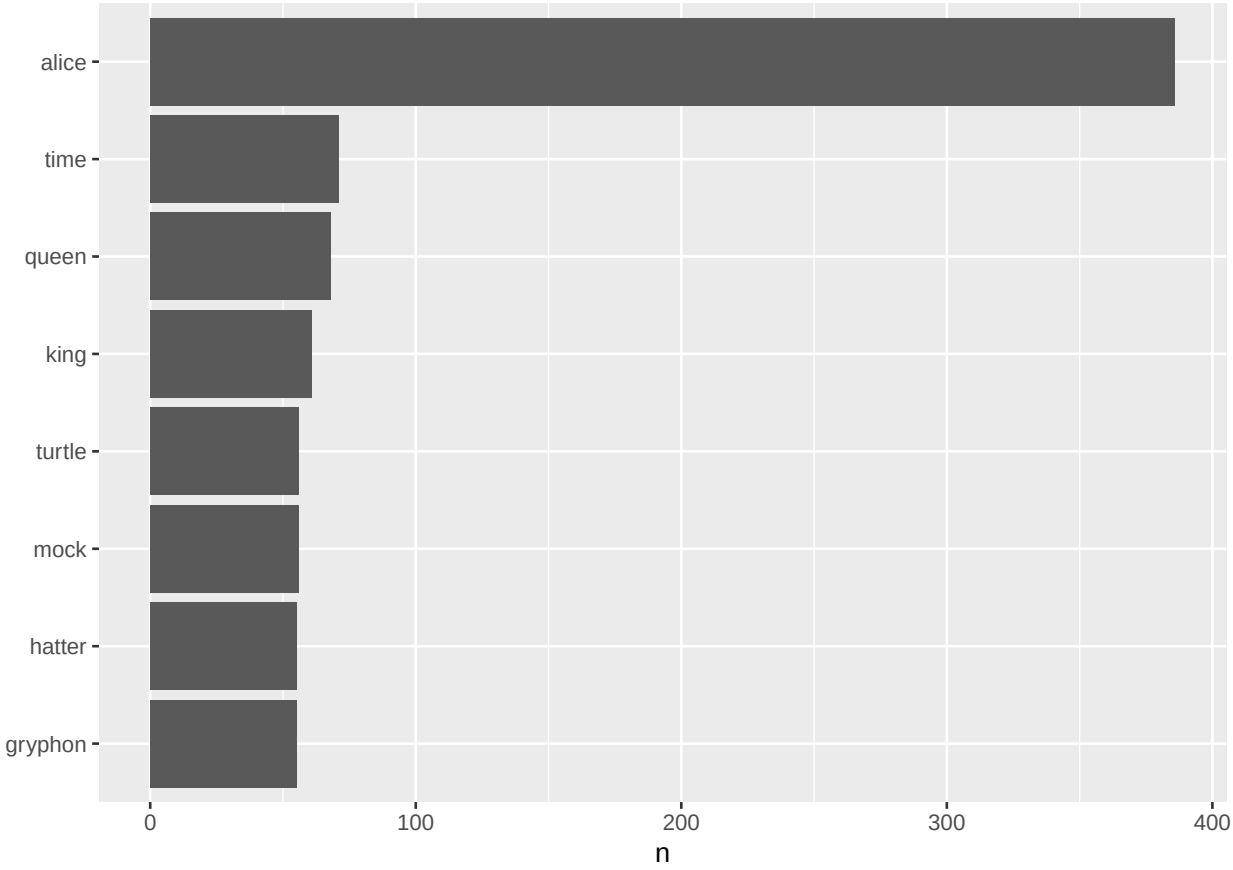

Figure 1: the eight most common words

I calculated the eight most frequent words in the book, and the results are as follows. It can be found that because Alice is the protagonist, the frequency of using her name is much higher than other words.

At the same time, there are 5 words that are the names of characters, such as Queen, King, Turtle, Hatter and Gryphon, which shows that they are the main supporting roles of the book.

Time is the second most frequent word in this book, reflecting Alice raced against time in her adventure.

