Report for MA615 Assignment-text mining

Team 2

11/18/2020

```
library(tnum)
library(tidyverse)
library(stringr)
library(ggplot2)
library(dplyr)
tnum.authorize()
```

Problem 1

Tag the sentences in which Darcy and Elizabeth co-occur. Counting these tags could give you a bar chart that shows how the counts of these sentences are distributed by chapter.

```
#check the database
tax=tnum.getDatabasePhraseList("subject",level=3)
```

```
#tag the sentences include "Darcy" in Pride and Prejudice
tnum.tagByQuery("*pride* has text= REGEXP(\"Darcy\")","reference:group2Darcy" )
```

```
## list(modifiedCount = 394, tagged = 394, removed = 0)
```

```
#tag the sentences include "Elizabeth" in Pride and Prejudice
tnum.tagByQuery("*pride* has text= REGEXP(\"Elizabeth\")","reference:group2Elizabeth" )
```

```
## list(modifiedCount = 610, tagged = 610, removed = 0)
```

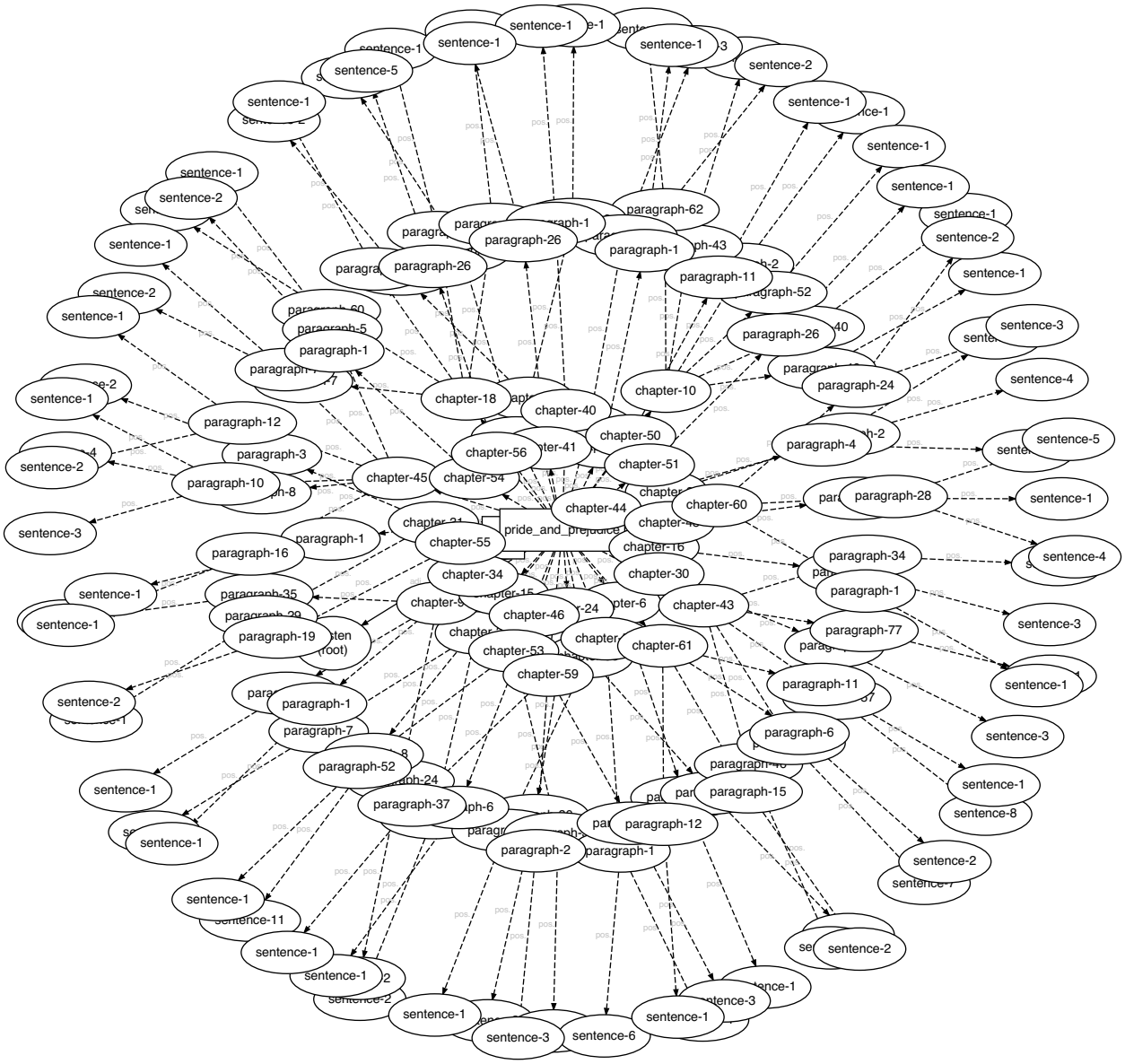
```
#create dataframe with Darcy sentences
num_darcy=tnum.query("*pride* has text= REGEXP(\"Darcy\")",max=500)
```

```
## Returned 1 thru 394 of 394 results
```

```
textdf=tnum.objectsToDf(num_darcy)

#create dataframe, sentences in which Darcy and Elizabeth co-occur
textdf_co_occur=filter(textdf,grepl('group2Elizabeth', tags))
#view(textdf_co_occur)
textdf_co_occur$string.value[1]
```

```
## [1] "\"Elizabeth Bennet had been obliged, by the scarcity of gentlemen, to sit down for two dances; a
```

```
#we can find both tags we created before are in the columns.
#generate a tree-plot for these sentences.
picco_occr=tnum.makePhraseGraphFromPathList(textdf_co_occur$subject)
tnum.plotGraph(picco_occr)
```



