Exploring and Tagging Text

Chenghao Meng,Chun Gao,Fan Feng,Jiaheng Li,Yinfeng Zhou

2020/11/18

```
# Load the package
pacman:: p_load(tnum,httr,tidyverse,sentimentr,tidytext)
tnum.authorize(ip="54.158.136.133")
```

1 Co-occurence of Elizabeth and Darcy

1.1 Data Preparation and Cleaning

```
# Take a glance at the db
pride_text <- tnum.getDatabasePhraseList("subject",level=3)
#pride_text
```

First, we will tag the sentences with the occurence of Elizabeth and Darcy.

```
# Tag "Elizabeth"
tnum.tagByQuery("*pride* has text= REGEXP(\"Elizabeth\")","reference:Group9Elizabeth" )
```

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

```
# Tag Darcy
tnum.tagByQuery("*pride* has text= REGEXP(\"Darcy\")","reference:Group9Darcy" )
```

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

After that, we will generate a dataframe with the occurence of Darcy based our tag.

Then, we will use the `filter` function in the `tidyverse` package to get the dataframe of sentences with the co-occurence of Elizabeth and Darcy.

```
# Dataframe with the occurance of Darcy
qdarcy <- tnum.query("@reference:Group9Darcy",max=394)
```

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

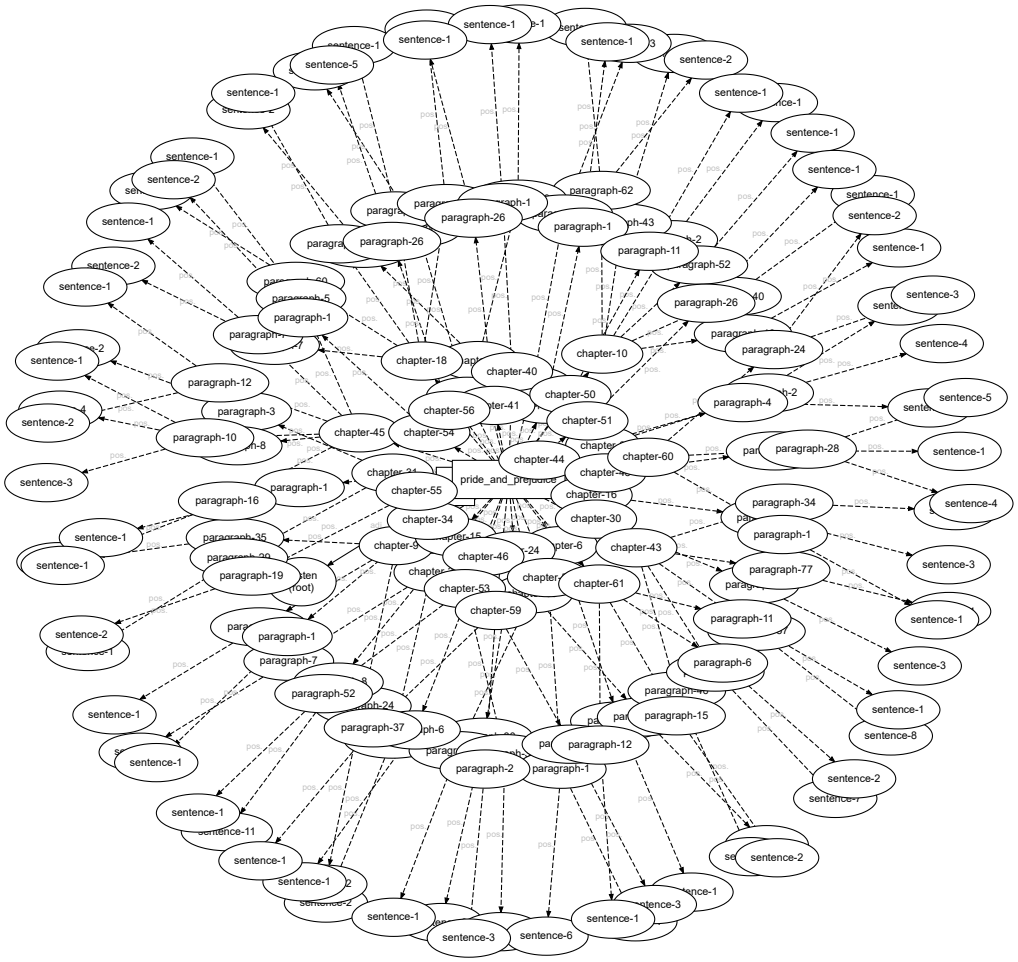
```
darcy_df <- tnum.objectsToDf(qdarcy)
#head(darcy_df)

# Dataframe: Co-occurence of Elizabeth and Darcy
co_occur <- filter(darcy_df,grepl('Group9Elizabeth',tags))
```

