MA615 Assignment4 Text Analysis

Kosuke Sasaki

2021/12/7

I. Bag of Words Analysis

Sentiment analysis based on AFINN, BIN, and NRC scale

I have chosen "The Jungle Book" as my book from the gutenberg ebooks. Then, three different lexicons, "Afinn", "Bing", and "Nrc" are used to calculate sentiment of the books. I counted up how many positive and negative words there are in defined sections of each book. I define an index to keep track of the book, and this index counts up sections of 80 lines of text. An estimate of the net sentiment (positive - negative) in each chunk of the book for each sentiment lexicon are shown in the Figure 1.

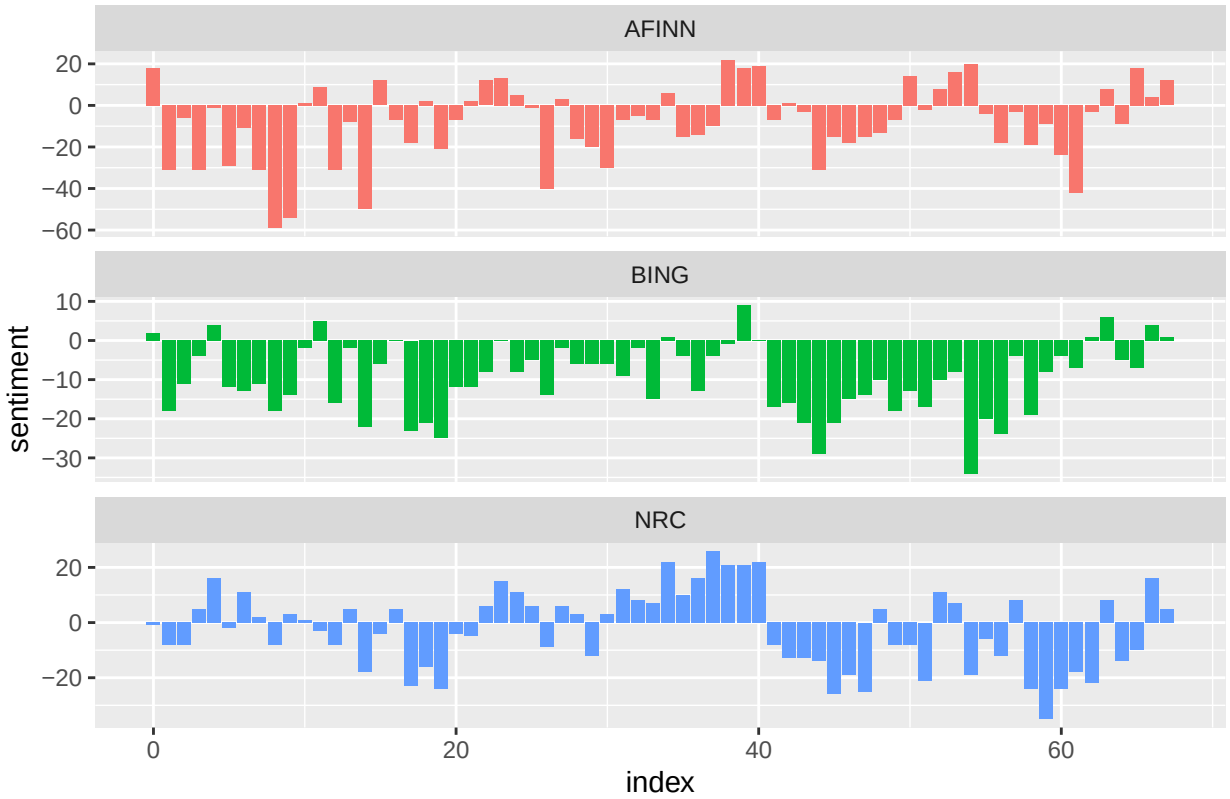

Fig.1 Comparison of three sentiment lexicons

The three different lexicons seem to produce results with some similar trajectories; we have more negative values than positive values throughout the book. Still, we have also some differences between them. The AFINN lexicon gives the largest absolute values, with high negative values. The Bing lexicon has very few positive values. The NRC lexicon seems to label the text more positively relative to the other two. These differences between the methods are seen when looking at other books; the NRC sentiment is high, the

AFINN sentiment has more variance, the Bing et al. sentiment appears to find longer stretches of similar text.[1]

Words count based on each sentiment scale

Then I compare the word counts which contribute to sentiment.