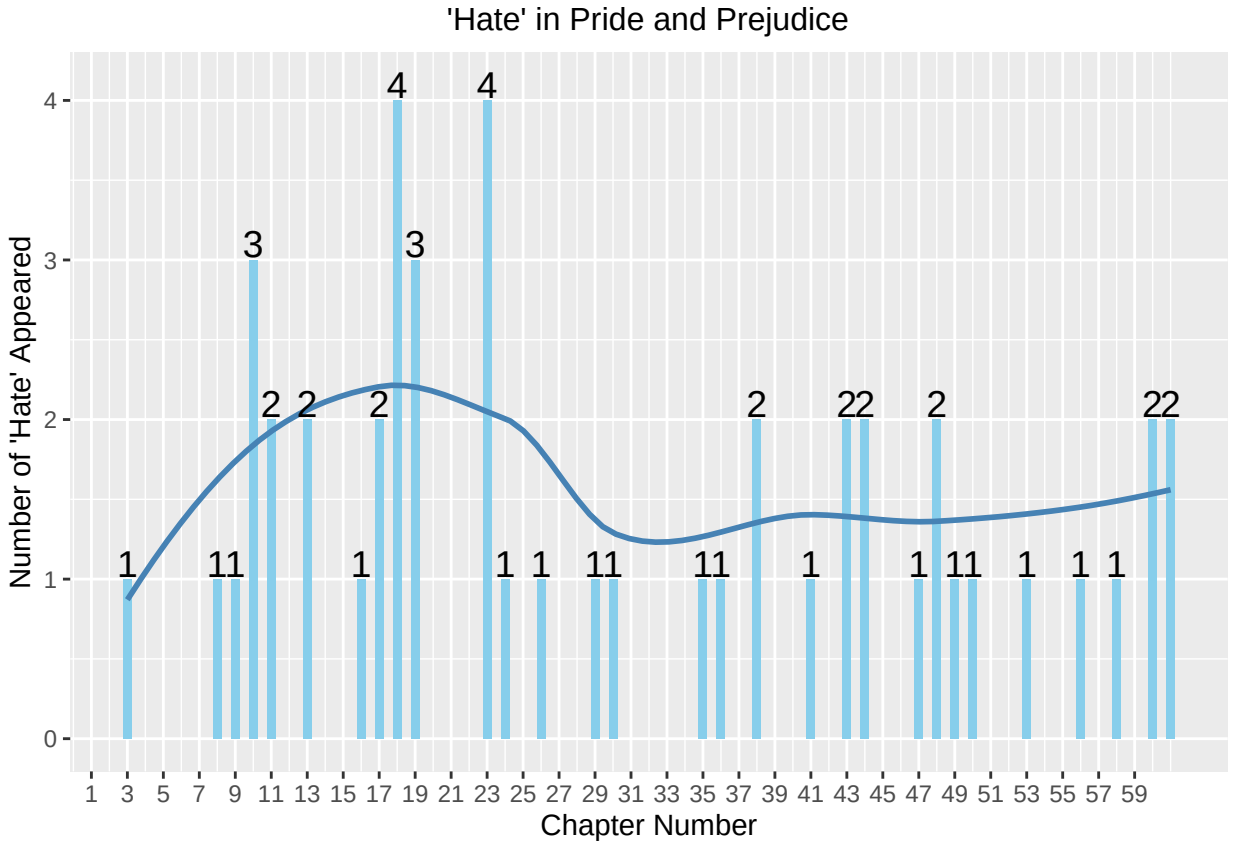
Problem 2

We are interested in the writer's emotion through the period that she finished her novels. And we will show some thoughts about data combining her life.

Pride and Prejudice(1813)

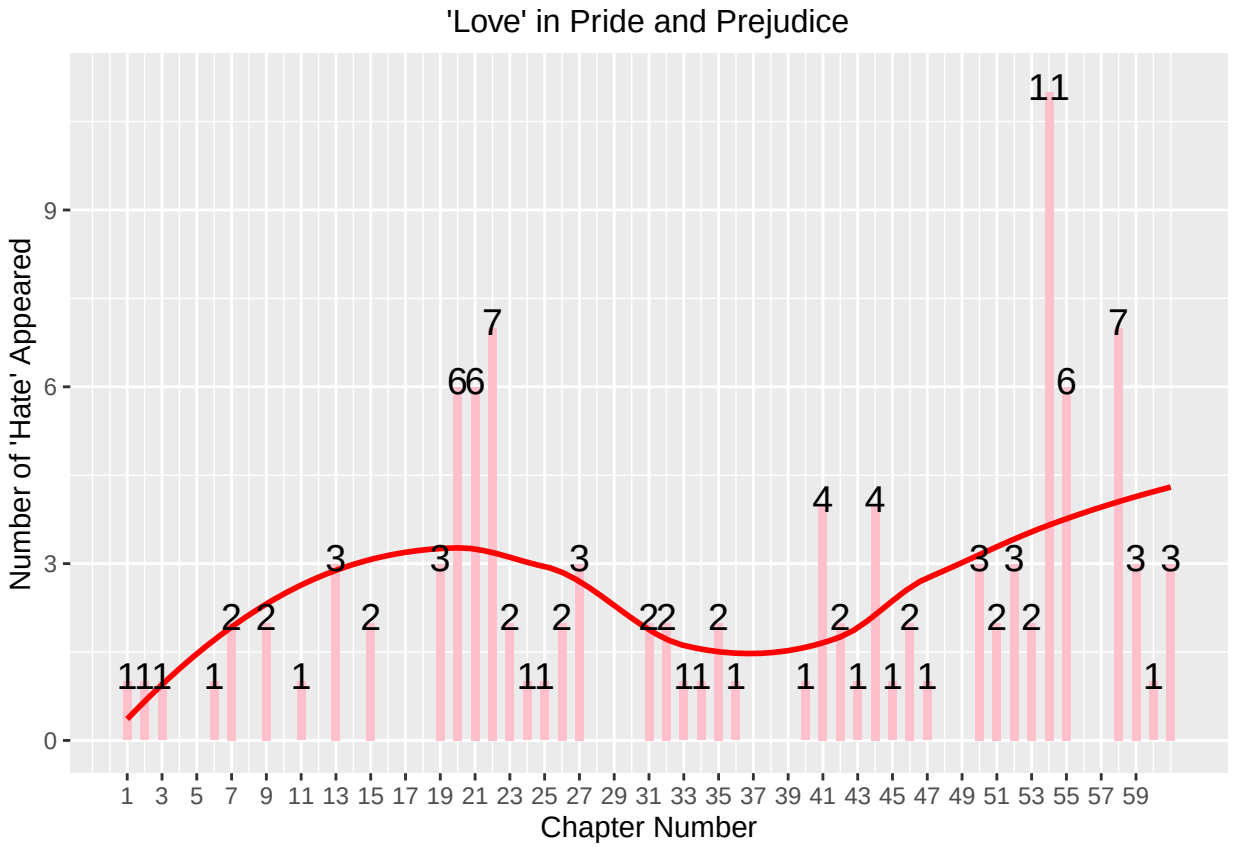
```
#Plot of word "hate"

#tag the word "hate" in the 'pride and prejudice'
tnum.tagByQuery("*pride* has text = REGEXP(\"hate\")","reference:group2hate" )
#extract all the sentences that contain 'hate'
pride.num_hate = tnum.query("*pride* has text = REGEXP(\"hate\")",max=500)
#make data frame from list of tnum objects
pride.textdf_hate=tnum.objectsToDf(pride.num_hate)
```

```
#find the number of chapters that contain the word 'hate'.
pride.text_hate = separate(pride.textdf_hate, col = subject,
  c("book","chapter","other"), sep = "/", remove = FALSE)

#count the total number of times word "hate" appears in each chapter
pride.count_hate = pride.text_hate %>%
  group_by(chapter) %>% summarise(count = n())
pride.count_hate = separate(pride.count_hate, col = chapter,
  c("Chapter", "number"), sep = "-", remove = FALSE)

#plot the word "hate" by chapters
ggplot(data = pride.count_hate, aes(x = sort(as.numeric(number)), y = count)) +
  geom_bar(stat = "identity",fill = "skyblue",width = 0.5) +
  geom_smooth(se=F,color="steelblue")+
  geom_text(aes(label=count,y=count+0.1),size=5,color="black")+
  scale_x_continuous( breaks = seq(1,60,2))+
  ggtitle("'Hate' in Pride and Prejudice")+
  labs(x = "Chapter Number", y = "Number of 'Hate' Appeared")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```




