# Text analysis

Yuli Jin

2021/11/24

## Task 1 Pick a book

I choose `The Burning Secret` as the my text analysis. This book was written by Zweug, Stefan.

## TASK 2 bag of word analysis

First, I use three types of sentiment analysis methods AFINN, Bing and NRC to plot barplot to compare these methods. From the graph below, the AFINN and Bing method fits better. Most of the polt in `The Burning Secret` is in negative. In this book, While being treated for asthma at a country spa, an American diplomat's lonely 12-year-old son is befriended and infatuated by a suave, mysterious baron. But soon his adored friend heartlessly brushes him aside and turns his seductive attentions to his mother. The boy's jealousy and feelings of betrayal become uncontrollable. The story is set in Austria in the 1920s. That is to say, at the beginning of the book, the sentiment of the book is positive, but soon it converts into negative sentiment. However, it is difficult to identify which of the two methods is better. In the following task, I use Bing method to conduct further analysis.

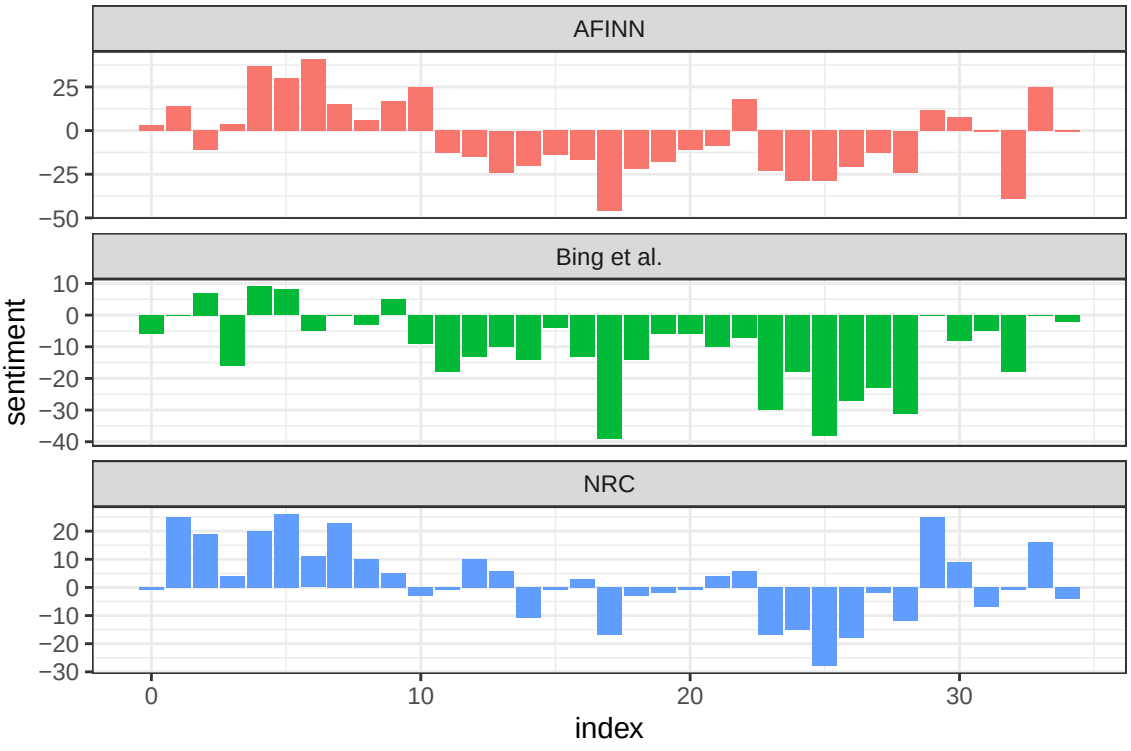



Figure 1: sentiment plot

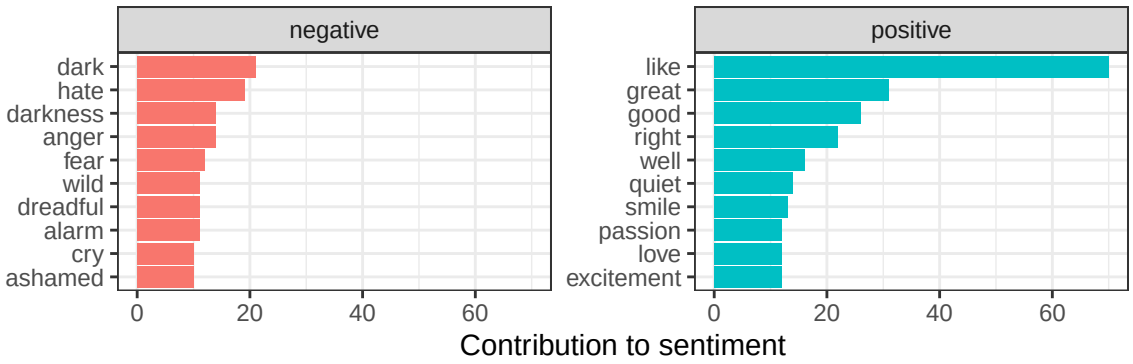



Figure 2: negative positive words count

Figure 2 shows negative and positive word count of each word. For the negative chart, dark is the most common words throughout the whole book. Hate and darkness rank the second and third place respectively. For the positive chart, like is the most common words throughout the whole book. Great and good rank the second and third place respectively.

