

Tagging Assignment

Group 6

2020/11/20

Introduction

Our work is to explore Jane Austen’s work using `tnum` fuctions to search and tag aspects of the work that reflect our thoughts and speculations.We choose the *Sense and Sensibility* to explore.

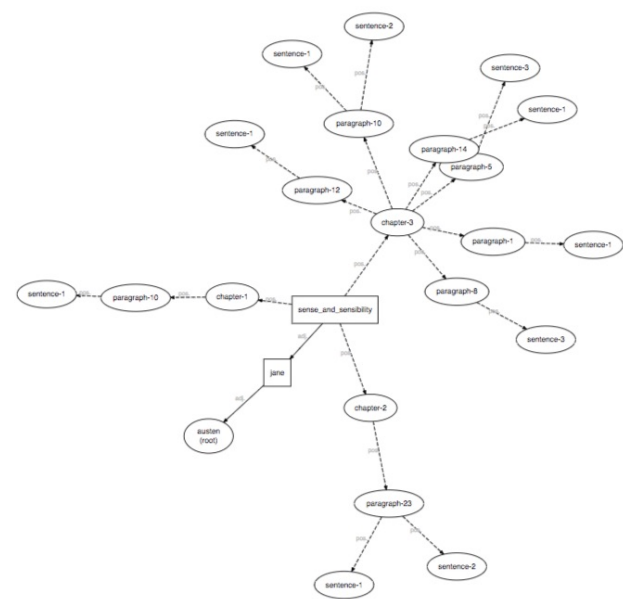
The novel follows the three Dashwood sisters as they must move with their widowed mother from the estate on which they grew up, Norland Park. Because Norland is passed down to John, the product of Mr. Dashwood’s first marriage, and his young son, the four Dashwood women need to look for a new home. They have the opportunity to rent a modest home, Barton Cottage, on the property of a distant relative, Sir John Middleton. There they experience love, romance, and heartbreak. The novel is likely set in southwest England, London, and Sussex between 1792 and 1797.

Our work are divided into 4 part – one topic for each part. # Part1:Positive Emotion Tag

Positive Emotion

Positive Emotion is considered a necessity in our life. In the book *Sense and Sensibility*, the author has a lot of emotional portrayals of the characters in the book. Therefore, I picked some words which can represent a positive emotion, and tag them with **ref:emotion**.

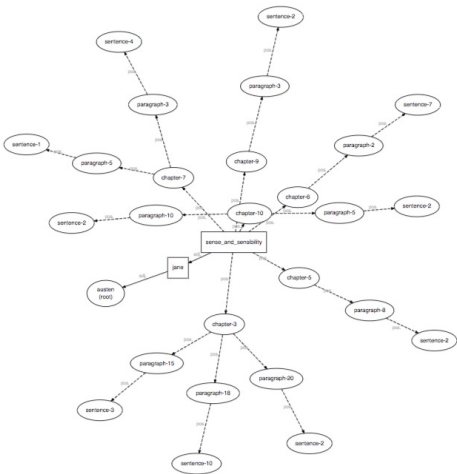
Additionally, I drew a treeplot to see how these tags are distributed in the book.



Happy

Happy is a kind of positive emotion, and it is considered the most direct way to express happiness. Therefore, I picked *happy* and tag them with **ref:happy**.

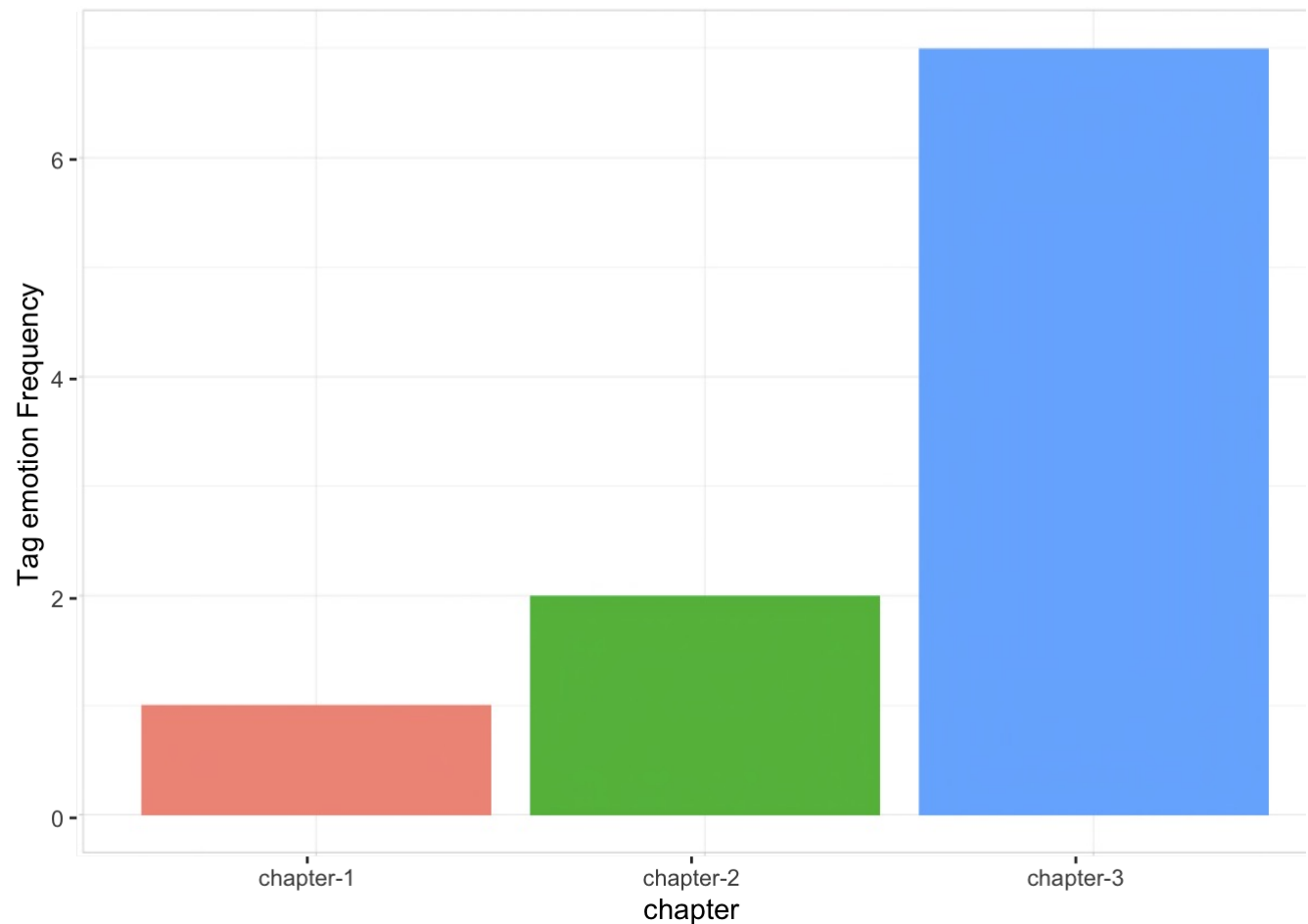
Also, a treeplot is attached.