```
#Plot of word "love"

#tag the word "love" in the 'pride and prejudice'
tnum.tagByQuery("*pride* has text = REGEXP(\"love\")","reference:group2love" )

#extract all the sentences that contain 'love'
pride.num_love=tnum.query("*pride* has text = REGEXP(\"love\")",max=500)
```

```
#make data frame from list of tnum objects
pride.textdf_love=tnum.objectsToDf(pride.num_love)

#find the number of chapters that contain the word 'love'
pride.text_love = separate(pride.textdf_love, col = subject,
  c("book","chapter","other"), sep = "/", remove = FALSE)

#count the total number of times word 'love' appears in each chapter
pride.count_love = pride.text_love %>% group_by(chapter) %>% summarise(count = n())
pride.count_love = separate(pride.count_love, col = chapter, c("Chapter", "number"),
  sep = "-", remove = FALSE)

#plot the word 'love' by chapters
ggplot(data = pride.count_love, aes(x = sort(as.numeric(number)), y = count)) +
  geom_bar(stat = "identity", fill = "pink",width = 0.5) +
  geom_smooth(se=F,color="red")+
  geom_text(aes(label=count,y=count+0.1),size=5,color="black")+
  scale_x_continuous( breaks = seq(1,60,2))+
  ggtitle("'Love' in Pride and Prejudice")+
  labs(x = "Chapter Number", y = "Number of 'Hate' Appeared")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```



Thoughts:

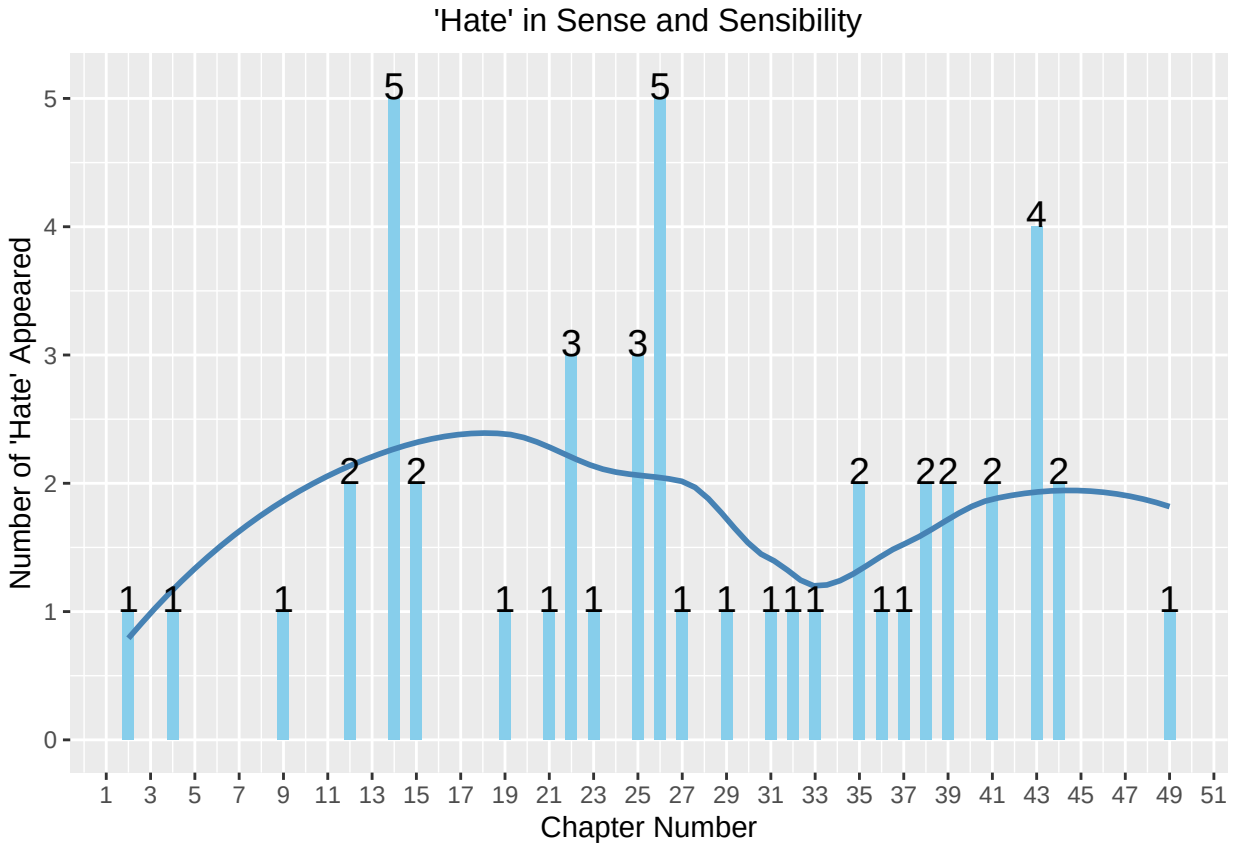
As can be seen from the figure, "hate" mainly appears in the first half of the novel, "love" mainly appears in the second half of the novel, and "love" appears the most in Chapter 54.