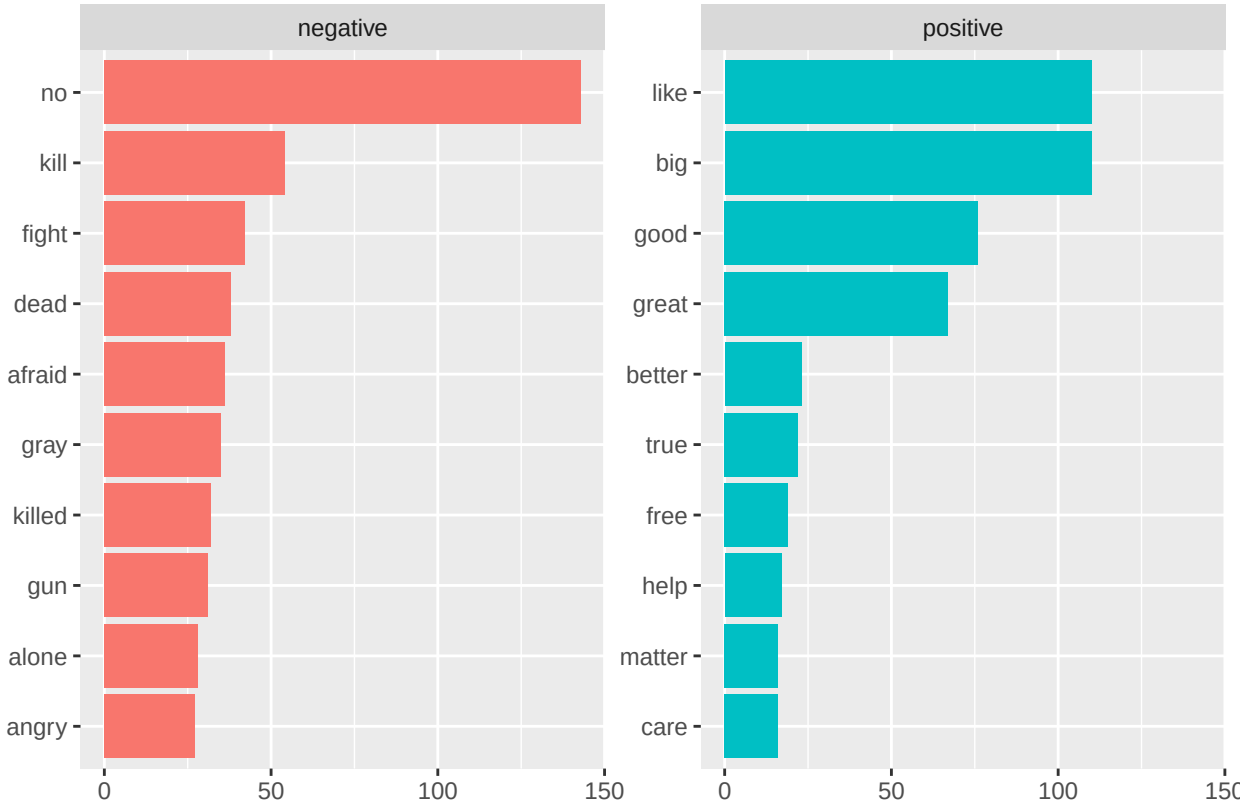

Fig.2−1 Contribution to sentiment (AFINN)