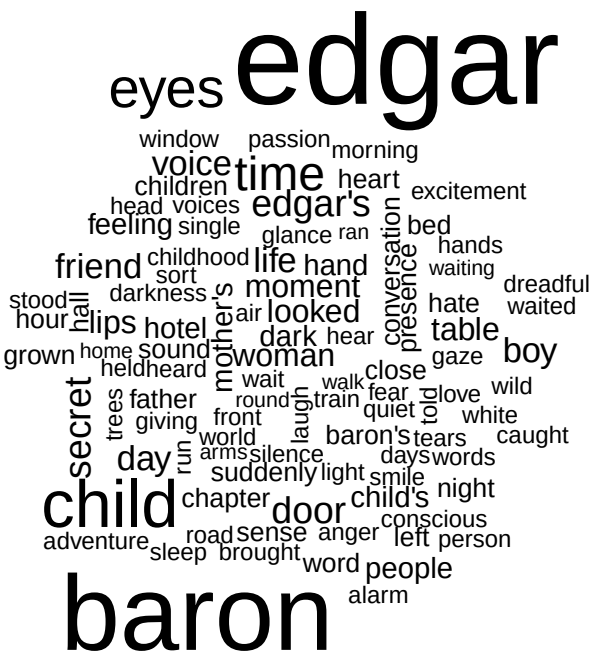



Figure 3: word cloud

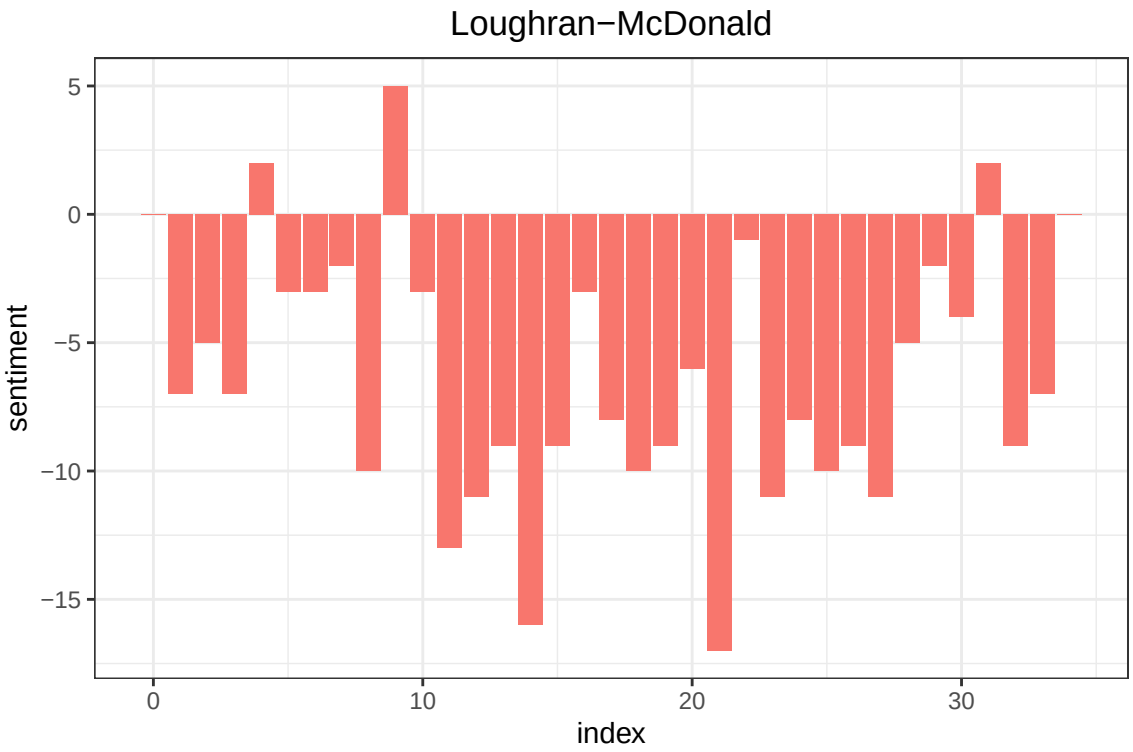
Figure 3 displays word cloud which shows the frequency. As we can see, baron, mother, edgar are the most frequency words among all the words. It is reasonable because they are the main characters in that fiction book. In task 3, I will use two of three characters to conduct further analysis.

Figure 4 generally converts Figure 2's information into word cloud


Figure 4: sentiment word cloud

### Task 2 extra credit

In `textdata` package, there are one extra lexicons available to use. This lexicons is called `Loughran-McDonald`. Here I use this new method and plot the similar graph to show the progression from start to finish of the book.




Then I provide some description. According to the introduction of https://emilhvitfeldt.github.io/textdata/reference/lexicon_loughran.html, this lexicon is created for use with financial documents. Therefore, the graph is complete different from previous lexicon graphs. Therefore, this lexicon cannot correctly reflect the exact sentiment of the book. After all, financial sentiment lexicon isn't necessarily suitable for fiction book.

## task 3 sentence-level analysis

### tnum

First, I put the book into tnum, the following table shows evidence of my tnum database.

```
# query heading text to display the head
q24<- tnum.query('zweig/test2/heading# has text',max=90)
df24 <- tnum.objectsToDf(q24) # turn the object to df
knitr::kable(df24 %>% select(subject:numeric.value)%>% head())
```