Considering the 2 plot, because there is a lot of prejudice between the hero and the heroine at the beginning,

the heroine thinks the hero is a very arrogant man, so in these chapters the word “hate” appear very frequently. In the second half of the book, as they get to know each other, the frequency of word “hate” flattens out.

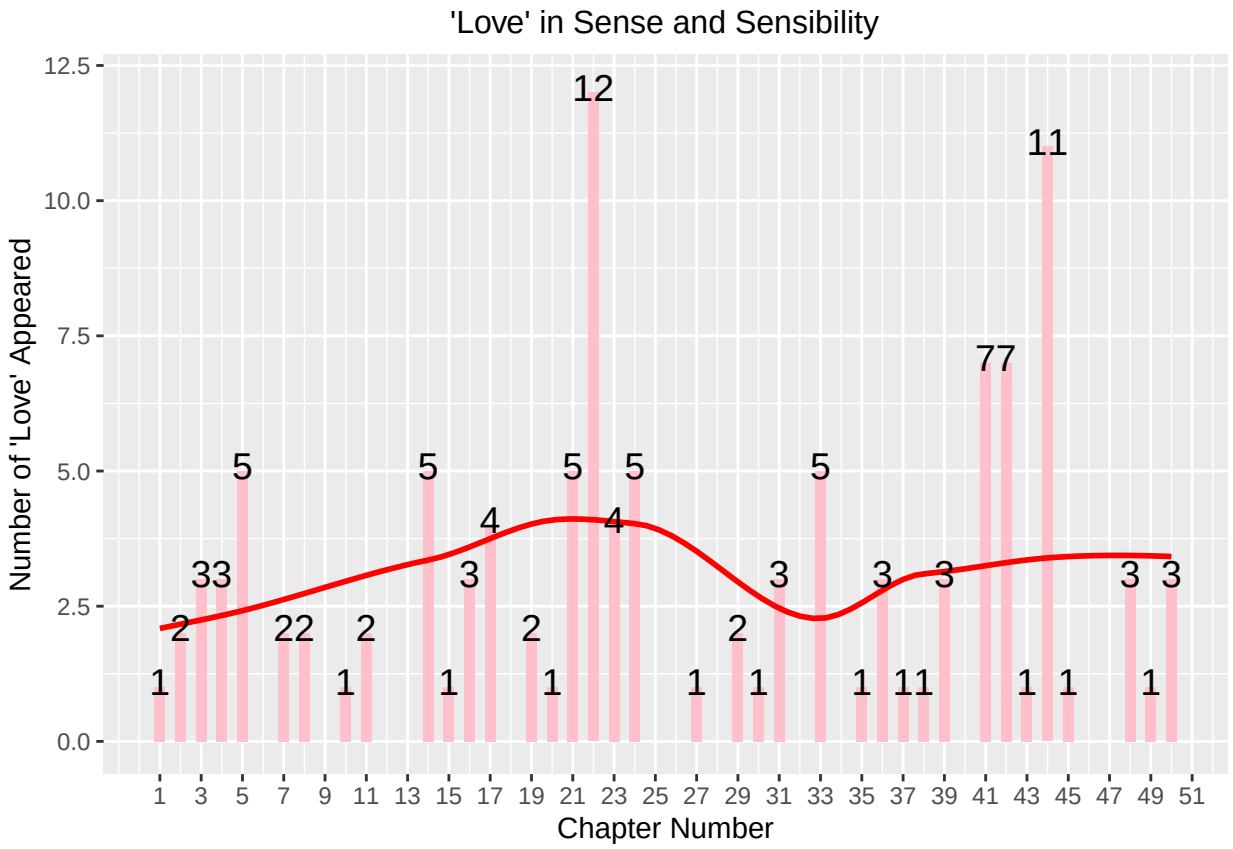
Sense and Sensibility(1811)

```
#Plot of word "hate"

#tag the word "hate" in the 'Sense and Sensibility'
tnum.tagByQuery("*sense* has text = REGEXP(\"hate\")","reference:group2hate" )

#extract all the sentences that contain 'hate'
sense.num_hate = tnum.query("*sense* has text = REGEXP(\"hate\")",max=500)
#make data frame from list of tnum objects
sense.textdf_hate=tnum.objectsToDf(sense.num_hate)

#find the number of chapters that contain the word 'hate'
sense.text_hate = separate(sense.textdf_hate, col = subject,
  c("book","chapter","other"), sep = "/", remove = FALSE)

#count the total number of times word 'hate' appears in each chapter
sense.count_hate = sense.text_hate %>% group_by(chapter) %>% summarise(count = n())
sense.count_hate = separate(sense.count_hate, col = chapter, c("Chapter", "number"),
  sep = "-", remove = FALSE)

#plot the word 'hate' by chapters
ggplot(data = sense.count_hate, aes(x = sort(as.numeric(number)), y = count)) +
  geom_bar(stat = "identity",fill = "skyblue",width = 0.5) +
  geom_smooth(se=F,color="steelblue")+
  geom_text(aes(label=count,y=count+0.1),size=5,color="black")+
  scale_x_continuous( breaks = seq(1,60,2))+
  ggtitle("'Hate' in Sense and Sensibility")+
  labs(x = "Chapter Number", y = "Number of 'Hate' Appeared")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```