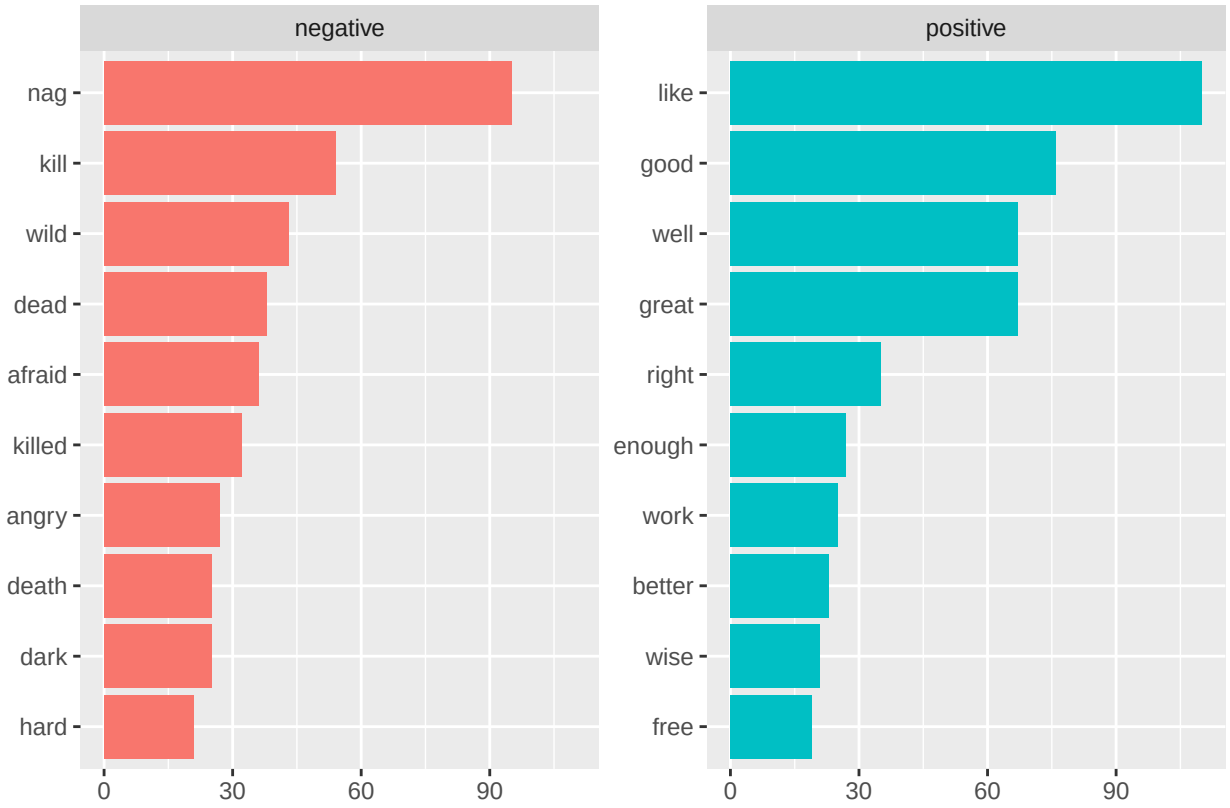

Fig.2−2 Contribution to sentiment (BING)

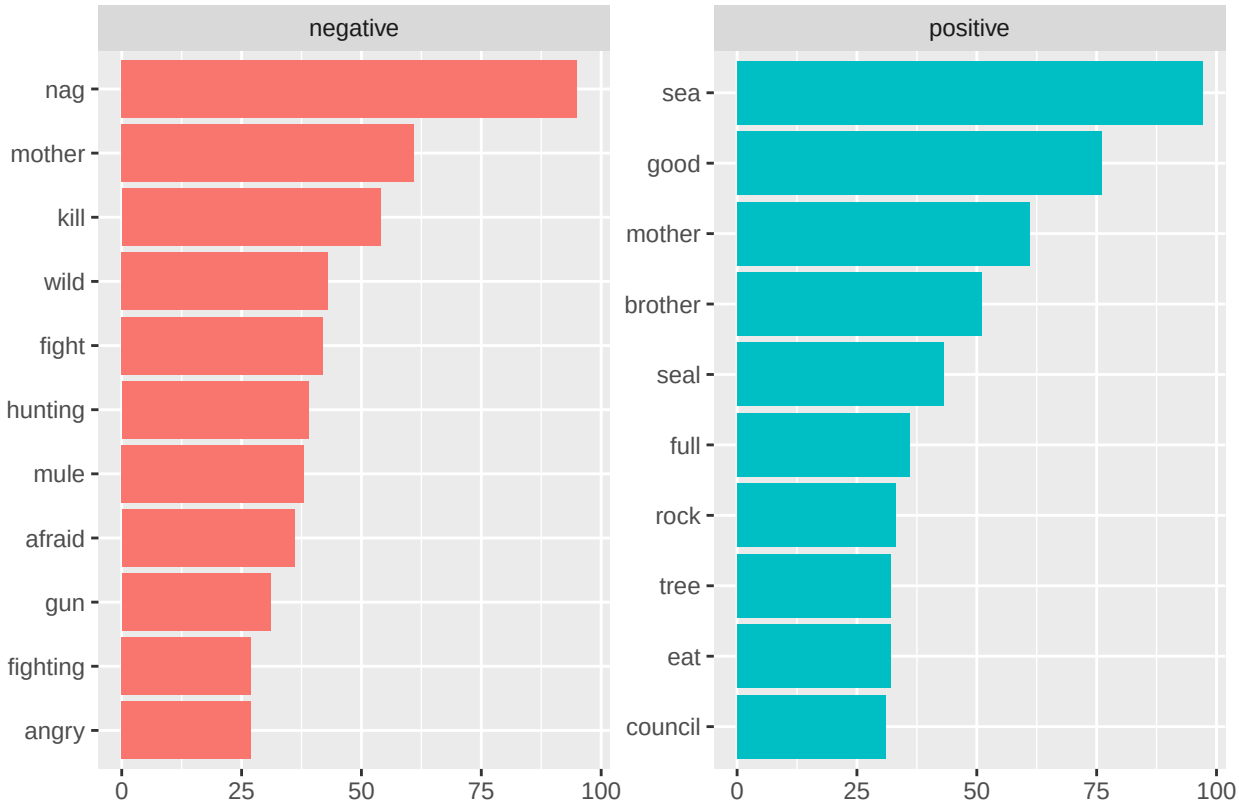

Fig.2−3 Contribution to sentiment (NRC)

As we can see from the figure2-1 to figure2-3, there are several words (kill, dead, like, good, so on so forth) common to the lexicons but the ratio of positive and negative words are different. In addition, the most frequent negative word “nag” on NRC scale is a name of a character and could lead to misleading result of sentiment analysis. Based on these results, there is a chance to choose a lexicon which does not match the word choices of this jungle book so that I should carefully choose the appropriate lexicon in order to draw figure 1 trajectory which best describes the outline of the book.