Task 2: bag of word analysis

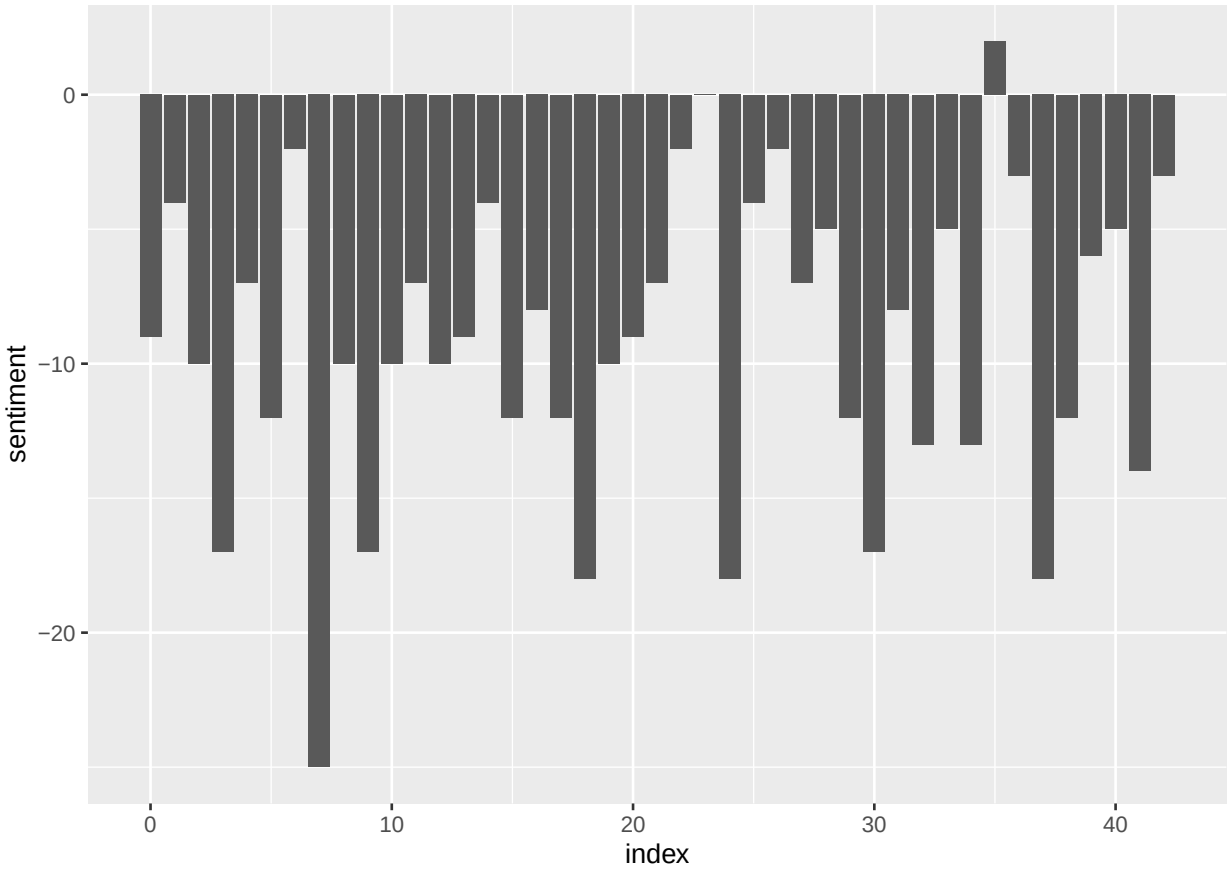

Figure 2: sentiment scores

Based on the plot trajectory of Alice's Adventures in Wonderland, I calculated the emotional score. The x-axis tracks the narrative time in the text part, and the y-axis is the difference between positive and negative emotions. It can be seen that almost all the plots in this book are negative emotions.

This book is about Alice falling into a rabbit hole by accident. She met a lot of strange people and animals, such as the playing cards behind a small door, the rough queen of hearts. In this world of fantasy and madness, Alice fights against violence and helps others.

So figure 2 coincides with the book's absurd, bizarre fictional scenes, weird characters, and thrilling adventure stories.