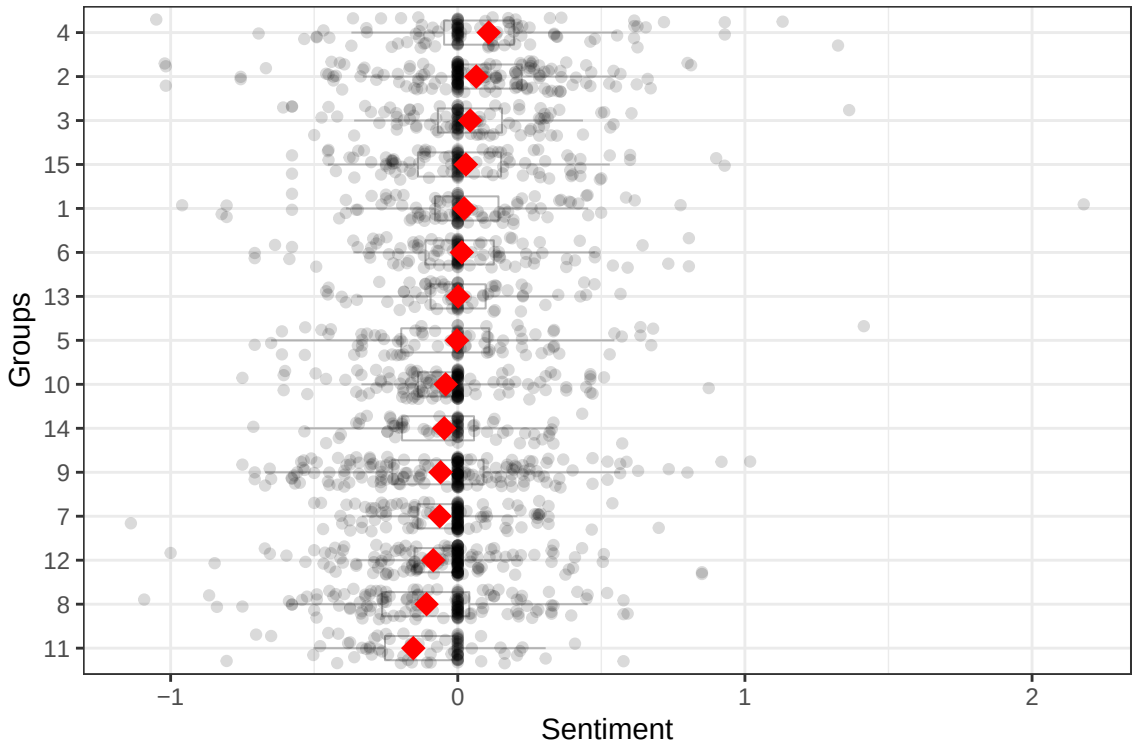
| subject | property | string.value | numeric.value |
|---|---|---|---|
| Zweig/test2/heading:0001 | text | “” | NA |
| Zweig/test2/heading:0002 | text | “” | NA |
| Zweig/test2/heading:0003 | text | “” | NA |
| Zweig/test2/heading:0004 | text | “” | NA |
| Zweig/test2/heading:0005 | text | “” | NA |
| Zweig/test2/heading:0006 | text | “” | NA |

```
# query section and heading text to display the head
q26<- tnum.query('zweig/test2# has text',max=60)
df26 <- tnum.objectsToDf(q26) # turn the object to df
df26 %>% select(subject:string.value)%>% head()
```

```
##                                          subject property
```

```
## 1                                Zweig/test2/heading:0001     text
## 2 zweig/test2/section:0001/paragraph:0001/sentence:0001     text
## 3 zweig/test2/section:0001/paragraph:0001/sentence:0002     text
## 4 zweig/test2/section:0001/paragraph:0001/sentence:0003     text
## 5 zweig/test2/section:0001/paragraph:0002/sentence:0001     text
## 6 zweig/test2/section:0001/paragraph:0002/sentence:0002     text
##
## 1
## 2
## 3
## 4 "Exacerbated voices called back and forth; then, with a puffing and a chugging and another shrill s
## 5
## 6
```

Then I use sentimentr to get sentiment score group by these scores with section to get the average result. The plot sort the average sentiment score from high to low.




**Compare this analysis with the analysis you did in Task TWO**

It is difficult to directly compare Sentimentr and Bing's score. Therefore, I apply `scale` function to keep two variable into the same criteria. Then I use ggplot to plot bar plot. From the Figure below, we can see that the trends, say positive and negetive direction, are mainly similar. But the exact number differs from two methods. It is difficult to identify which side is more optimistic. However, in some sections, say section 1,3,7,9,11,13,14, bing method is more optimistic than sentimentr method. When it comes to other sections, sentimentr method is more optimistic than bing method. But generally, these two methods' score have similar positive and negative trends after scaling.