Additional lexicon analysis

In addition to the lexicon of Afinn, Bing, and Nrc, I tried to apply nrc_eil lexicon to sentiment analysis as below.

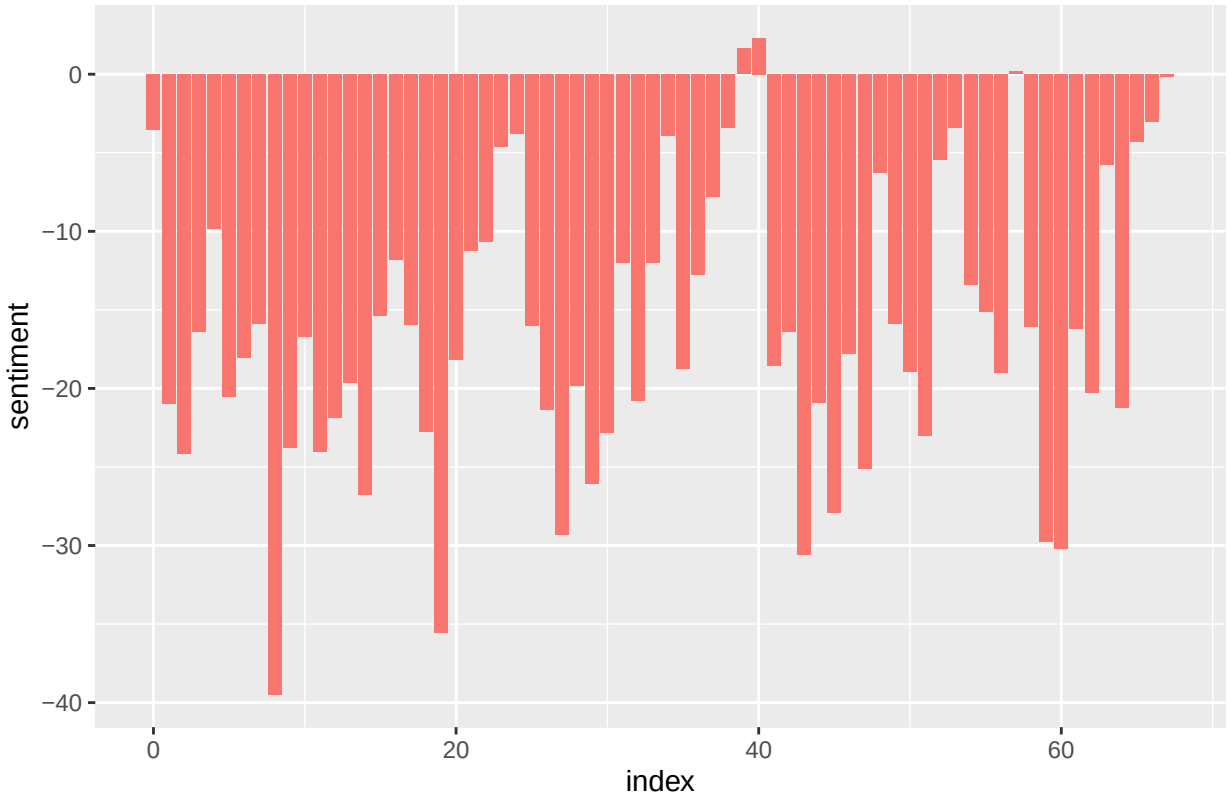

Fig.3−1 NRC_EIL trajectory

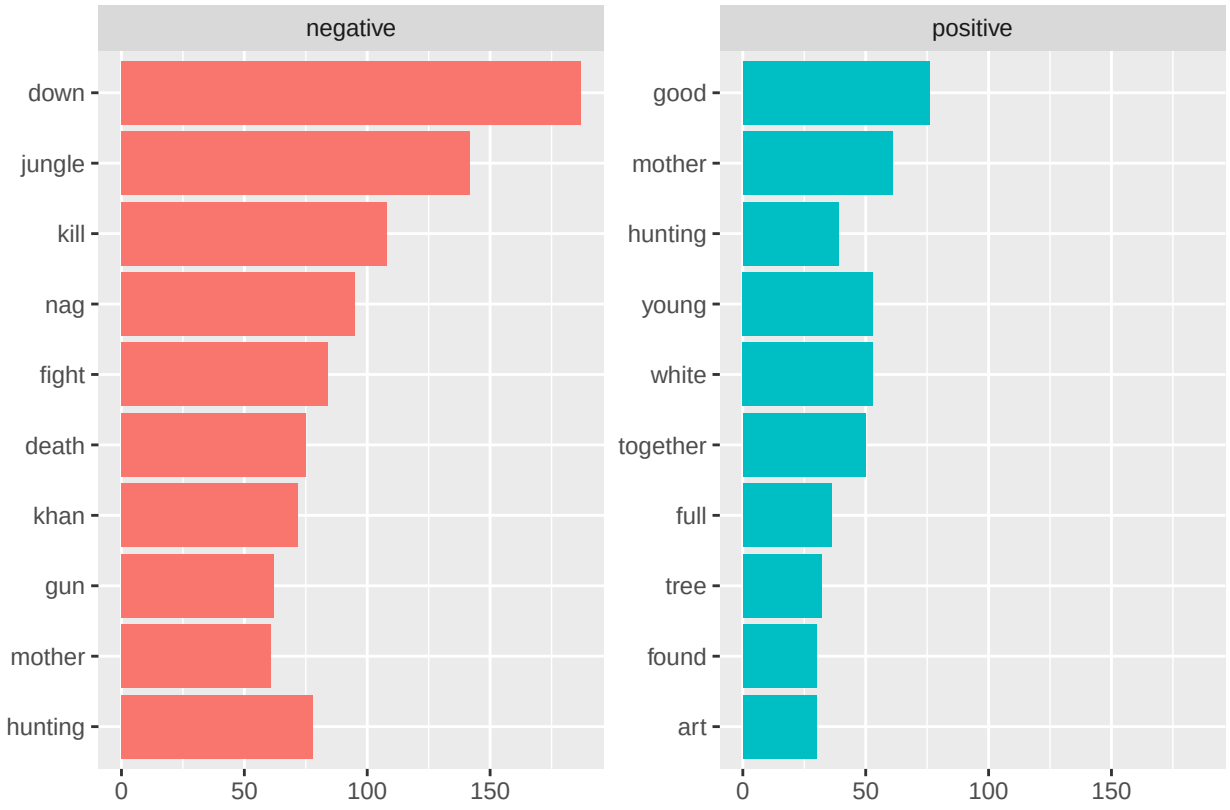

Fig.3−2 Contribution to sentiment (NRC_EIL)

As you can see Figure3-1 and 3-2, nrc_eil lexicon shows much more negative words than positive words relative to other lexicons . This is because nrc_eil lexicon categorize words into 4 types, “joy”, “anger”, “fear”, and “sadness”, and I assume “joy” is the positive category and the others as negative category, which leads to more negative sentiment values assigned to the book.

##Comparison of visualization of lexicons and plotline The plotline of this book is basically not so happy because the main character “Mowgli” has experienced a lot of hardship, even though he and his friends sometimes experienced success, throughout the story. In that sense, NRC lexicon would not be appropriate because it shows rather a positive trajectory, and it also includes the negative word “nag”, which is actually used as a character name in the book. The trajectories of Afinn and Bing lexicon seem to be aligned with the storyline but Bing one shows too many negative sentiment values considering there is still some success. So I would choose Afinn lexicon as the most appropriate one for this book.

II. Paragraph-level of Analysis

In this section, I will do sentiment analysis using “truenumber” and “sentimentr” as below.