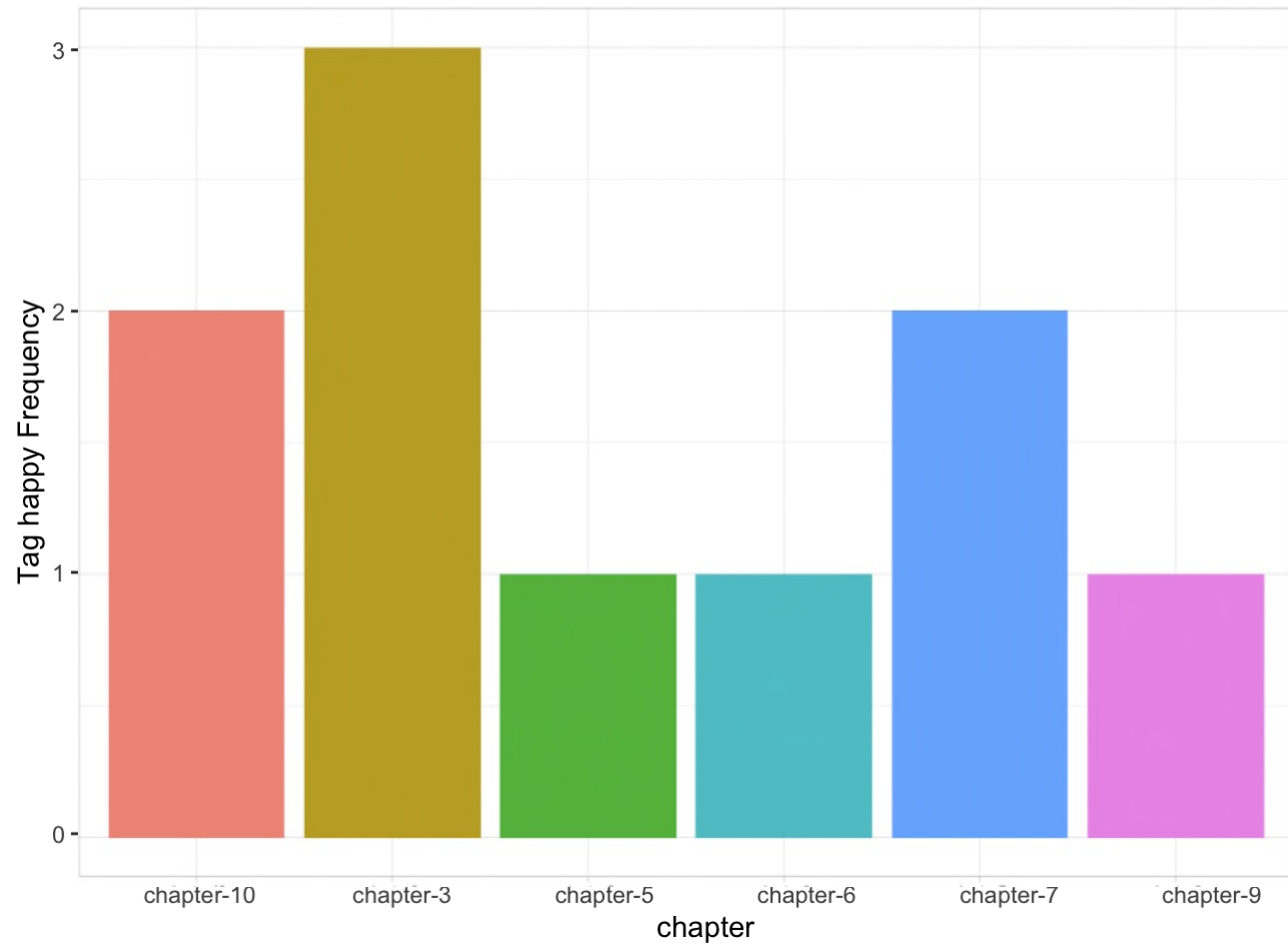
Bar chart

I use bar plot to show the frequency of each tag in each chapter.

```
emo_sep <- separate(data = emo,col = subject,into = c("book","chapter","paragrah","sentence"),sep = "/")
ggplot(data = emo_sep,aes(x=chapter))+
  geom_histogram(stat = "count")+
  labs(x="chapter",y="Tag emotion Frequency")
```



```
hap_sep <- separate(data = hap,col = subject,into = c("book","chapter","paragrah","sentence"),sep = "/")
ggplot(data = hap_sep,aes(x=chapter))+
  geom_histogram(stat = "count")+
  labs(x="chapter",y="Tag happy Frequency")
```



From the plots, we can intuitively see the distribution of positive emotions in each chapter. # Part2: Sense and Sensibility Tag Elinor represents the “sense” half and Marianne represents the “sensibility” half of Austen’s title Sense and Sensibility. Therefore, their actions could display sense and sensibility. We choose some words which represent sense and sensibility respectively and show how they are distributed by chapter.

Passion and Love

Firstly, we tag *passion* and *love* with sensibility.

```
# tag passion and love with sensibility
# query1 <- tnum.query("*sense* has * = REGEXP(\"passion\")")
# query2 <- tnum.query("*sense* has * = REGEXP(\"love\")")
tnum.tagByQuery("*sense* has * = REGEXP(\"passion | love\")",adds = ("preformance:sensibility"))
tag_sensibility<-tnum.query("@preformance:sensibility",max=200)
performance_sensibility<-tnum.objectsToDf(tag_sensibility)
```

Passion and Love

Besides, we tag *control*,*help* and *hesitate* with sense.

```
#tag control, help and hesitate with sense
# query4 <- tnum.query("*sense* has * = REGEXP(\"control\")")
# query6 <- tnum.query("*sense* has * = REGEXP(\"help\")")
# query7 <- tnum.query("*sense* has * = REGEXP(\"hesitate\")")
tnum.tagByQuery("*sense* has * = REGEXP(\"control | help | hesitate\")",adds = ("preformance:sense"))
tag_sense<-tnum.query("@preformance:sense",max=20)
performance_sense<-tnum.objectsToDf(tag_sense)
```