1.2 Explortary Data Analysis

1.2.1 Tree Plot

Using the **co_occur** dataframe, a tree plot will be generated to visualize the co-occurence of Elizabeth and Darcy in different chapters and sentences within the *Pride and Prejudice*.

```
# Tree plot
plot_tree_co_occr=tnum.makePhraseGraphFromPathList(co_occur$subject)
tnum.plotGraph(plot_tree_co_occr)
```



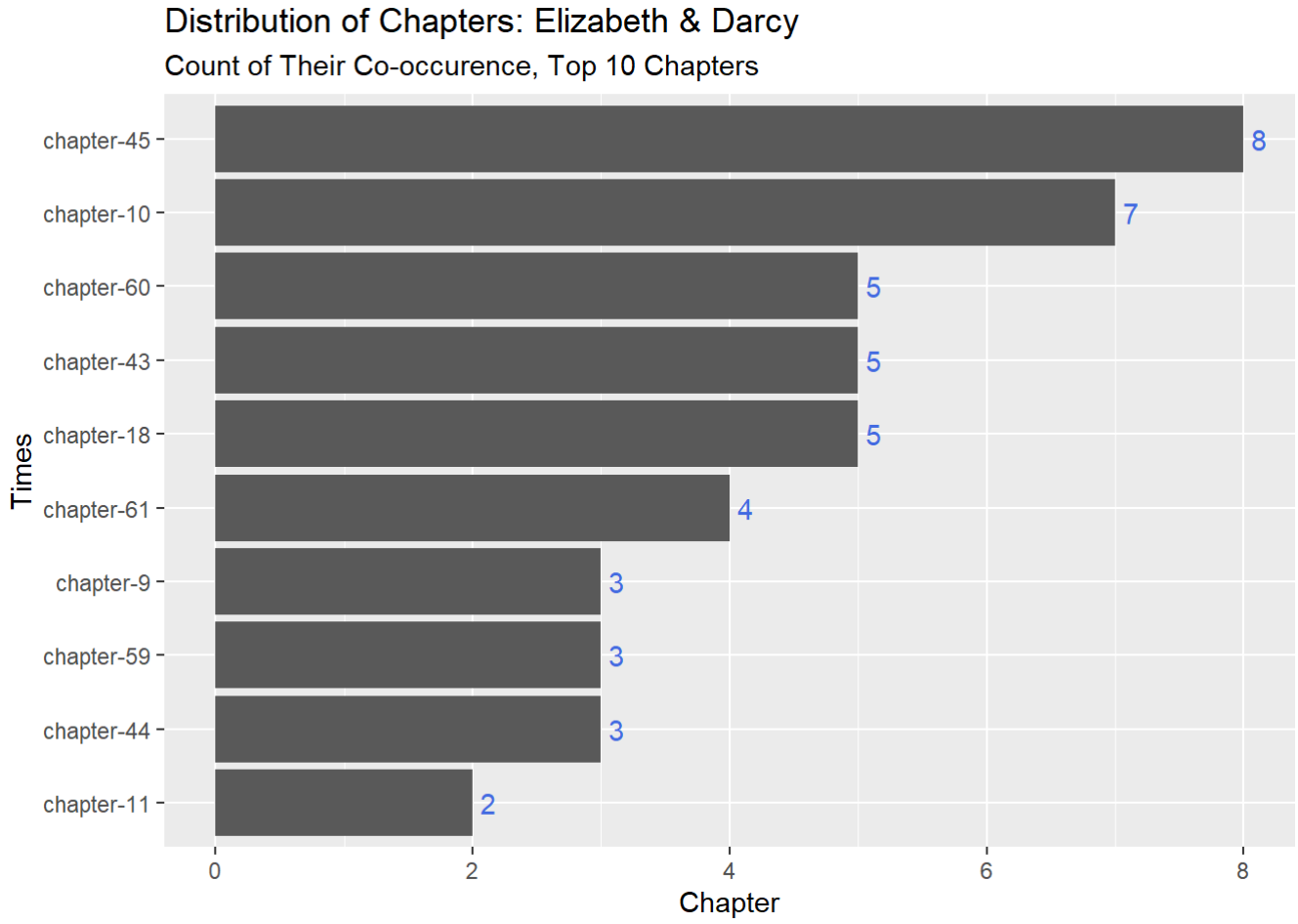
1.2.2 Distribution of Co-occurence

The tree plot gives us a sense that Elizabeth and Darcy have appeared together in a lot of chapters in *Pride and Prejudice*. However, we are still not sure of the distribution of their co-occurence in different chapters. As a result, we will continue exploring the dataframe **co_occur**.

```
# Get the info of chapter,paragraph,sentence
co_occur_info <- co_occur %>%
  separate(subject,c("book","chapter","paragraph","sentence"),sep="/") %>%
  select(book,chapter,paragraph,sentence,string.value)
```

When having the information of the chapter, paragraph and sentence of their co-occurence in the book, we will explore their top 10 co-occurence in different chapters