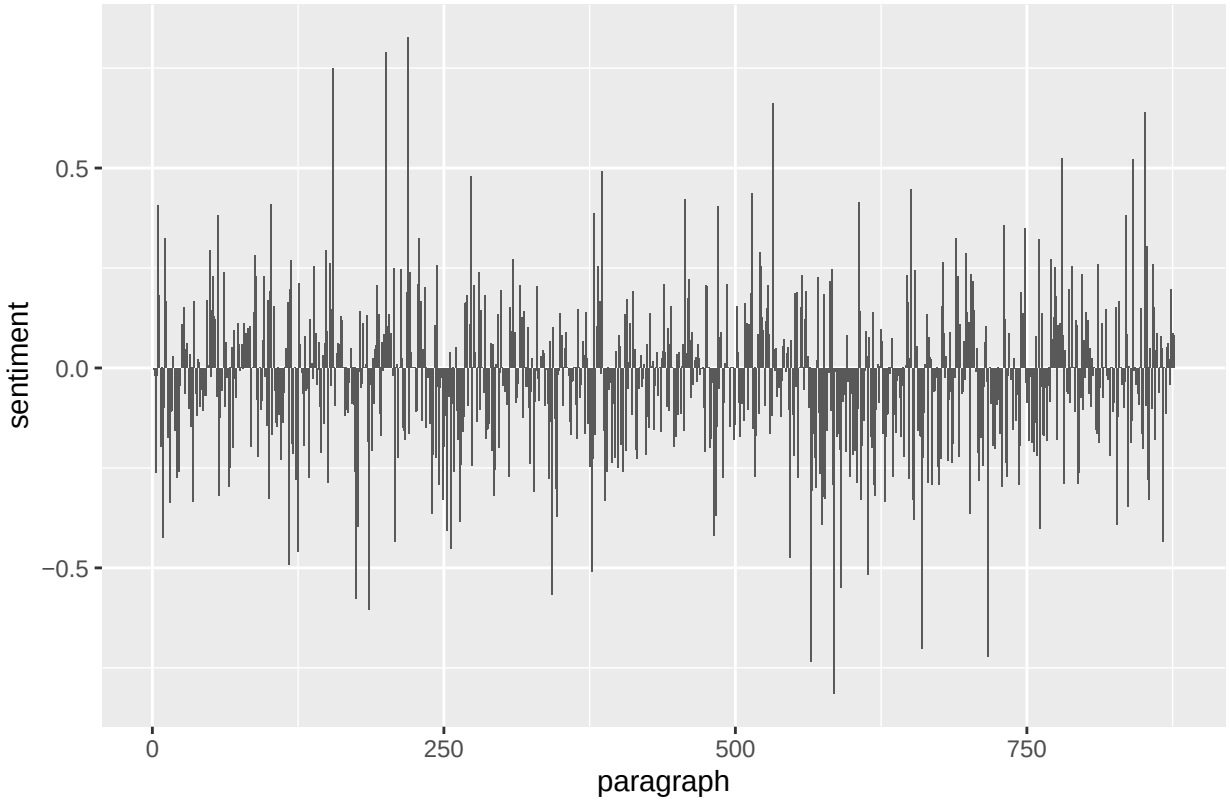

Fig.4 Sentimentr trajectory by paragraph

For this analysis, I calculated sentiment value for each paragraph, while I used keep track of the book for each section of 80 lines of text in the "I. Bag of Words Analysis" section. In total, there are 876 paragraphs and the trajectory of sentiment value by sentimentr analysis is shown above as Figure4.

comparison between sentimentr and lexicons as Paragraph-level Analysis

I will compare the trajectories between sentimentr and lexicons as paragraph-level analysis. To do that, I re-do the bag of words analysis for each lexicon as below.