Bar chart

Next, we use bar chart to show how they are distributed by chapter.

```
# bar chart: show how they are distributed by chapter.

# sensibility
## seperate subject and extract chapter
performance_sensibility1 <- performance_sensibility %>% separate(subject, c("book", "chapter", "paragraph", "sentence"), sep = "/")
performance_sensibility1$chapter %<>% str_replace("chapter-", "") %<>% as.numeric()
performance_sensibility1$paragraph%<>% str_replace("paragraph-", "") %<>% as.numeric()
performance_sensibility1$sentence%<>% str_replace("sentence-", "") %<>% as.numeric()

count_sensibility <- performance_sensibility1 %>% group_by(chapter) %>% summarise(count = sum(string.value != ""))

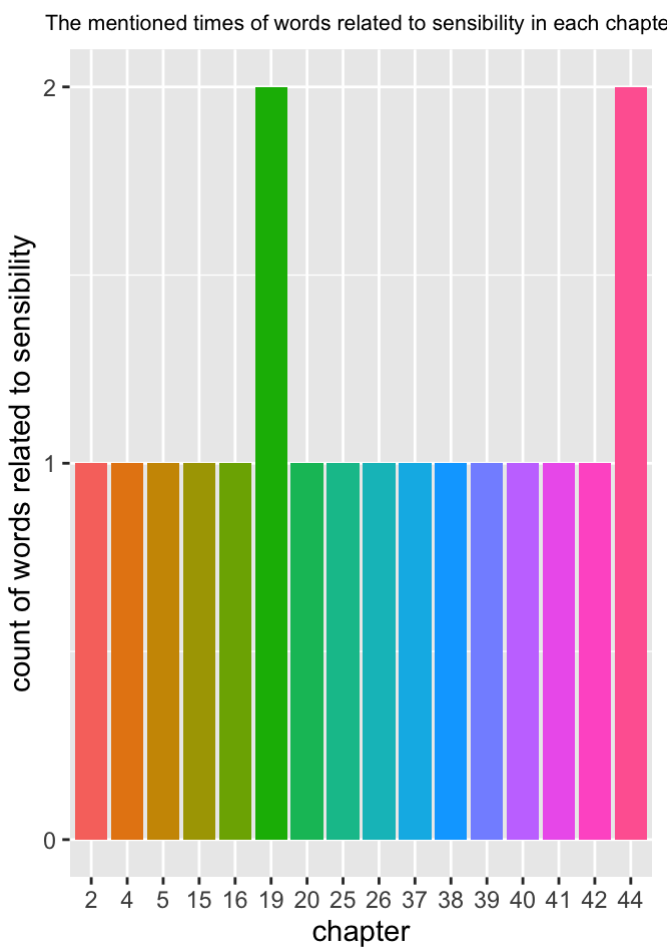
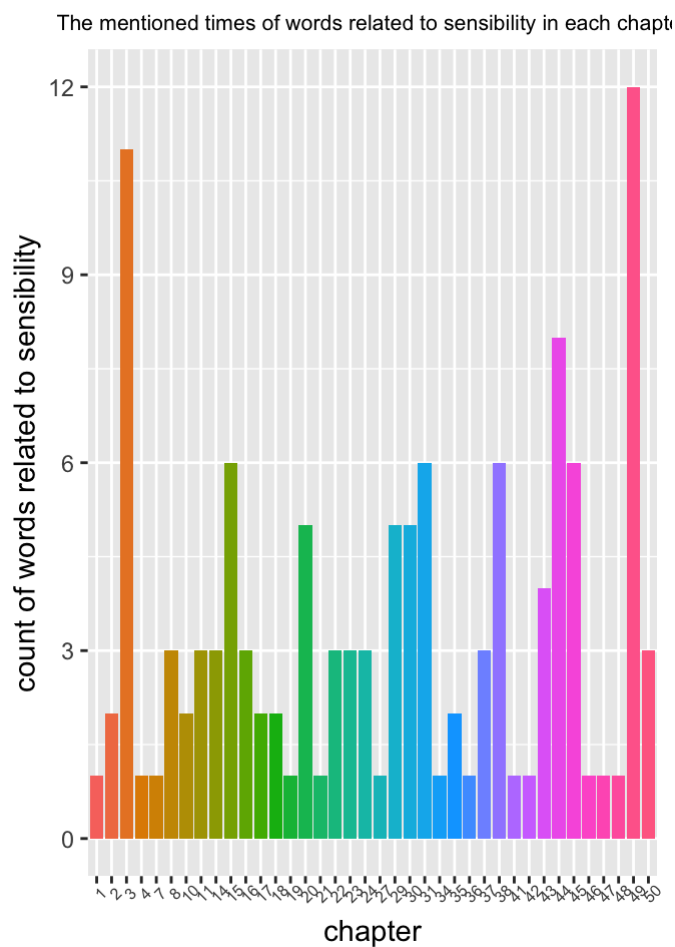
## bar chart
p1<-ggplot(data = count_sensibility, mapping = aes(x = factor(chapter), y = count, fill = factor(chapter))) +
  geom_bar(stat = "identity") +
  guides(fill = F) +
  xlab("chapter") +
  ylab("count of words related to sensibility") +
  scale_y_continuous(limits=c(0,12), breaks=seq(0,12,3))+
  ggtitle("The mentioned times of words related to sensibility in each chapter") +
  theme(plot.title = element_text(hjust = 0.5,size = 8),
        axis.text.x=element_text(angle=45,size=6))

# sense
## seperate subject and extract chapter
performance_sense1 <- performance_sense %>% separate(subject, c("book", "chapter", "paragraph", "sentence"), sep = "/")
performance_sense1$chapter %<>% str_replace("chapter-", "") %<>% as.numeric()
performance_sense1$paragraph%<>% str_replace("paragraph-", "") %<>% as.numeric()
performance_sense1$sentence%<>% str_replace("sentence-", "") %<>% as.numeric()

count_sense <- performance_sense1 %>% group_by(chapter) %>% summarise(count = sum(string.value != ""))

## bar chart
p2<-ggplot(data = count_sense, mapping = aes(x = factor(chapter), y = count, fill = factor(chapter))) +
  geom_bar(stat = "identity") +
  guides(fill = F) +
  xlab("chapter") +
  ylab("count of words related to sensibility") +
  scale_y_continuous(limits=c(0,2), breaks=seq(0,2,1))+
  ggtitle("The mentioned times of words related to sensibility in each chapter") +
  theme(plot.title = element_text(hjust = 0.5, size = 8))

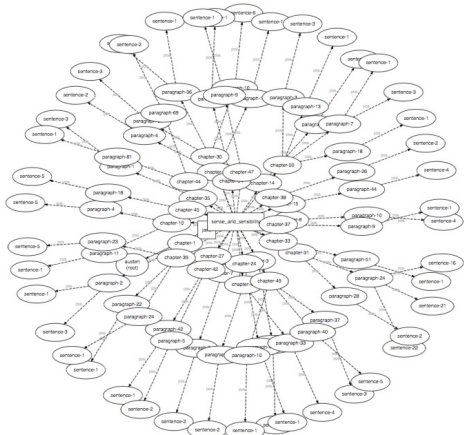
cowplot::plot_grid(p1, p2, nrow = 1)
```



Based on the plots, we can find that words related to sensibility are more than words related to sense. I think people who are emotional are more likely to express their thoughts. However, people who are sensible prefer to suppresses their emotions and hide their thoughts. # Part3: Marriage Tag

Marriage

We chose this tag because this novel(Sense and Sensibility) is based on the marriages of the time.The protagonist's emotional experience and marriage are the key to the development of the story.So I think by analyzing this tag, we can see the development of the plot of the whole book. First,We have marked all the chapters and sentences related to marriage.And then draw tree plot and barplot to see how they're distributed throughout the book.