Comparing the three sentiment dictionaries

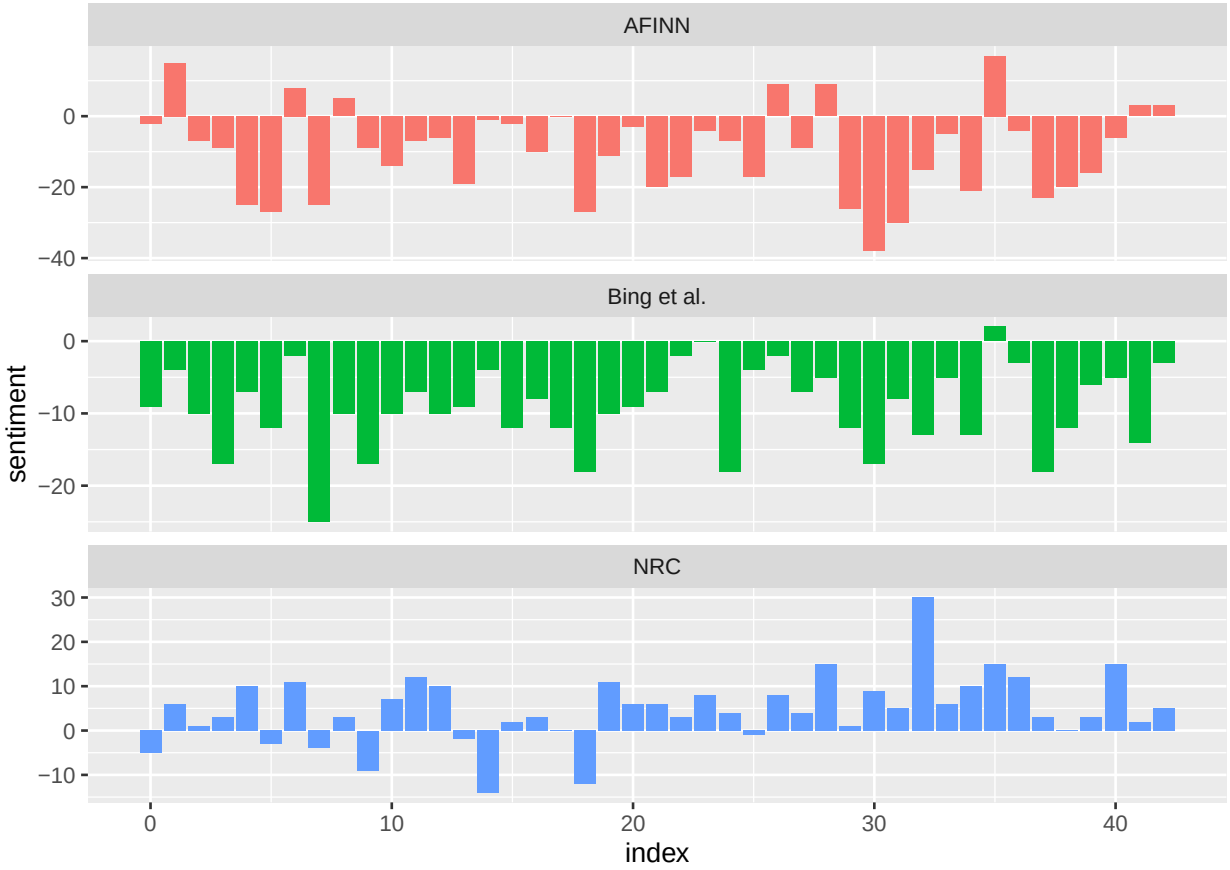

Figure 3: comparing the three sentiment dictionaries

According to text mining with R, I imported AFINN dictionary, Bing dictionary and NRC dictionary. It can be seen that the results of AFINN and Bing tend to be consistent, but the results of NRC are more positive, and the result of AFINN has more variance.