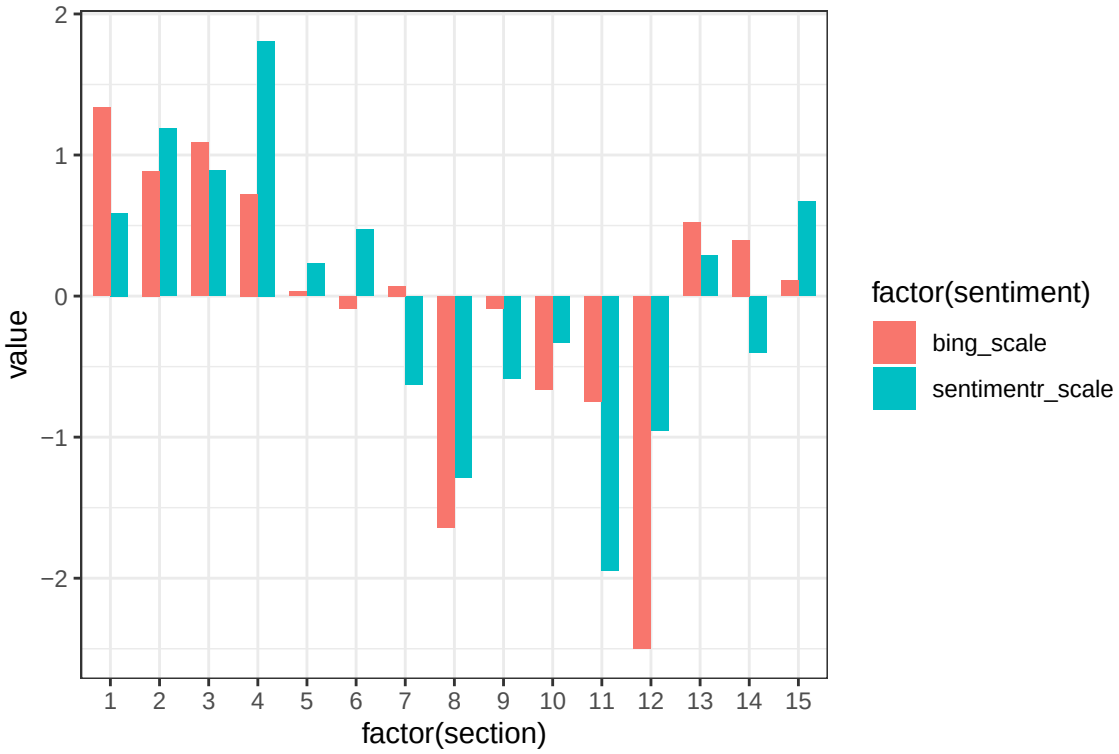



Figure 5: sentiment comparison

**EXTRA CREDIT: character analysis**

Baron and Edger are two main character among the fiction book. I Pick these two characters from my book. The following table in the count number of times each character appears in each chapter:

| chapter | baron | edgar |
|---|---|---|
| 1 | 10 | 2 |
| 2 | 22 | 17 |
| 3 | 18 | 16 |
| 4 | 13 | 12 |
| 5 | 9 | 5 |
| 6 | 16 | 18 |
| 7 | 12 | 13 |
| 8 | 15 | 24 |
| 9 | 13 | 19 |
| 10 | 8 | 14 |
| 11 | 10 | 12 |
| 12 | 5 | 18 |
| 13 | 0 | 11 |
| 14 | 1 | 13 |
| 15 | 1 | 14 |

The following table is the count of number of times both characters appear in the same paragraphs.

| section | paragraph | both_appear |
|---|---|---|
| 2 | 4 | 1 |
| 2 | 28 | 1 |
| 2 | 35 | 1 |
| 2 | 36 | 1 |
| 2 | 40 | 1 |
| 3 | 1 | 1 |
| 3 | 16 | 1 |
| 4 | 3 | 1 |
| 4 | 7 | 1 |
| 4 | 11 | 1 |
| 4 | 12 | 1 |
| 4 | 23 | 1 |
| 4 | 28 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |
| 6 | 3 | 1 |
| 6 | 21 | 1 |
| 7 | 5 | 1 |
| 7 | 7 | 1 |
| 7 | 8 | 1 |
| 7 | 9 | 1 |
| 8 | 1 | 1 |
| 8 | 6 | 1 |
| 8 | 11 | 1 |
| 8 | 29 | 1 |
| 8 | 31 | 1 |
| 8 | 35 | 1 |
| 9 | 2 | 1 |
| 9 | 21 | 1 |
| 9 | 32 | 1 |
| 9 | 41 | 1 |
| 10 | 11 | 1 |
| 11 | 5 | 1 |
| 11 | 16 | 2 |

## Reference

https://www.gutenberg.org/ebooks/45755
https://emilhvitfeldt.github.io/textdata/reference/lexicon_loughran.html