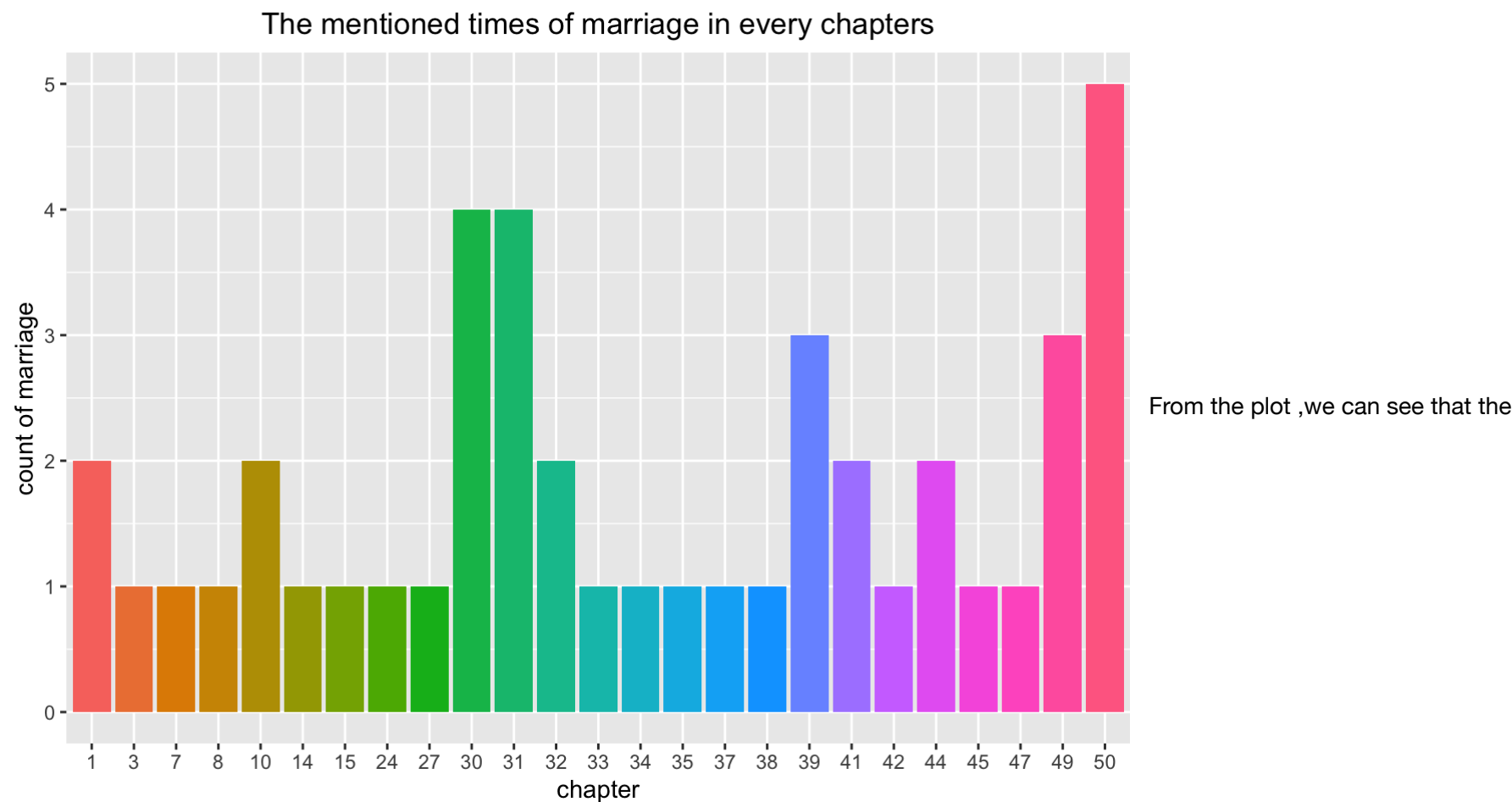


Bar chart

```
# barplot
marriage %<>% tidyr::separate(subject, c("book", "chapter", "paragraph", "sentence"), sep = "/")
marriage$chapter %<>% str_replace("chapter-", "") %<>% as.numeric()
marriage$paragraph %<>% str_replace("paragraph-", "") %<>% as.numeric()
marriage$sentence %<>% str_replace("sentence-", "") %<>% as.numeric()

countmarriage <- marriage %>% group_by(chapter) %>% summarise(count = sum(string.value != ""))

ggplot(data = countmarriage, mapping = aes(x = factor(chapter), y = count, fill = factor(chapter))) +
  geom_bar(stat = "identity") +
  guides(fill = F) +
  xlab("chapter") +
  ylab("count of marriage") +
  ggtitle("The mentioned times of marriage in every chapters") +
  theme(plot.title = element_text(hjust = 0.5))
```



whole book contains the topic of marriage.And the in last chapter,marriage is mentioned more often than any other chapters. # Part4: NRC Sentiment Tag

Data Organization

Observe the book structure

```
tnum.getDatabasePhraseList("subject", pattern= "*",levels=5)
```

There are 50 chapters in the *Sense and Sensibility*.

Tag Elinor,Marianne, Willoughby and Edward

```
#tnum.query("*Sense* has text = REGEXP(\"Elinor\")")
tnum.tagByQuery("*Sense* has text = REGEXP(\"Elinor\")", adds=("w_elinor"))
tag_elinor=tnum.query("@w_elinor",max=653)
elinor=tnum.objectsToDf(tag_elinor)

#tnum.query("*Sense* has text = REGEXP(\"Marianne\")")
tnum.tagByQuery("*Sense* has text = REGEXP(\"Marianne\")", adds=("w_marianne"))
tag_marianne=tnum.query("@w_marianne",max=524)
marianne=tnum.objectsToDf(tag_marianne)

cooccur = filter(marianne,grepl('w_elinor',tags))

#tnum.query("*Sense* has text = REGEXP(\"Willoughby\")")
tnum.tagByQuery("*Sense* has text = REGEXP(\"Willoughby\")", adds=("w_willoughby"))
tag_willoughby=tnum.query("@w_willoughby",max=200)
willoughby=tnum.objectsToDf(tag_willoughby)

#tnum.query("*Sense* has text = REGEXP(\"Edward\")")
tnum.tagByQuery("*Sense* has text = REGEXP(\"Edward\")", adds=("w_edward"))
tag_edward=tnum.query("@w_edward",max=254)
edward=tnum.objectsToDf(tag_edward)

M_W=filter(willoughby,grepl('w_marianne',tags))
E_E=filter(edward,grepl('w_elinor',tags))
```

Separate the subject column

```
mydf=function(df){
  a=df %>%
  separate(subject,c("book","chapter","paragraph","sentence"),sep="/")%>%
  separate(chapter,c("lable","chapter"),sep="-")%>%
  separate(sentence,c("label1","sentence"),sep="-")%>%
  separate(paragraph,c("label2","paragraph"),sep="-")%>%
  select(book,chapter,paragraph,sentence,string.value)
  a$chapter = as.numeric(a$chapter)
  a$paragraph= as.numeric(a$paragraph)
  a$sentence =as.numeric(a$sentence)
  return(a)
}

df_elinor = mydf(elinor)
df_marianne = mydf(marianne)
df_cooccur = mydf(cooccur)
df_MW = mydf(M_W)
df_EE = mydf(E_E)
```

EDA: Sentiment of Marianne and Elinor

Add the NRC sentiment to the data

```
#### add sentiment to elinor and marianne
st_elinor=get_nrc_sentiment(df_elinor$string.value)
st_df_elinor=cbind(df_elinor,st_elinor)

st_marianee=get_nrc_sentiment(df_marianne$string.value)
st_df_marianne=cbind(df_marianne,st_marianee)

st_MW=get_nrc_sentiment(df_MW$string.value)
st_df_MW=cbind(df_MW,st_MW)

st_EE=get_nrc_sentiment(df_EE$string.value)
st_df_EE=cbind(df_EE,st_EE)

#### group_by and summarise for plot
plot_st_df =function(st_df){
  b=st_df %>%
  group_by(chapter,paragraph) %>%
  summarise(anger=sum(anger),anticipation=sum(anticipation),disgust=sum(disgust),fear=sum(fear),joy=sum(joy),sadness=sum(sadness),surprise=sum(surprise),trust=sum(trust),negative=sum(negative),positive=sum(positive))
  b$index=1:nrow(b)
  return(b)
}

plot_st_elinor=plot_st_df(st_df_elinor)
plot_st_marianne=plot_st_df(st_df_marianne)
plot_st_MW=plot_st_df(st_df_MW)
plot_st_EE=plot_st_df(st_df_EE)

#### join elinor and marianne by chapter and paragraph
plot_st=left_join(plot_st_elinor, plot_st_marianne,by=c("chapter","paragraph"))
plot_st$positive=plot_st$positive.y-plot_st$positive.x
plot_st$negative=plot_st$negative.y-plot_st$negative.x
plot_st$anger=plot_st$anger.y-plot_st$anger.x
plot_st$anticipation=plot_st$anticipation.y-plot_st$anticipation.x
plot_st$disgust=plot_st$disgust.y-plot_st$disgust.x
plot_st$fear=plot_st$fear.y-plot_st$fear.x
plot_st$joy=plot_st$joy.y-plot_st$joy.x
plot_st$sadness=plot_st$sadness.y-plot_st$sadness.x
plot_st$surprise=plot_st$surprise.y-plot_st$surprise.x
plot_st$trust=plot_st$trust.y-plot_st$trust.x
plot_st$total_st=plot_st$anger+plot_st$anticipation+plot_st$disgust+plot_st$fear+plot_st$joy+ plot_st$sadness +plot_st$surprise +plot_st$trust
plot_st$pn=plot_st$positive+plot_st$negative
#+ plot_st$negative +plot_st$positive
plot_st$index=1:nrow(plot_st)
```