The most common positive and negative words

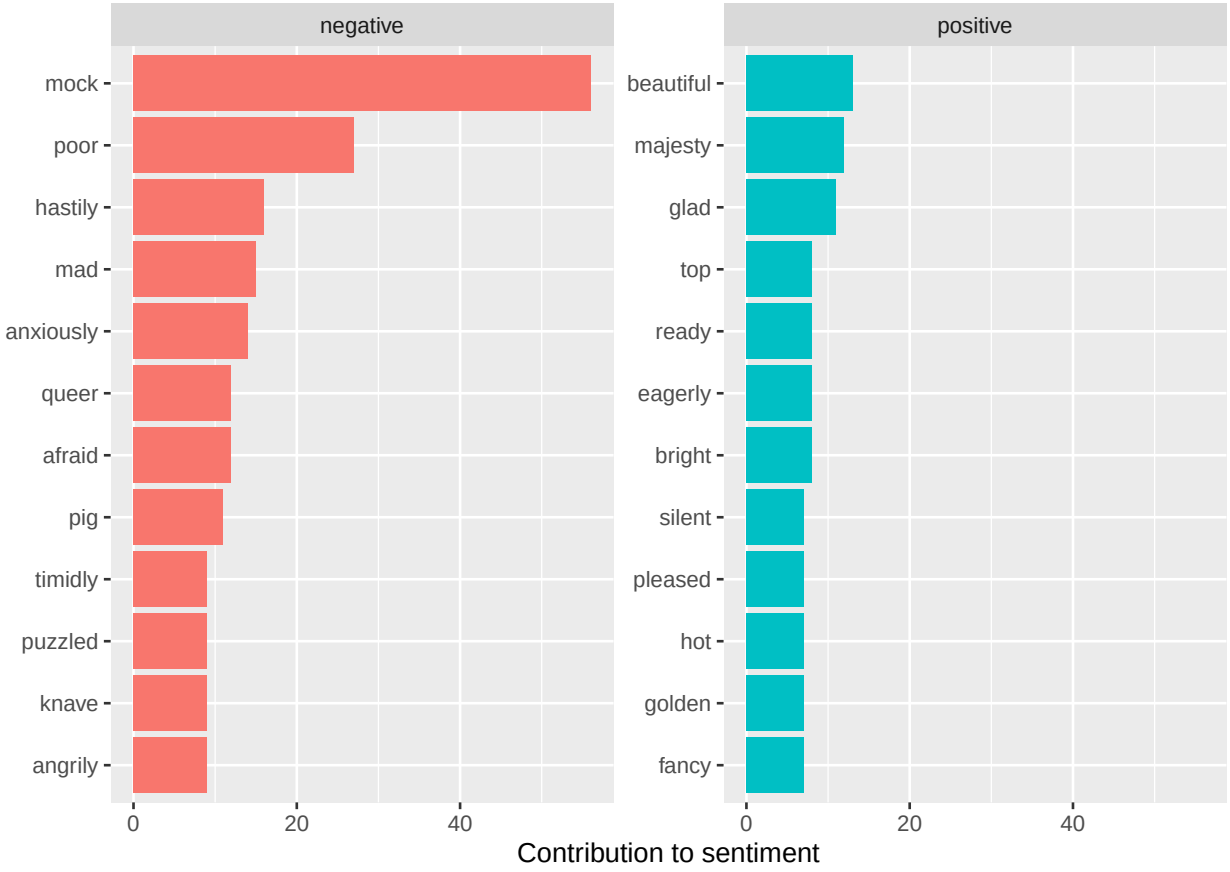

Figure 4: the most common positive and negative words

Figure 4 shows the 12 most frequently positive and negative words. We can see that negative words are more than positive words in the book. Among them, mock is usually used to describe the action state of the characters, and express the absurd content of the story. And beautiful is more used to express the environment, reflecting that Alice has entered a fantasy and extraordinary world of fiction.

Word cloud

Subsequently, I drew two word cloud plots of words frequency. These are the visualization of the previous plot.