```
#Plot of word "love"

#tag the word "love" in the 'Sense and Sensibility'
tnum.tagByQuery("*sense* has text = REGEXP(\"love\")","reference:group2hate" )

#extract all the sentences that contain 'love'
sense.num_love = tnum.query("*sense* has text = REGEXP(\"love\")",max=500)
#make data frame from list of tnum objects
sense.textdf_love=tnum.objectsToDf(sense.num_love)

#find the number of chapters that contain the word 'love'
sense.text_love = separate(sense.textdf_love, col = subject,
  c("book","chapter","other"), sep = "/", remove = FALSE)

#count the total number of times word 'love' appears in each chapter
sense.count_love = sense.text_love %>% group_by(chapter) %>% summarise(count = n())
sense.count_love = separate(sense.count_love, col = chapter, c("Chapter", "number"),
  sep = "-", remove = FALSE)

#plot the word 'love' by chapters
ggplot(data = sense.count_love, aes(x = sort(as.numeric(number)), y = count)) +
  geom_bar(stat = "identity",fill = "pink",width = 0.5) +
  geom_smooth(se=F,color="red")+
  geom_text(aes(label=count,y=count+0.1),size=5,color="black")+
  scale_x_continuous( breaks = seq(1,60,2))+
  ggtitle("'Love' in Sense and Sensibility")+
  labs(x = "Chapter Number", y = "Number of 'Love' Appeared")+
```

```
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```



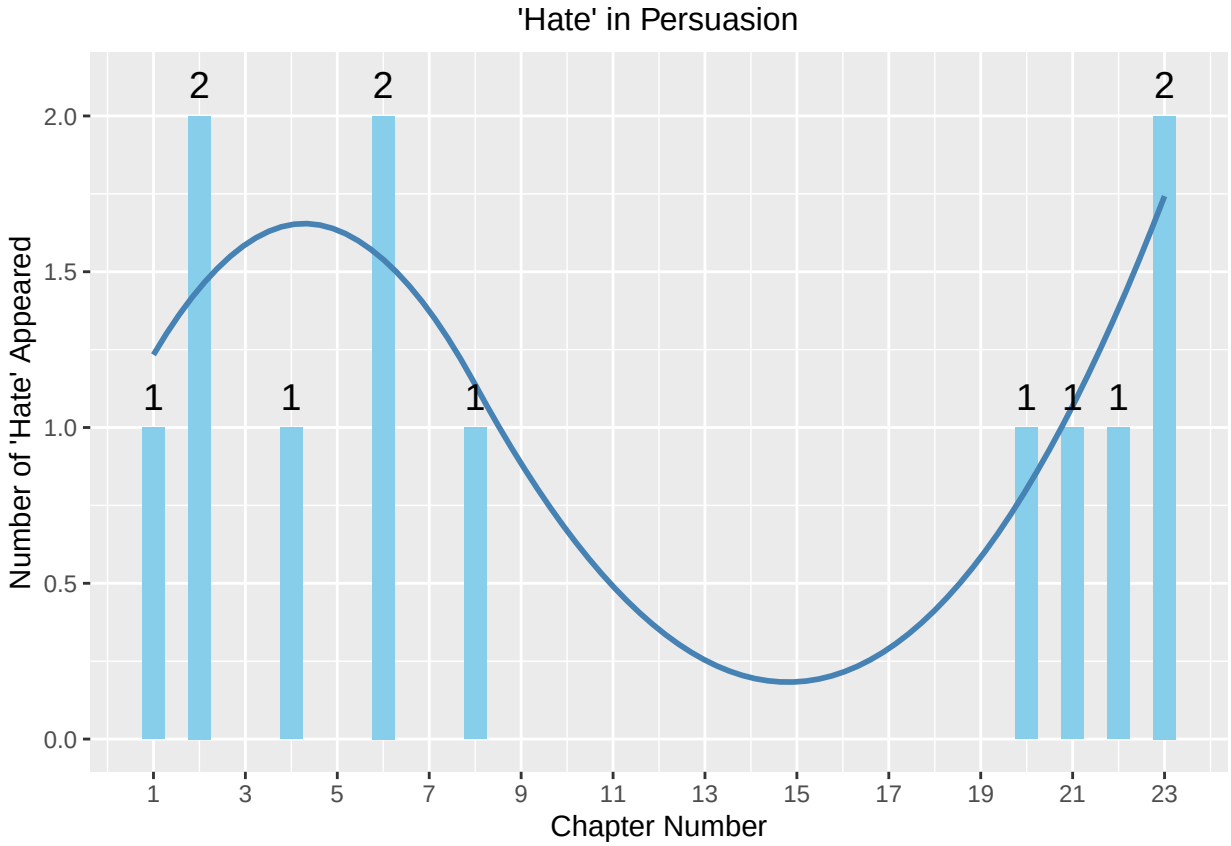
Thoughts:

As we can see from the plot, It is difficult for us to find a trend for word “hate” and “Love”, However, We can see that both “love” and “hate” appears reached the peak in chapters 43-44 and Chapter 21-25. We can infer that the book Sense and Sensibility reached the pinnacle of the plot in these chapters.