Compare the total sentiment word counts of Marianne and Elinor

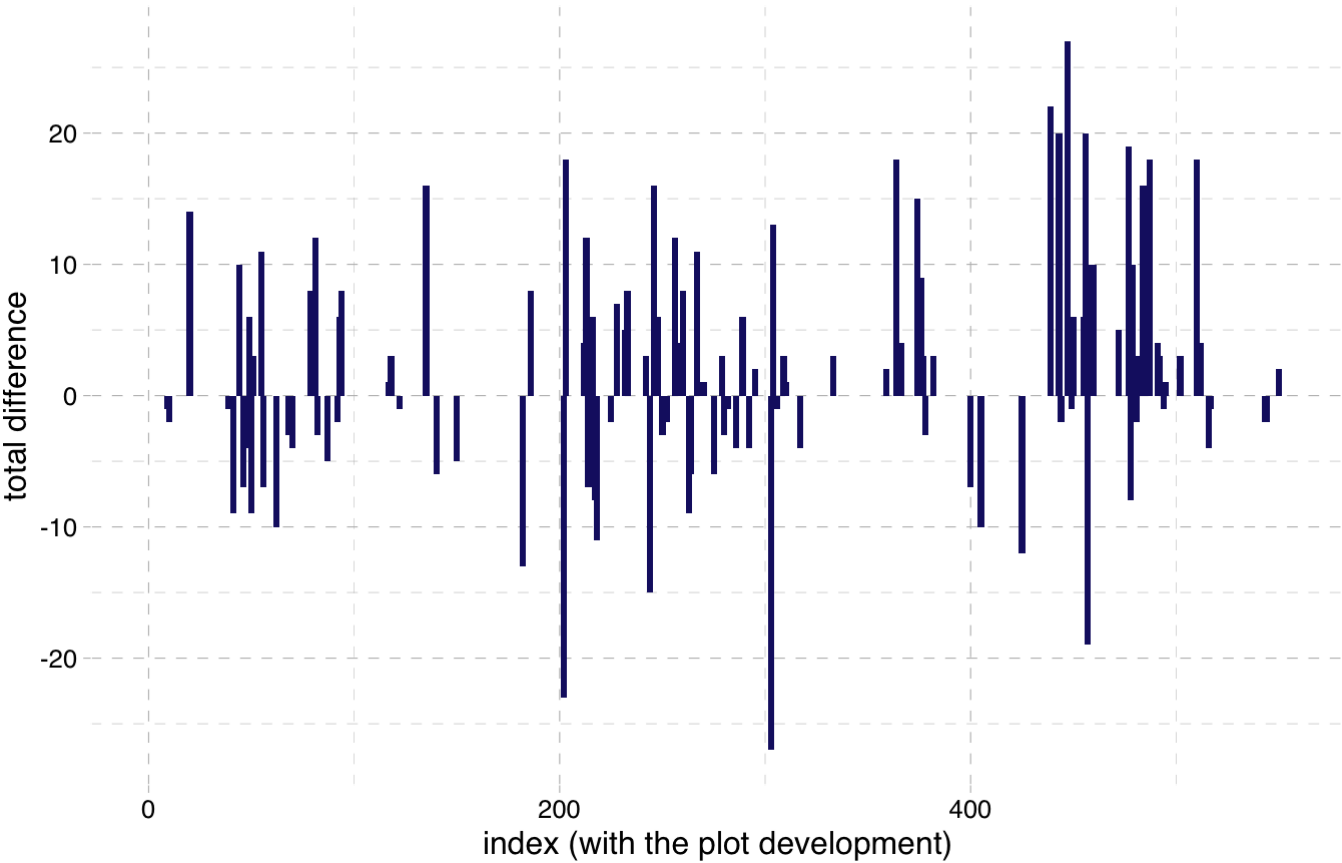
```
## Adding missing grouping variables: `chapter`
## Adding missing grouping variables: `chapter`
```

```
## Warning: position_dodge requires non-overlapping x intervals
```

```
## Warning: Removed 354 rows containing missing values (geom_bar).
```

The total sentiment word counts of Marianne is 229 more than Elinor

Sense and Sensibility



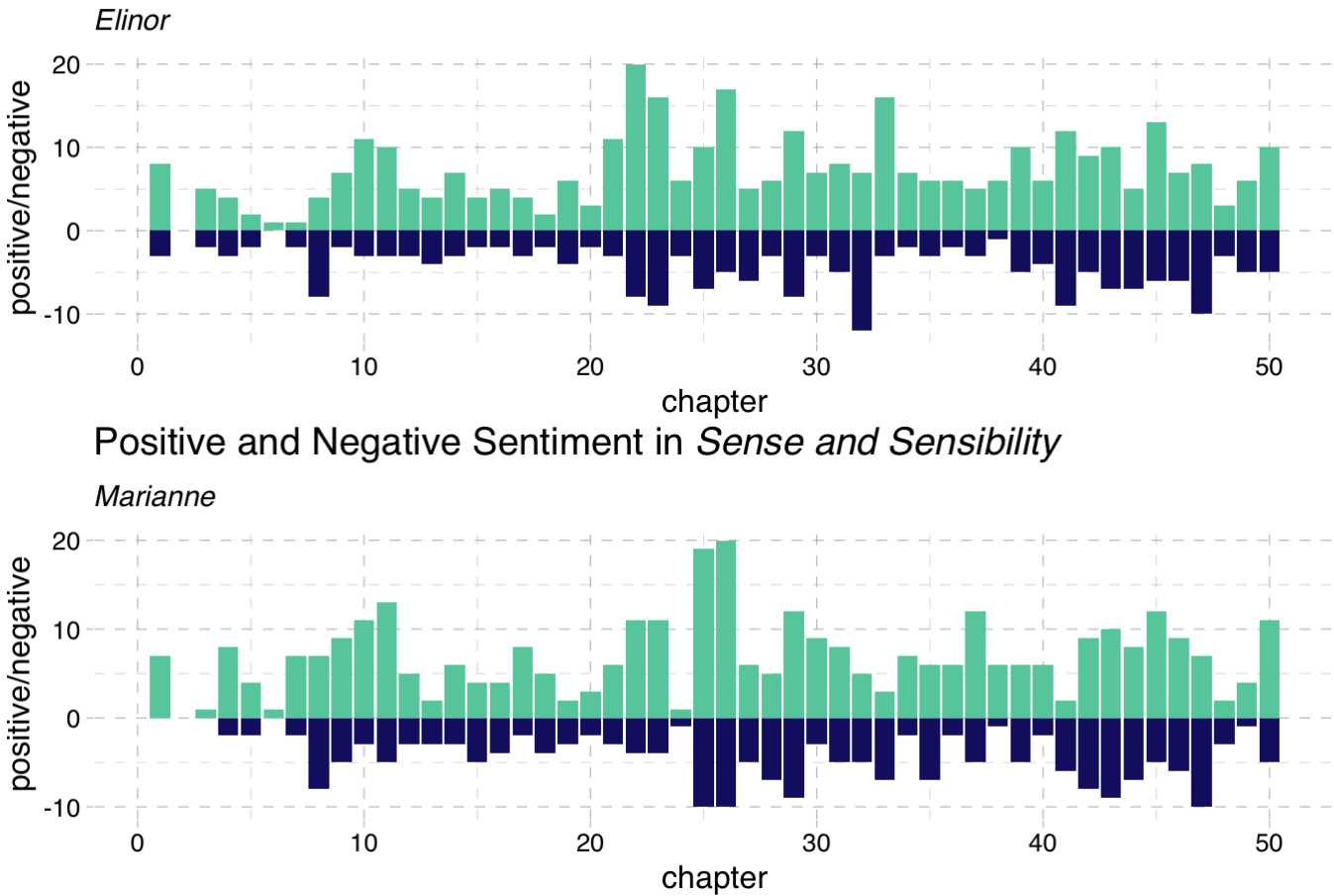
According to the plot, we can observe the different value of the sentiment word counts between Marianne and Elinor. When the bar is above the $y = 0$ axis, it means Marianne is related with more sentiment words.

Connecting with the characters in this book, I think this result is appropriate. Elinor Dashwood is the sensible and reserved eldest daughter, and she represents the “sense” half of Austen’s title *Sense and Sensibility*. Marianne Dashwood is the romantically inclined and eagerly expressive second daughter, and her emotional excesses identify her as the “sensibility” half of Austen’s title. Obviously, Marianne related with more sentiment words is consistent with the characters’ personalities.

Explore the sentiment along with plot development

Now, I want to research the sentiment trends of these two characters along with some important plots.