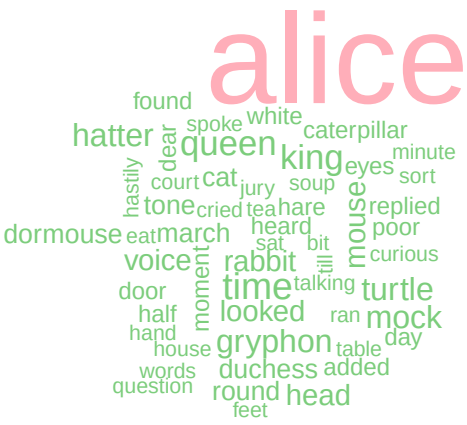

Figure 5: sentiment word cloud version 1

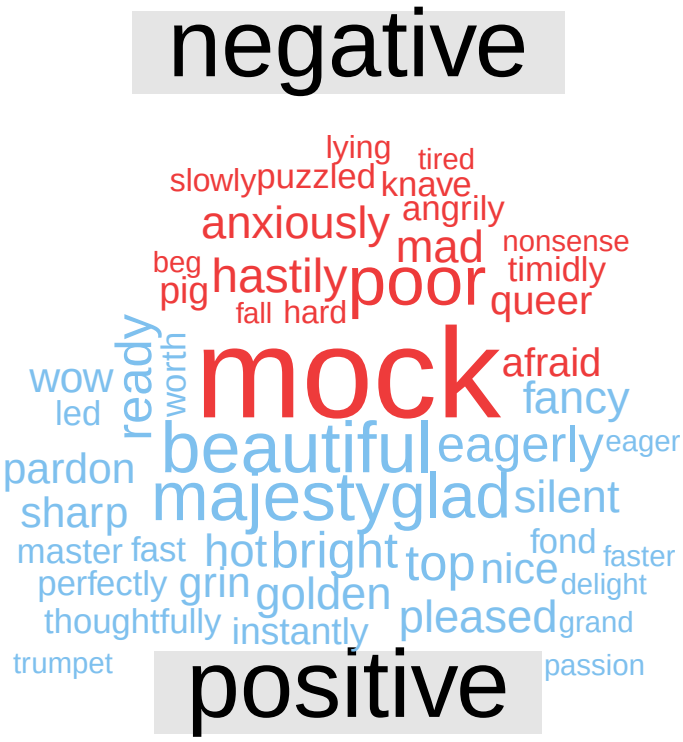

Figure 6: sentiment word cloud version 2

Extra credit: another lexicon

Coincidentally, I found a new method "loughran" when the system reported an error, which said "Error in match.arg(lexicon) : 'arg' should be one of "bing", "afinn", "loughran", "nrc"".

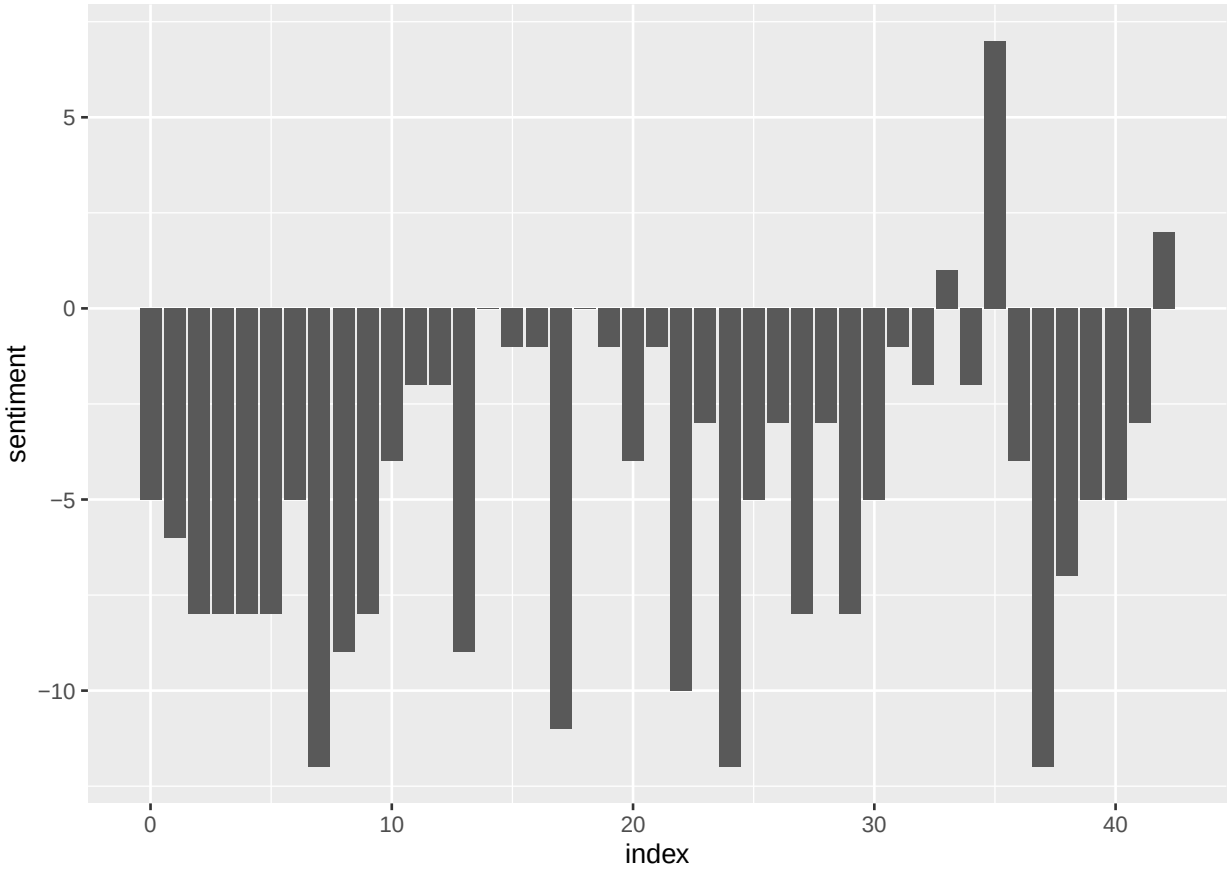

Figure 7: sentiment scores using loughran

Figure 7 shows the sentiment score along the story line using loughran dictionary. The results are almost all negative emotions, consistent with the results of other dictionaries.