Persuasion(1818)

```
#Plot of word "hate"

#tag the word "hate" in the 'Persuasion'
tnum.tagByQuery("*persuasion* has text = REGEXP(\"hate\")","reference:group2hate" )

#extract all the sentences that contain 'hate'
persuasion.num_hate = tnum.query("*persuasion* has text = REGEXP(\"hate\")",max=500)
#make data frame from list of tnum objects
persuasion.textdf_hate=tnum.objectsToDf(persuasion.num_hate)

#find the number of chapters that contain the word 'hate'
persuasion.text_hate = separate(persuasion.textdf_hate, col = subject,
  c("book","chapter","other"), sep = "/", remove = FALSE)

#count the total number of times word 'hate' appears in each chapter
persuasion.count_hate = persuasion.text_hate %>% group_by(chapter) %>% summarise(count = n())
```

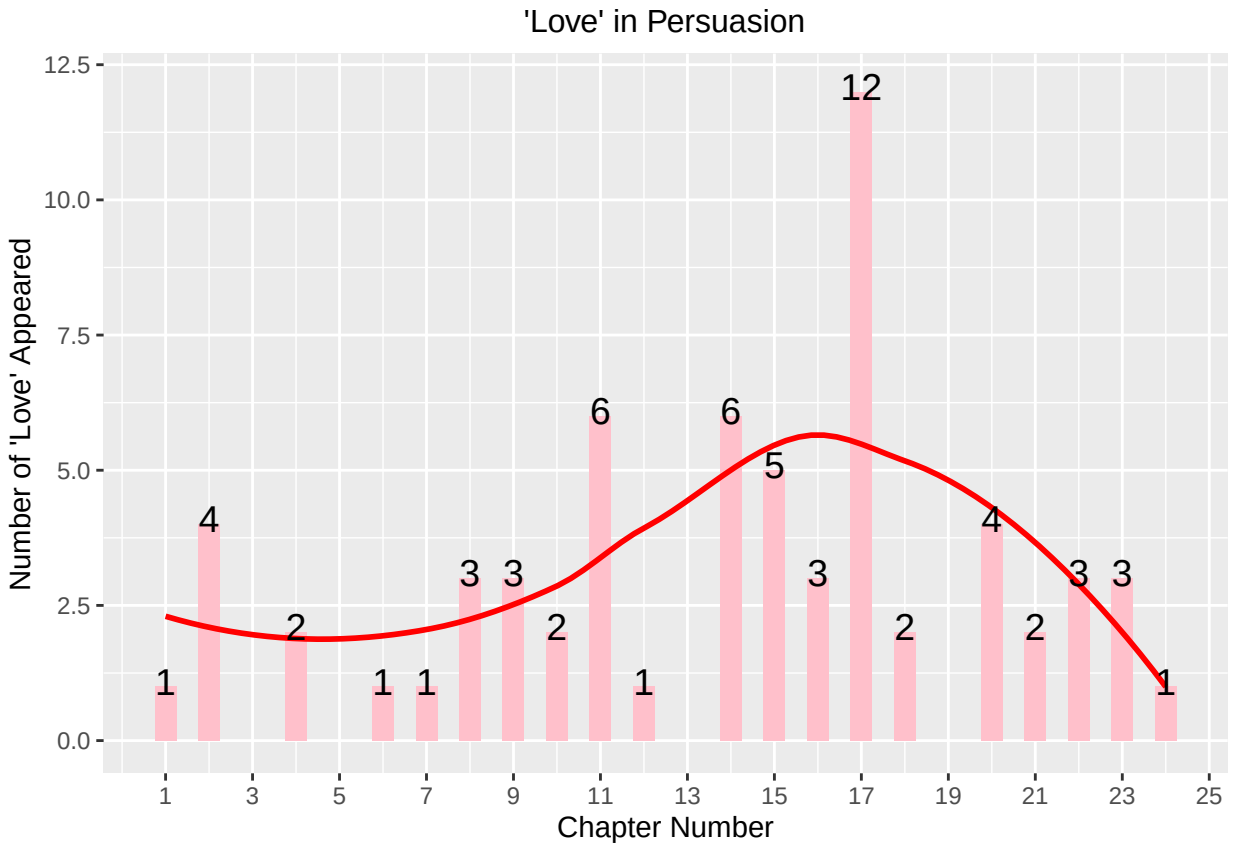
```
persuasion.count_hate = separate(persuasion.count_hate, col = chapter, c("Chapter", "number"),
  sep = "-", remove = FALSE)

#plot the word 'hate' by chapters
ggplot(data = persuasion.count_hate, aes(x = sort(as.numeric(number)), y = count)) +
  geom_bar(stat = "identity",fill = "skyblue",width = 0.5) +
  geom_smooth(se=F,color="steelblue")+
  geom_text(aes(label=count,y=count+0.1),size=5,color="black")+
  scale_x_continuous( breaks = seq(1,60,2))+
  ggtitle("'Hate' in Persuasion")+
  labs(x = "Chapter Number", y = "Number of 'Hate' Appeared")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```



Thoughts:In the book "Persuasion", we can clearly see that for the word'hate' in the first part appears more, and then it is basically relieved. It's the opposite of 'love'.

```
#Plot of word "love"

#tag the word "love" in the 'Persuasion'
tnum.tagByQuery("*persuasion* has text = REGEXP(\"love\")","reference:group2hate" )
#extract all the sentences that contain 'love'
persuasion.num_love = tnum.query("*persuasion* has text = REGEXP(\"love\")",max=500)
#make data frame from list of tnum objects
persuasion.textdf_love=tnum.objectsToDf(persuasion.num_love)

#find the number of chapters that contain the word 'love'
persuasion.text_love = separate(persuasion.textdf_love, col = subject,
```

```
  c("book","chapter","other"), sep = "/", remove = FALSE)

#count the total number of times word 'love' appears in each chapter
persuasion.count_love = persuasion.text_love %>% group_by(chapter) %>% summarise(count = n())
persuasion.count_love = separate(persuasion.count_love, col = chapter,
  c("Chapter", "number"), sep = "-", remove = FALSE)

#plot the word 'love' by chapters
ggplot(data = persuasion.count_love, aes(x = sort(as.numeric(number)), y = count)) +
  geom_bar(stat = "identity",fill = "pink",width = 0.5) +
  geom_smooth(se=F,color="red")+
  geom_text(aes(label=count,y=count+0.1),size=5,color="black")+
  scale_x_continuous( breaks = seq(1,60,2))+
  ggtitle("'Love' in Persuasion")+
  labs(x = "Chapter Number", y = "Number of 'Love' Appeared")+
  theme(plot.title = element_text(hjust = 0.5, size = 12))
```



Thoughts:

'love' in the first two chapters are more frequent, then gradually decreased, and in the middle and late stages, it rose sharply until the seventeenth chapter, it reach the highest point and then return to calm. This is consistent with the plot of the novel.

This is the book work of Jane Austen. Due to the influence of romanticism, the emotion of this book has a big wave. At first, Annie's father thought that Wentworth was born in a poor family, so the relationship between the Annie and Wentworth is gradually changed from being in love to a bit of hatred. Later, when Colonel Wentworth returned to his hometown, these two person met again by chance, and Annie had never forgotten Wentworth. So in the end, Annie and Wentworth fall in love again.

There is a little confusion at the end of the picture, which may be due to the author's death before the book was published. So there may be some problems in the last few chapters, we don't know. But after discarding the last few chapters, overall, the word frequency of 'love' and 'hate' matches the sentiment of the chapters of the book.

Comparison

Now we want to analyze the ratio of the frequency of the words "love" and "hate" in each book written by Jane Austen1. We call it love_hate_ratio.