Positive and Negative Sentiment in *Sense and Sensibility*



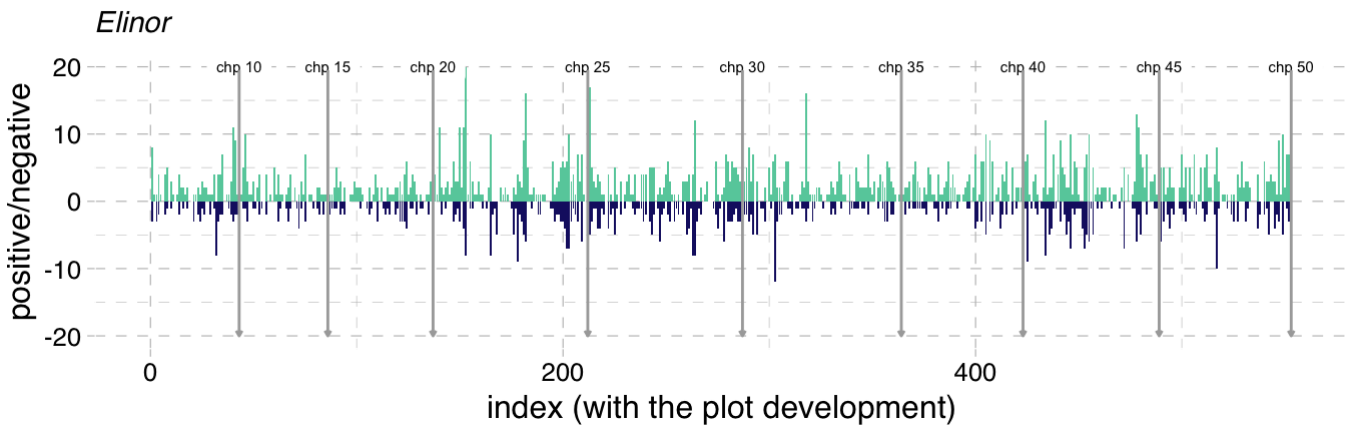
We can match the sentiment words counts with certain important plots in the book, for examples:

- Chp9: more positive — Marianne met Willoughby
- Chp25-26: more positive — Marianne and Elinor left home with Mrs. Jennings, and Marianne wished to meet Willoughby
- Chp33: more negative — Marianne and Elinor not happy with Mrs.Ferras in a party
- Chp43:more negative — Marianne illd seriously for a long time
- Chp47: more negative — Elinor knew Edward had married with Lucy and felt not good
- Chp50: more positive — Marianne and Elinor both got married with loved one

Marianne and Elinor are both positive figures in the *Sense and Sensibility*, they are kind and nice girls. And since this book is a not serious, dark story, Marianne and Elinor have more positive sentiment than negative sentiment in general.

We can also see the same sentiment trends in these following plots, which xlim is divided into more detailed segments.

Positive and Negative Sentiment in *Sense and Sensibility*

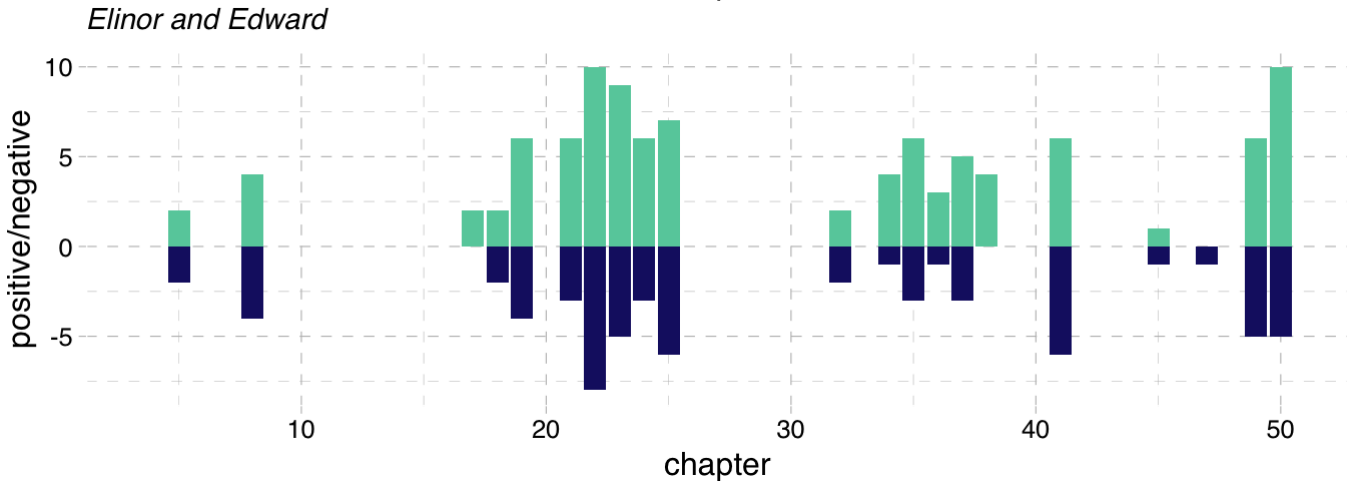
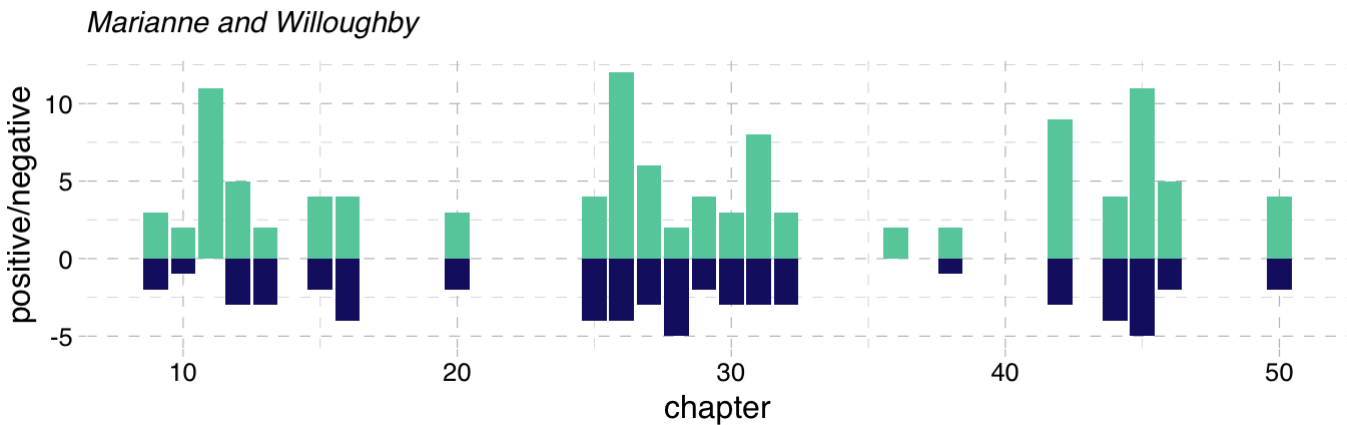