Comparing the four sentiment dictionaries

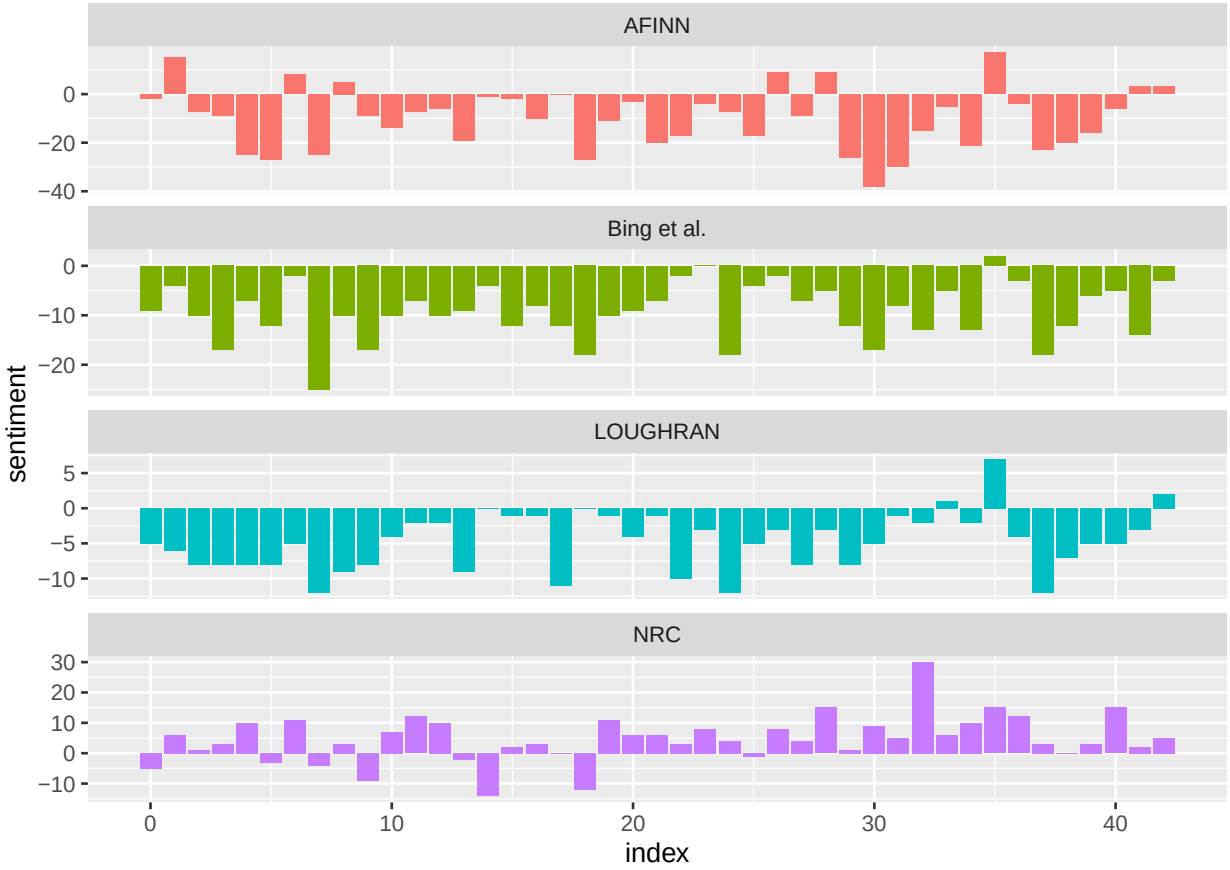

Figure 8: comparing the four sentiment dictionaries

Finally, I drew a comparison chart of the four methods. I found that loughran is like a neutralized version of AFINN and Bing. Maybe this is a more reliable sentiment analysis package.

Reference

[https://www.gutenberg.org/ebooks/11] [https://www.tidytextmining.com/sentiment.html] [https://www.rdocumentation.org/packages/textdata/versions/0.4.1] [https://www.rdocumentation.org/packages/textdata/versions/0.4.1/topics/lexicon_loughran]