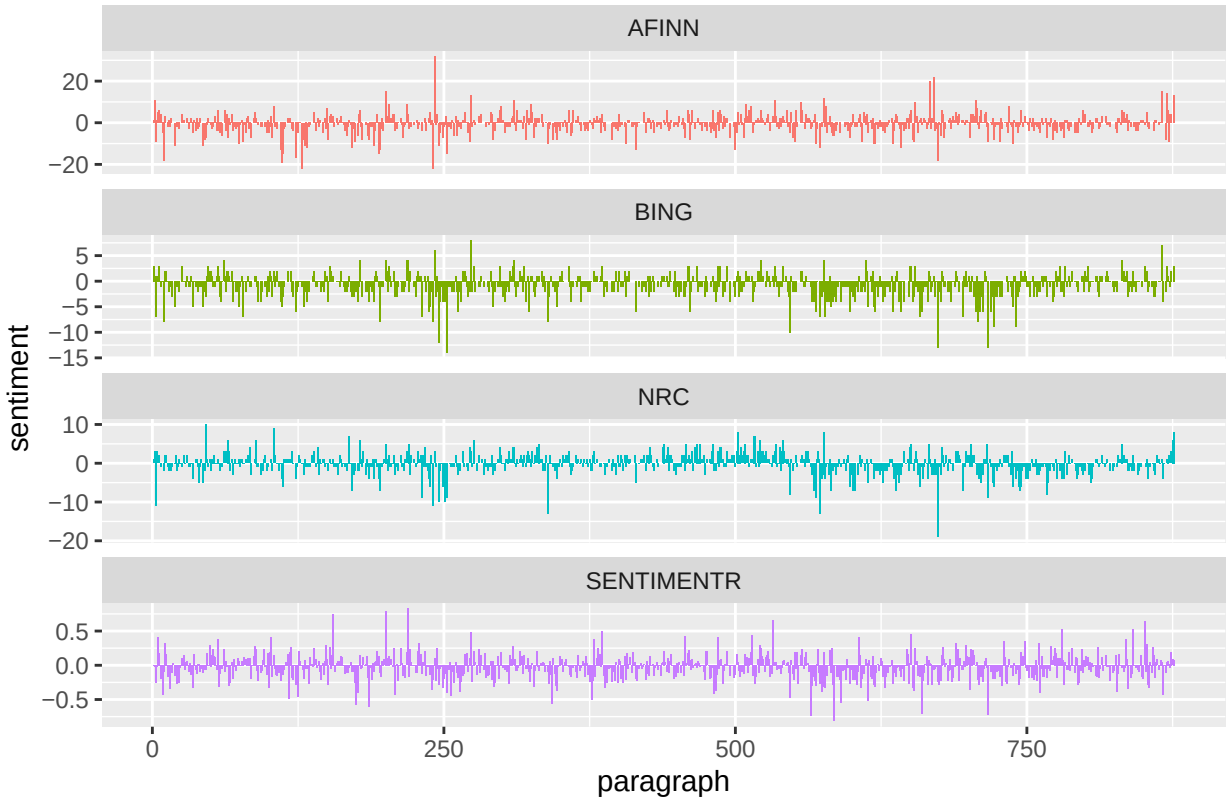

Fig.5 Comparison of lexicons and sentimentr by paragraph

Based on the figure5 above, the trajectory calculated by sentimentr for each graph looks similar to the ones by three lexicons. Even though, when you take a closer look, NRC trajectory shows more positive values relative to the other trajectories, which we could see in the comparison of three lexicons in the first section of this report. Compared to Afinn and Bing trajectories, sentimentr one is small in absolute value but the shape is very similar to those two trajectories, just as in-between.
Given these graphs, I could say Afinn, Bing and sentimentr are all appropriate for the sentiment analysis for this book.

III. Character Analysis

Finally, I will calculate sentiment value related to “Mowgli”, the main character of this book, and “Shere Khan”, Mowgli’s opponent.