```
#get the number of love/hate in each book
emma_love=tnum.query("*emma* has text = REGEXP(\"love\")",max=500)
```

```
## Returned 1 thru 181 of 181 results
```

```
emma_hate=tnum.query("*emma* has text = REGEXP(\"hate\")",max=500)
```

```
## Returned 1 thru 28 of 28 results
```

```
pride_love=tnum.query("*pride* has text = REGEXP(\"love\")",max=500)
```

```
## Returned 1 thru 111 of 111 results
```

```
pride_hate=tnum.query("*pride* has text = REGEXP(\"hate\")",max=500)
```

```
## Returned 1 thru 49 of 49 results
```

```
persuasion_love=tnum.query("*persuasion* has text = REGEXP(\"love\")",max=500)
```

```
## Returned 1 thru 65 of 65 results
```

```
persuasion_hate=tnum.query("*persuasion* has text = REGEXP(\"hate\")",max=500)
```

```
## Returned 1 thru 12 of 12 results
```

```
sense_love=tnum.query("*sense* has text = REGEXP(\"love\")",max=500)
```

```
## Returned 1 thru 118 of 118 results
```

```
sense_hate=tnum.query("*sense* has text = REGEXP(\"hate\")",max=500)
```

```
## Returned 1 thru 48 of 48 results
```

```
park_love=tnum.query("*park* has text = REGEXP(\"love\")",max=500)
```

```
## Returned 1 thru 185 of 185 results
```

```
park_hate=tnum.query("*park* has text = REGEXP(\"hate\")",max=500)
```

```
## Returned 1 thru 48 of 48 results
```

```
abbey_love=tnum.query("*abbey* has text = REGEXP(\"love\")",max=500)
```

```
## Returned 1 thru 77 of 77 results
```

```
abbey_hate=tnum.query("*abbey* has text = REGEXP(\"hate\")",max=500)
```

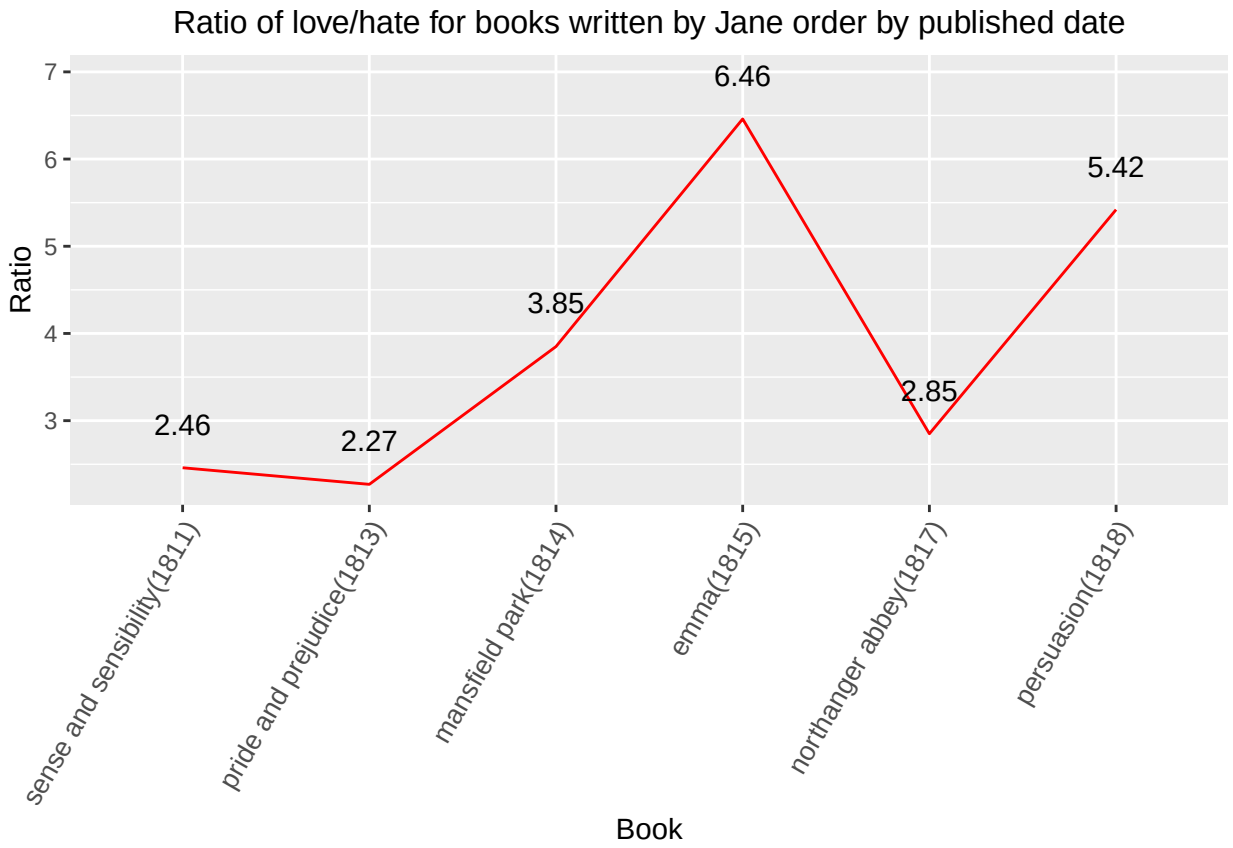
```
## Returned 1 thru 27 of 27 results
```

Now we create a data frame to analyze the love_hate_ratio for Jane Austin's books