Positive and Negative Sentiment in *Sense and Sensibility*

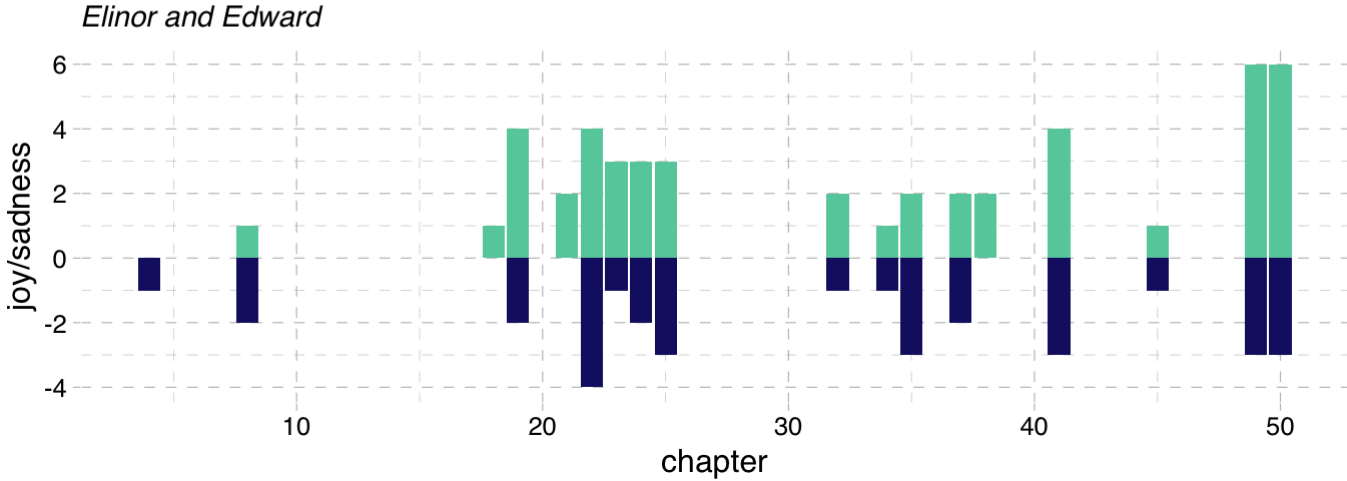
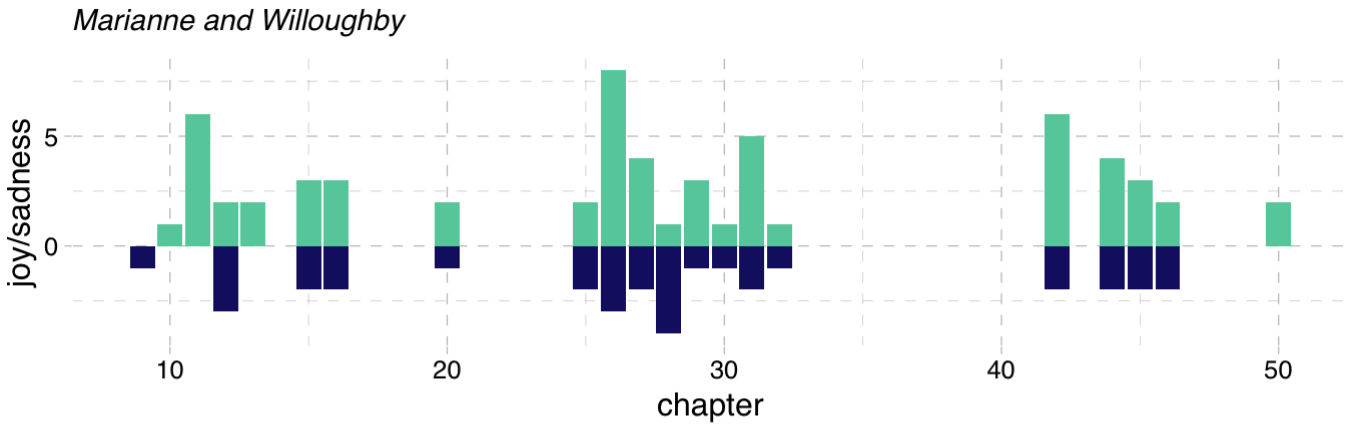


