Camille Text Mining - Task 1 & 2

Wuji Shan

12/8/2021

Book: Camille (LA DAME AUX CAMILIAS)
Author: Alexandre Dumas

I. Task One

For Task One, from The Gutenberg Project, I picked the novel Camille (LA DAME AUX CAMILIAS) written by Alexandre Dumas. The book size is 390 kB.

Novel Summary:

Camille (LA DAME AUX CAMILIAS) is a novel and masterpiece written by the French writer Alexandre Dumas. The story tells a tortuous and sad love story between a young man Armand and a social star Marguerite in Parisian high society. Marguerite was a poor girl in the countryside; after she came to Paris, she became a famous social star adored by nobles. A young man Armand fell in love with Marguerite; his love over two years impressed Marguerite and they traveled to the suburbs living a life with respect and love. However, when Armand's father knew this situation, he asked Armand to return Paris and Marguerite to leave his son. Considering Armand's future, Marguerite agreed and chose to leave him. Finally, Marguerite passed away due to illness, and after her death, Armand just knew the truth why she left him and her respectable heart.

II. Task Two: Sentiment Analysis

For Task Two - Sentiment Analysis, I chose three sentiment lexicons AFINN, BING, and NRC to compare their sentiment analyses and graphs corresponding to the narrative of my book. Five figures are shown in this part, illustrating distribution of words related to sentiment in book Camille from several angles. The conclusion got is that BING lexicon is the best fit between the plotline of the book.

Figure 1:

It's a plot showing sentiment plot of BING lexicon. From Figure 1, it is obvious the first ten chapters and final ten chapters show the negative pattern of sentiment, but the mid ten chapter show the positive sentiment pattern, following the overall development of the novel storyline, which will be explained in detail in Figure 2 description.

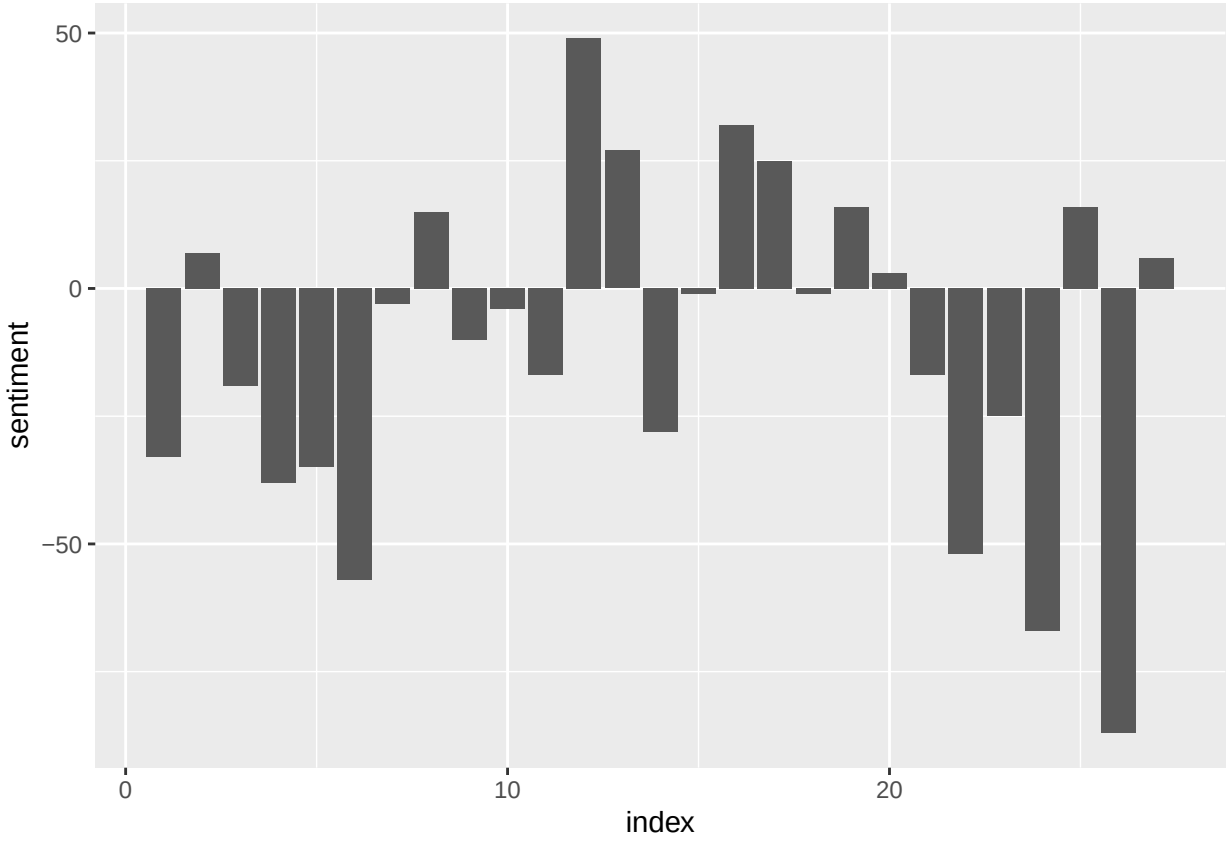


Figure 2:

It's a plot binding estimate for each sentiment lexicons (nrc, bing, afinn) together. From Figure 2, it is obvious that the pattern of the sentiment change is corresponding to the development of the novel narrative. At the beginning, the sentiment's bias is negative because Marguerite was from countryside and her original family was very poor. During the mid-development of the story, Marguerite and Armand traveled together living a life with respect and love, so the sentiment now is positive. At the end part, this couple were forced to forced to leave each other with misunderstanding, leading to the negative sentiment trend. Although NRC did very well in the positive mid part, considering the whole pattern change of the story, BING is the best fit between the plotline of the book.

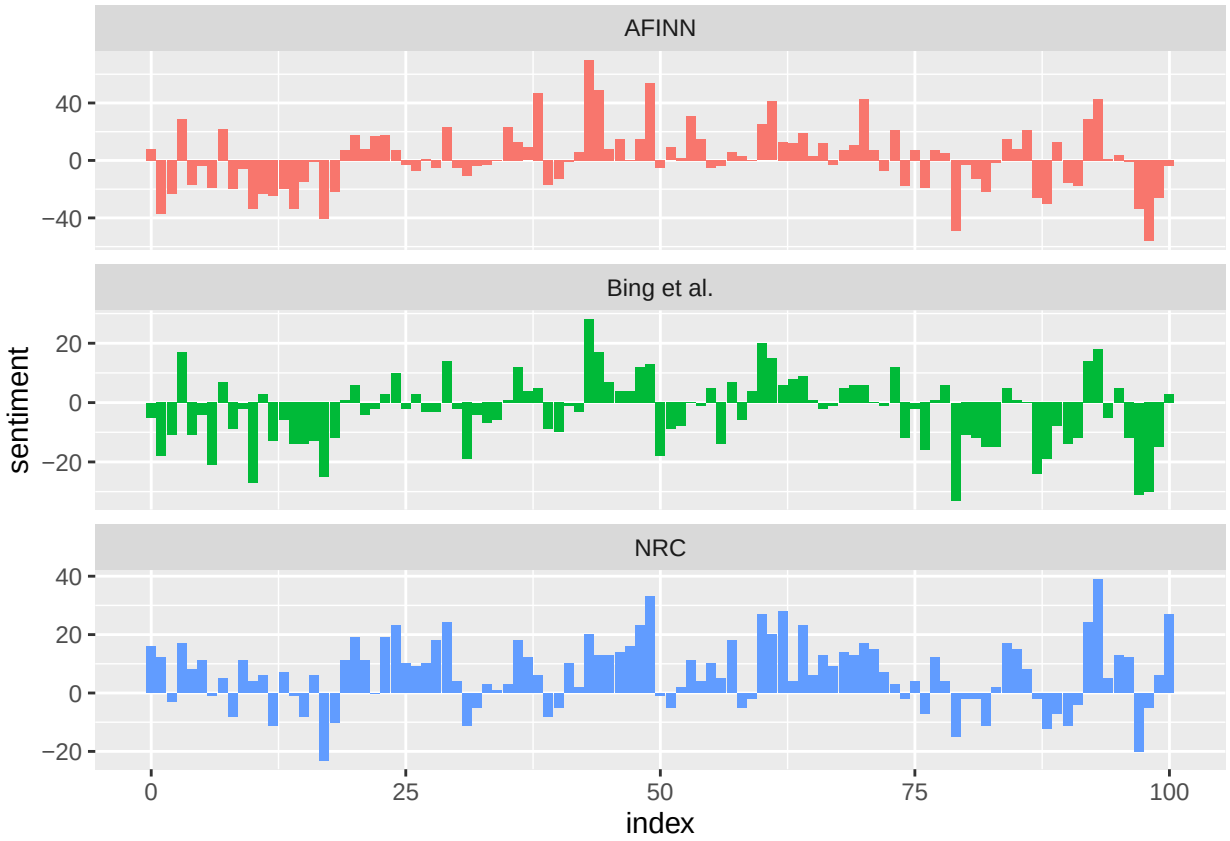


Figure 3:

It's a plot showing the number of words contributed to each sentiment (positive and negative). Figure 3 shows only the top ten frequent positive and negative sentiment words in the book. We can observe that positive words contributes more to sentiment than negative words. Additionally, in the book Camille, the most frequent negative word is "poor", which appears about 55 times; the most frequent positive word is "love", which appears about 270 times. One interesting finding is that 4 of top ten frequent positive words are all related to love, which reflects more the topic of the novel.