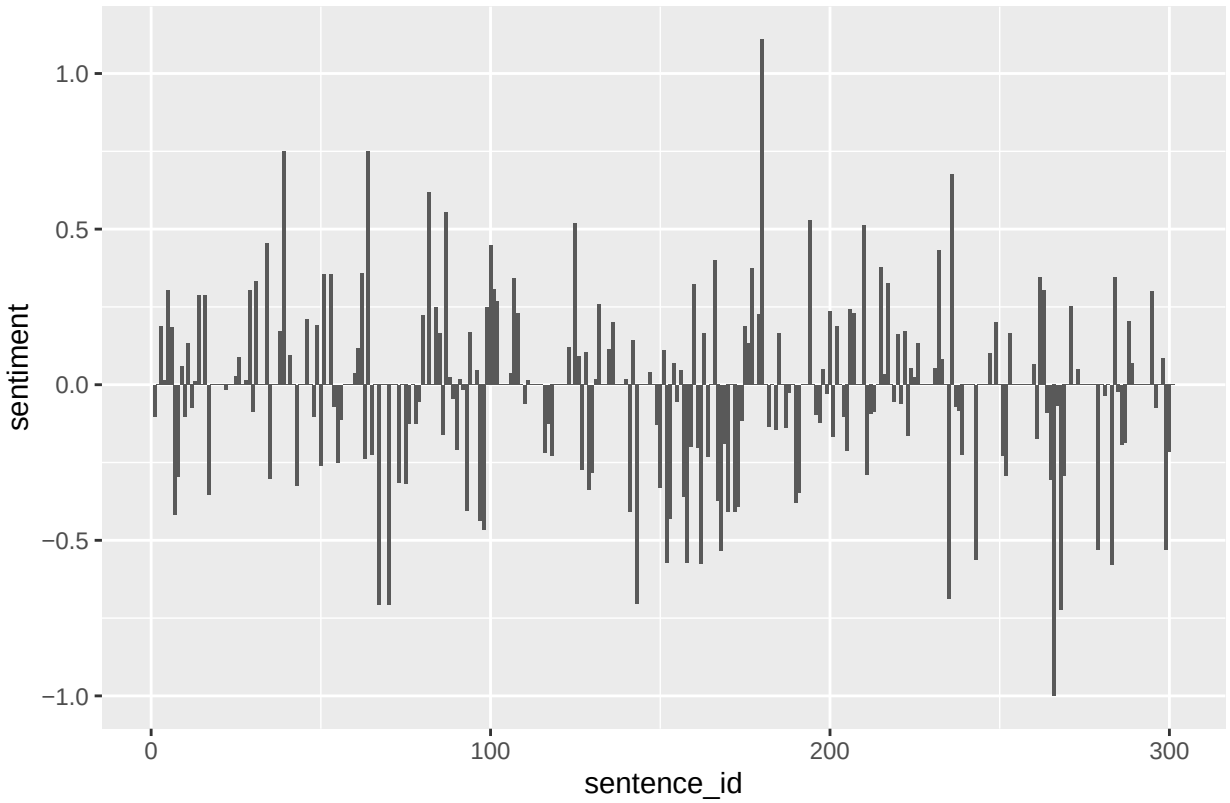

Fig.6−1 Mowgli sentiment trajectory by paragraph

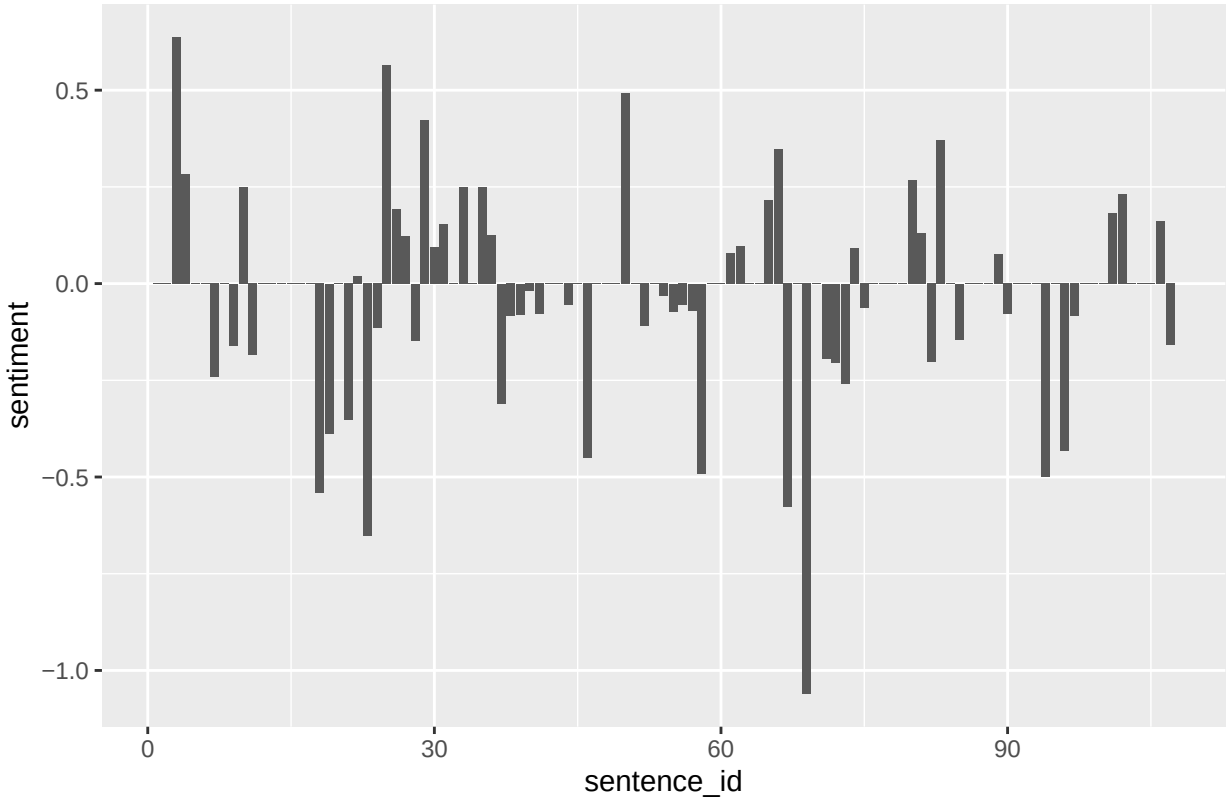

Fig.6−2 Shere Khan sentiment trajectory by paragraph

```
##    element_id word_count        sd ave_sentiment
## 1:          1       4882 0.2522549   -0.01984604
```

```
##    element_id word_count        sd ave_sentiment
## 1:          1       1487 0.2352948    -0.0403314
```

As you can see from Figure6-1 and 6-2, the sentences related to Mowgli are much more than the ones related to Shere Khan. It is reasonable because Mowgli is the main character whereas Shere Khan is his opponent and the number of appearance in the story is much less than that of Mowgli. It is also reasonable that trajectory of Shere Khan seems to be slightly more negative than the one of Mowgli, and the average of sentiment value of Shere Khan based on sentimentr is -0.04, which is less than that of Mowgli, -0.02, which you can see in the above tables.
In conclusion, sentiment analylsis based on sentimentr function works really well for this book.

Reference

1. Julia, S. (2021) *Text Mining with R: A Tidy Approach*[online]. O'Reilly: https://www.tidytextmining.com/index.html [accessed 7 December 2021]