```
#list of book ordered by the published time
book = c(
  "sense and sensibility(1811)",
```

```
  "pride and prejudice(1813)",
  "mansfield park(1814)",
  "emma(1815)",
  "northanger abbey(1817)",
  "persuasion(1818)"
)
book= factor(book, levels=unique(book))

#frequency of word "love" appeared in each book
love = c(length(sense_love),
         length(pride_love),
         length(park_love),
         length(emma_love),
         length(abbey_love),
         length(persuasion.num_love)
         )

#frequency of word "hate" appeared in each book
hate = c(length(sense_hate),
         length(pride_hate),
         length(park_hate),
         length(emma_hate),
         length(abbey_hate),
         length(persuasion.num_hate)
         )

#compute the ratio
ratio = love/hate

#Gain the data frame
emotion = data.frame(book, ratio, love, hate)
for (i in 1:6) {
  emotion$sum[i] <- sum(emotion$love[i]+emotion$hate[i])
}
emotion$ratio <- round(emotion$ratio,2)
```

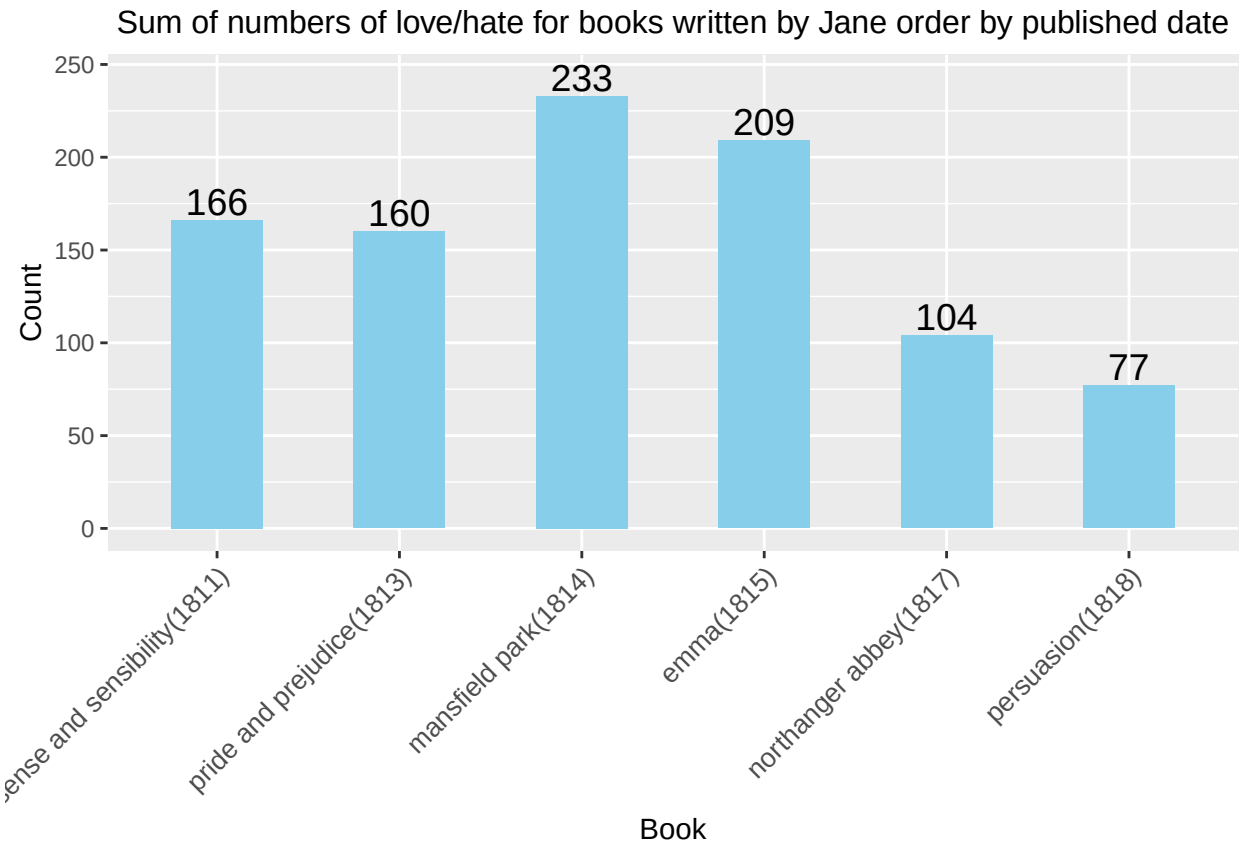
```
#We plot the love/hate ratio for books written by Jane ordered by published date.
ggplot(data = emotion)+
  geom_path(aes(x = book, y = ratio, group = 1),color="red") +
  geom_text(aes(x=book,label=ratio,y=ratio+0.5),size=4,color="black")+
  ggtitle("Ratio of love/hate for books written by Jane order by published date")+
   theme(axis.text.x = element_text(angle = 60, hjust = 1,size = 10))+
  theme(plot.title = element_text(hjust = 0.5, size = 12))+
  xlab("Book")+ylab("Ratio")
```



Thoughts:

As we can see from the graph, love/ hate ratio increase dramatically from 1811 to 1815, During this time, Jane Austen experienced a happy and stable time. But after 1815, Jane Austen's physical condition began to decline and became seriously ill, so at this time,the ove/ hate ratio in her books decrease.

```
#plot Sum of numbers of love/hate
ggplot(data = emotion)+ aes(x=book,y=sum)+
  geom_bar(stat = "identity",fill = "skyblue",width = 0.5)+
  geom_text(aes(label=sum,y=sum+10),size=5,color="black")+
  ggtitle("Sum of numbers of love/hate for books written by Jane order by published date") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,size = 10))+
  theme(plot.title = element_text(hjust = 0.5, size = 12))+
  xlab("Book")+ylab("Count")
```



Thoughts:

Also, from the plot, We can see that the frequency of emotional vocabulary reached its peak in the novels written by Jane Austen in 1814 and 1815. This may be because she had the most peaceful and beautiful life during this period