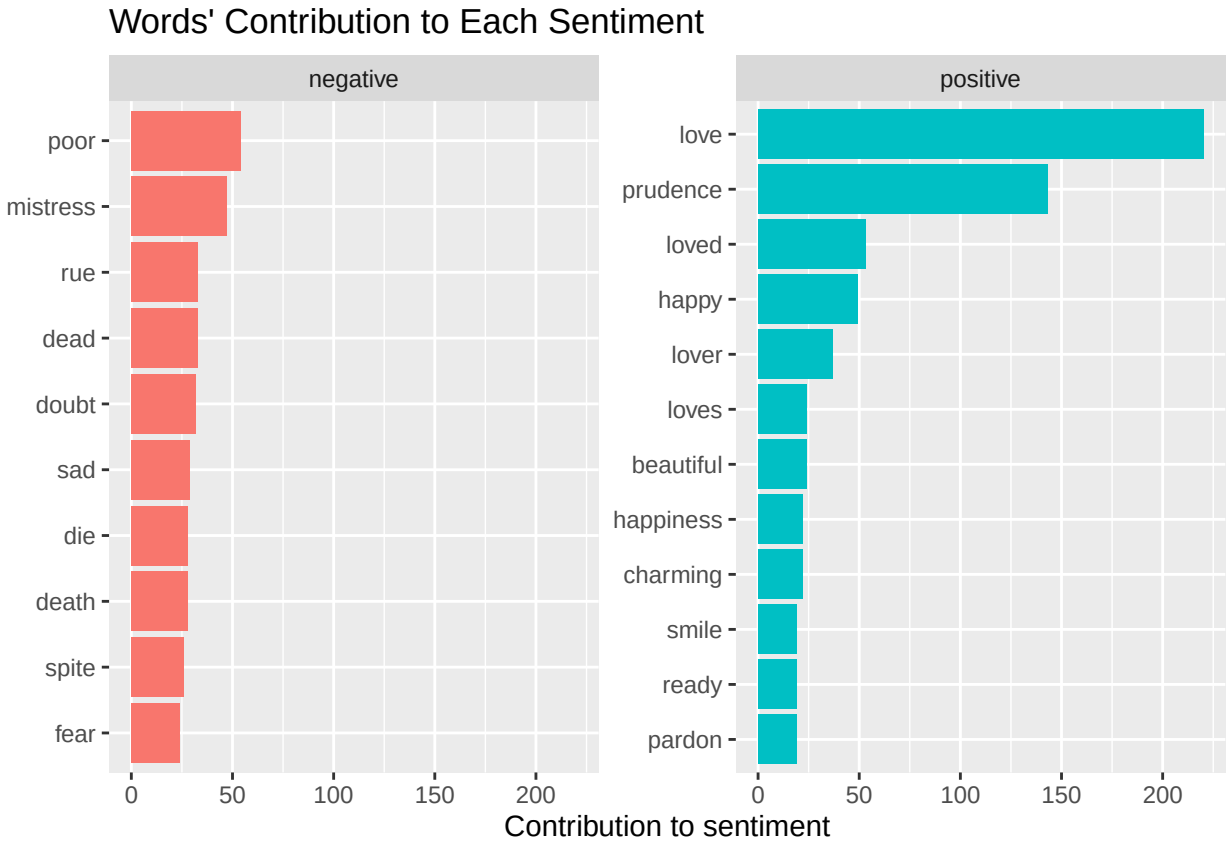


Figure 4:

It's a Word Cloud graph showing the most common words in the book: marguerite (name of the main character), love, prudence, woman, father, time and etc.

```
## Joining, by = "word"
```

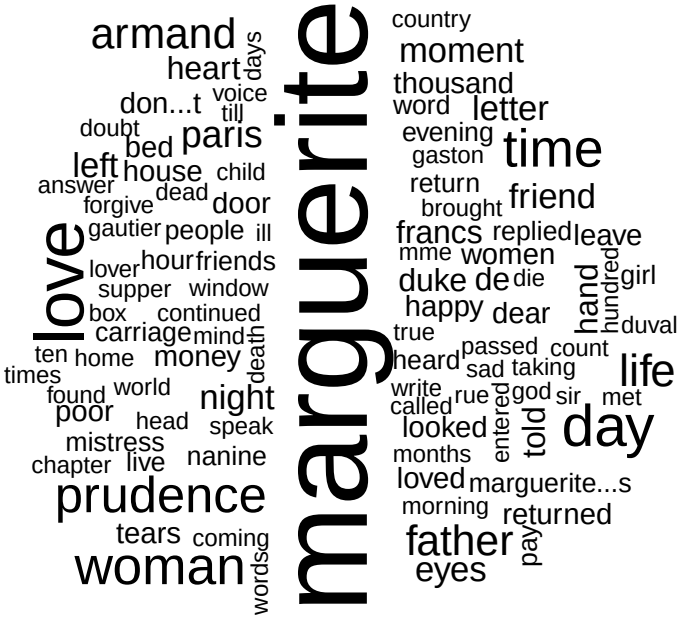


Figure 5:

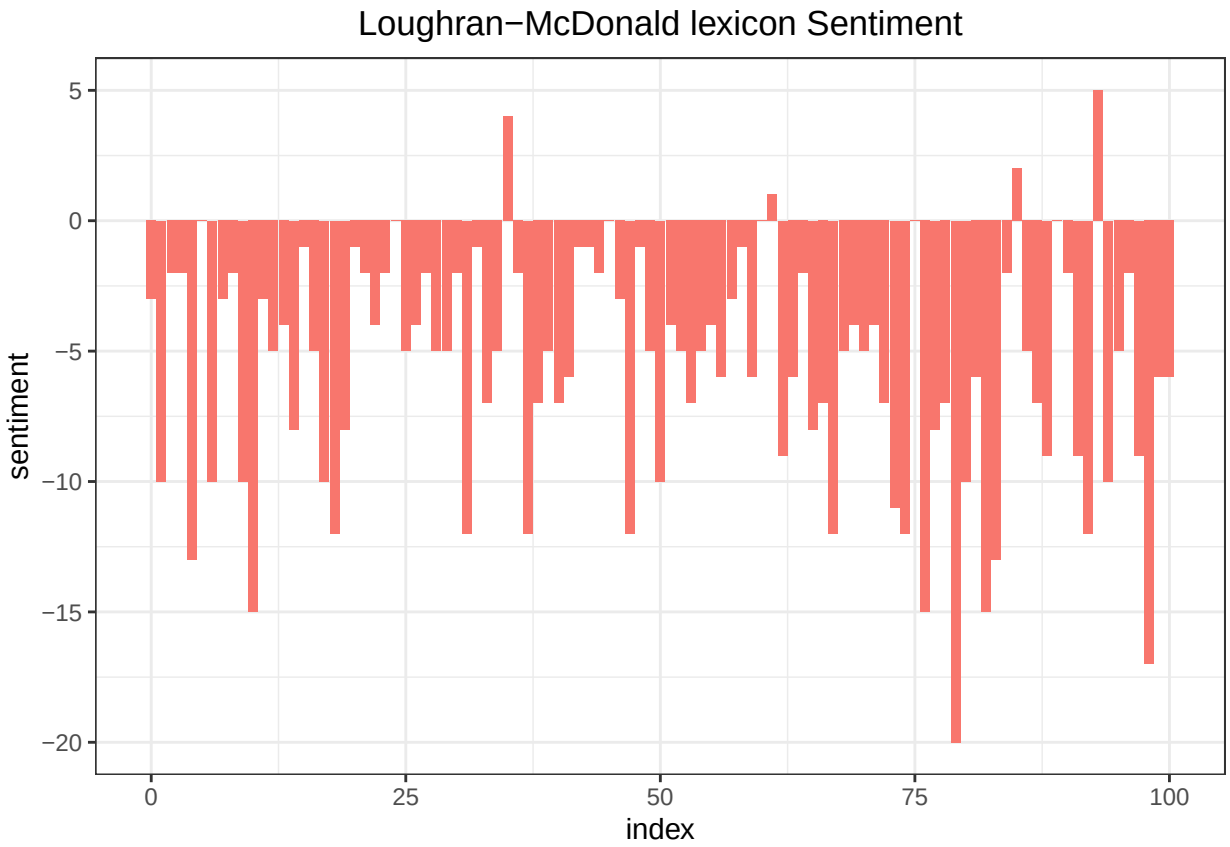
It's a Word Cloud graph showing the most common positive and negative words via making the sentiment analysis to tag positive and negative words. We can observe that positive words - with gray color - are more than negative words, which corresponds to responses of figure 1, 2 and 3 as well.

```
## Joining, by = "word"
```

Task TWO Extra Credit

Here, I decided to use Loughran-McDonald sentiment lexicon in textdata package.

```
## Joining, by = "word"
```



From the sentiment plot shown via using Loughran-McDonald lexicon, we can observe that most of sentiment are all negative, so it does not fit the storyline better than other three lexicons: afinn, bing, and nrc.