Co-occurence: Marianne & Willoughby | Elinor & Edward

Positive and Negative Sentiment

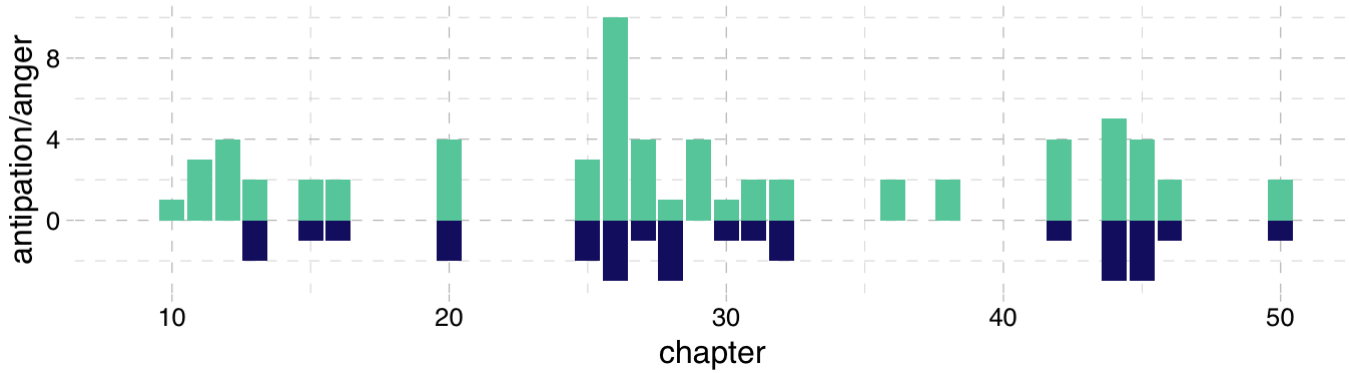


Positive and Negative Sentiment

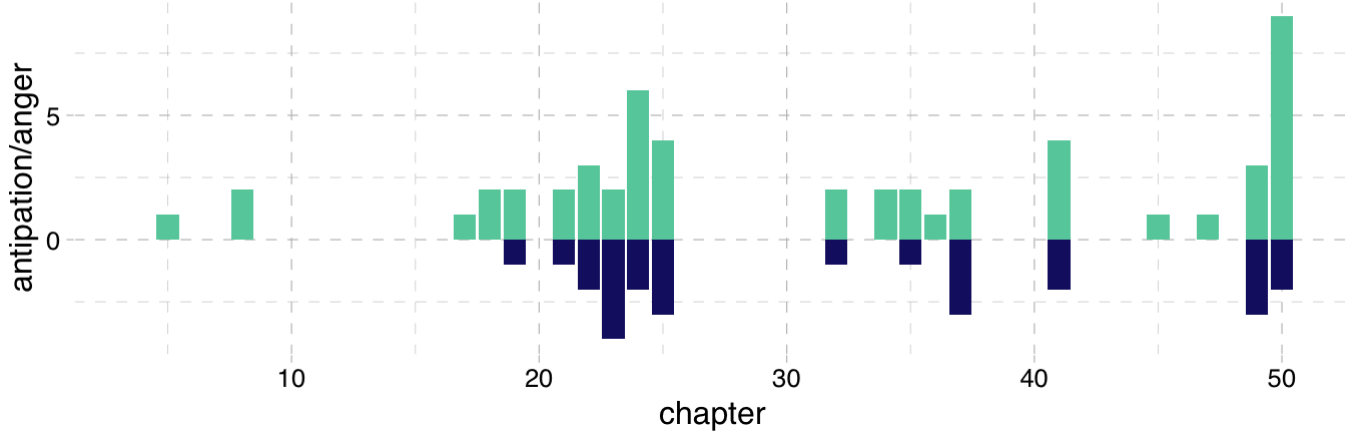


Positive and Negative Sentiment

Marianne and Willoughby

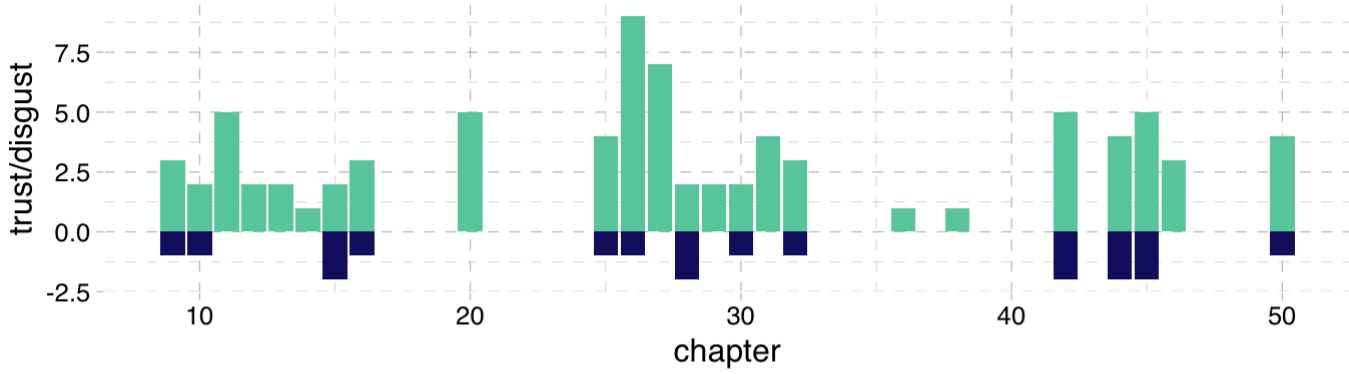


Elinor and Edward

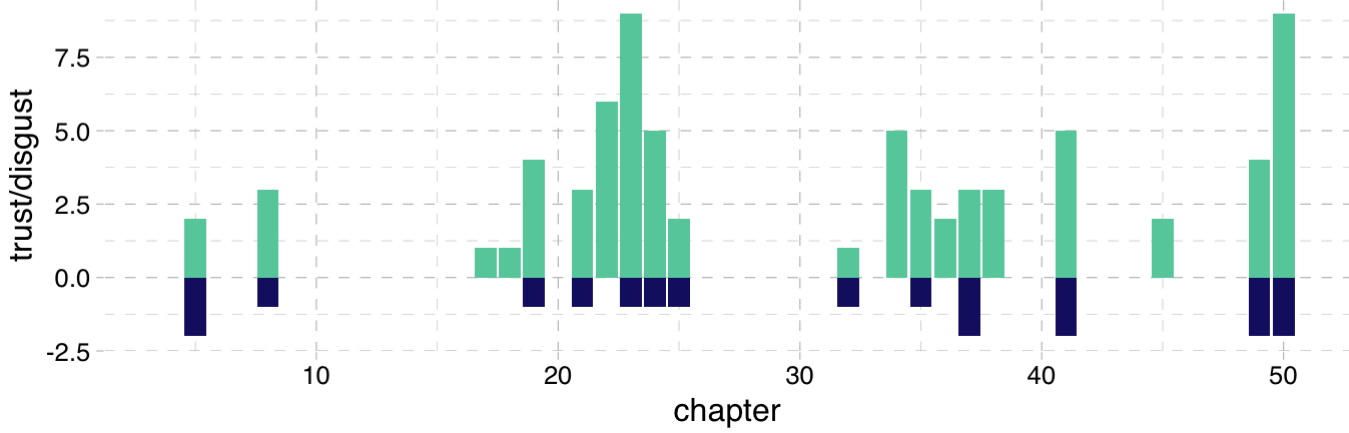


Positive and Negative Sentiment

Marianne and Willoughby



Elinor and Edward



Marianne was attracted to young, handsome, romantically spirited Willoughby.

Elinor become attached to Edward Ferrars, the brother-in-law of her elder half-brother, John.

We can observe some interesting things in these plots. For examples,

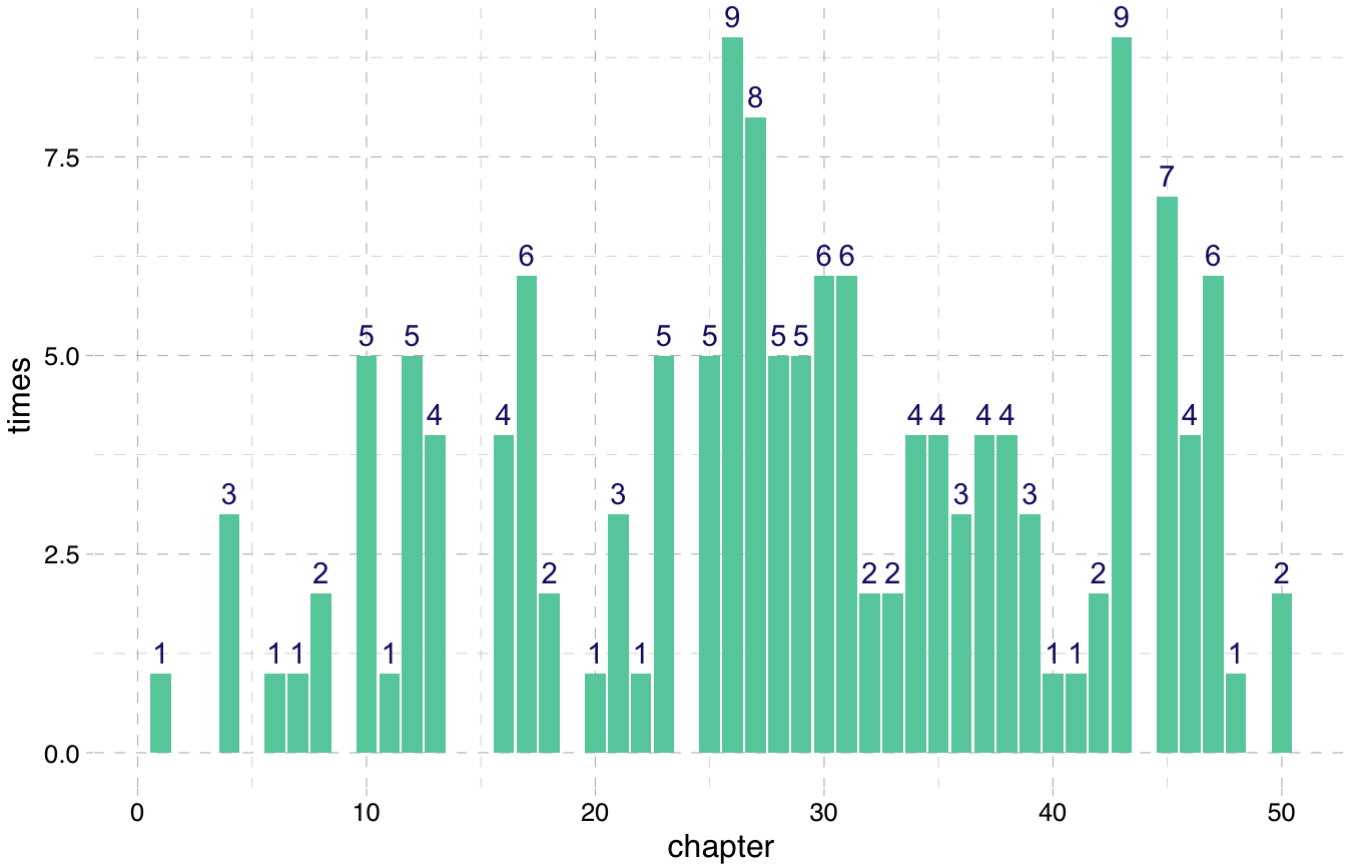
- In the **trust** sentiment plot: Marianne and Willoughby's trust sentiment increased first, then decreased. Elinor and Edward's trust sentiment increased first, then decreased, finally increased a lot. These two observations are consistent with the content of the book. Willoughby let Marianne down in the later period and married Brandon, while Elinor and Edward got married eventually after experiencing some twists and turns.
- In the **joy** sentiment plot: In the final chapter, the joy sentiment words bars are very high since Edward proposed to Elinor and had a happy ending.

There are many other interesting things to be discovered.

Co-occurence of Elinor and Marianne

Marianne and Elinor Co-occurence

chapter distrubution



According to the plot, Elinor and Marianne co-occur most times in chapter 26 and 43. And they co-occur many times in chapter27 and 45 too.

In this book, Marianne and Elinor were always together and supported each other. I would like to use the last paragraph of *Sense and Sensibility* as the end:

- Between Barton and Delaford, there was that constant communication which strong family affection would naturally dictate;—and among the merits and the happiness of Elinor and Marianne, let it not be ranked as the least considerable, that though sisters, and living almost within sight of each other, they could live without disagreement between themselves, or producing coolness between their husbands.*

Acknowledge

- tnum package (<https://github.com/Truenumbers/tnum>)
we use this package for text analysis
- get start with tnum (<https://github.com/Truenumbers/tnum/tree/master/tnum/inst/Rmd>)
- jane austen package (<https://github.com/juliasilge/janeaustenr>)
This package provides access to the full texts of Jane Austen’s 6 completed, published novels.
- ggthemes package (<https://www.rdocumentation.org/packages/ggthemes/versions/3.5.0>)