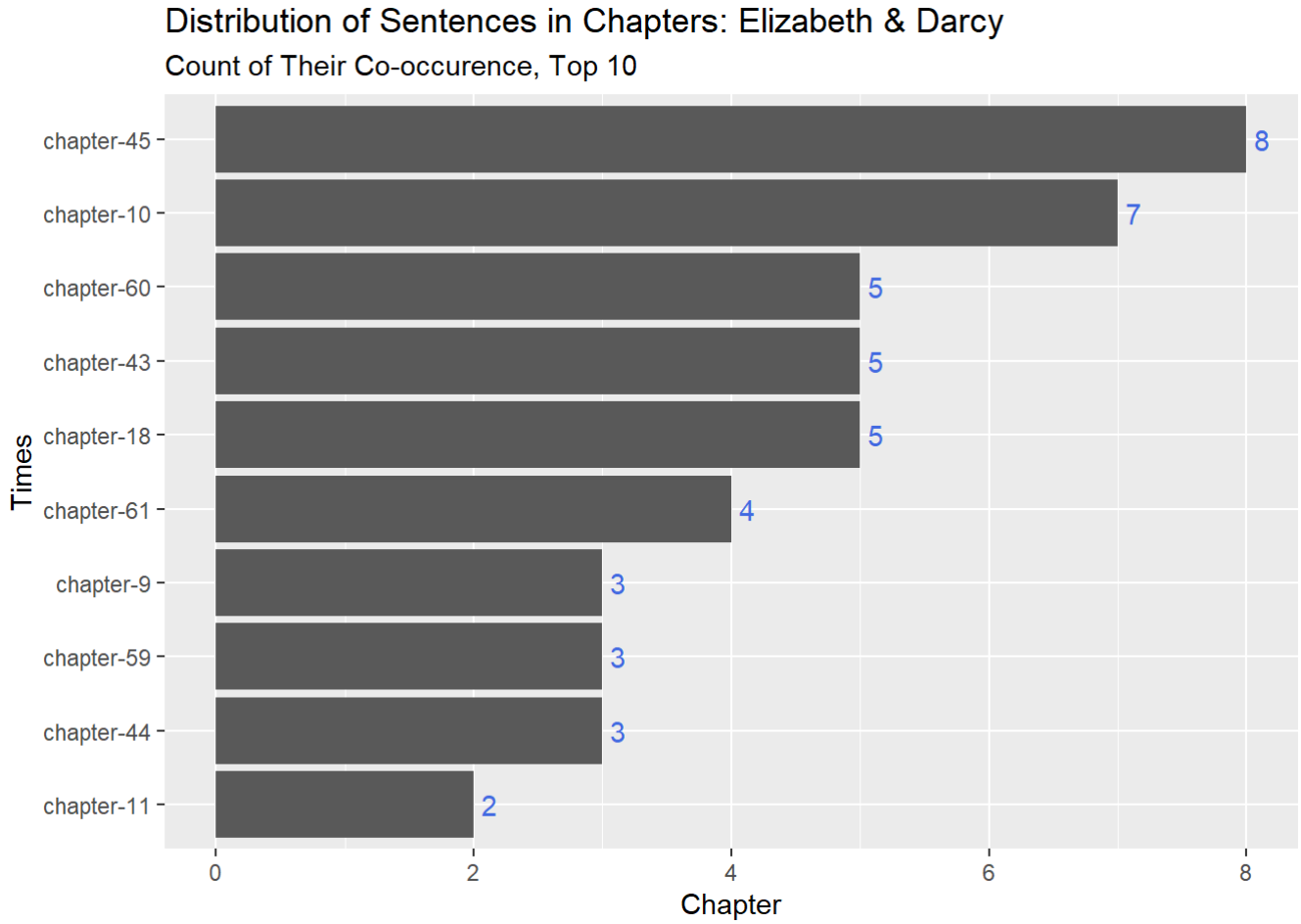
```
# Top 10 co-occurence in chapters
co_occur_info %>% group_by(chapter) %>%
  summarise(times=n()) %>%
  arrange(desc(times)) %>% head(10) %>%

  ggplot() +
  geom_bar(aes(x=reorder(chapter,times),y=times),stat = "identity") +
  geom_text(aes(x=reorder(chapter,times),y=times,label=times),hjust=-0.5,col="royalblue") +
  xlab("Times") + ylab("Chapter") +
  ggtitle("Distribution of Chapters: Elizabeth & Darcy",subtitle = "Count of Their Co-occurenc
e, Top 10 Chapters") +
  coord_flip()
```



The bar plot shows that Elizabeth and Darcy appears together 8 times in chapter 45, which is the largest number of their co-occurence.

We also have the interest to explore how the counts of these sentences with their co-occurence are distributed by chapter.

```
co_occur_info %>% group_by(chapter,paragraph,sentence) %>%
  summarise(times=n()) %>% group_by(chapter)%>%
  summarise(sum_times=sum(times))%>%
  arrange(desc(sum_times)) %>% head(10) %>%


ggplot() +
  geom_bar(aes(x=reorder(chapter,sum_times),y=sum_times),stat = "identity") +
  geom_text(aes(x=reorder(chapter,sum_times),y=sum_times,label=sum_times),hjust=-0.5,col="royal
blue") +
  xlab("Times") + ylab("Chapter") +
  ggtitle("Distribution of Sentences in Chapters: Elizabeth & Darcy",subtitle = "Count of Their
Co-occurence, Top 10") +
  coord_flip()
```



We find out that in each sentence, there is at most 1 co-occurence of `Darcy` and `Elizabeth`. Therefore, the distribution of co-occurence sentences is the same as the previous distribution.

2 Money-related Information

After exploring the co-occurence of Elizabeth and Darcy, we would also like to explore the money-related information in the *Pride and Prejudice*.

2.1 Data Preparation

We will use the regular expression to get the sentences relating to money in the query to produce a dataframe.

```
# Dataframe-money_df
#tnum.tagByQuery("*pride* has text == REGEXP(\"money\")", "reference:Group9money")
qmoney <- tnum.query("*pride* has * = REGEXP(\"money\")",max=1000)
```

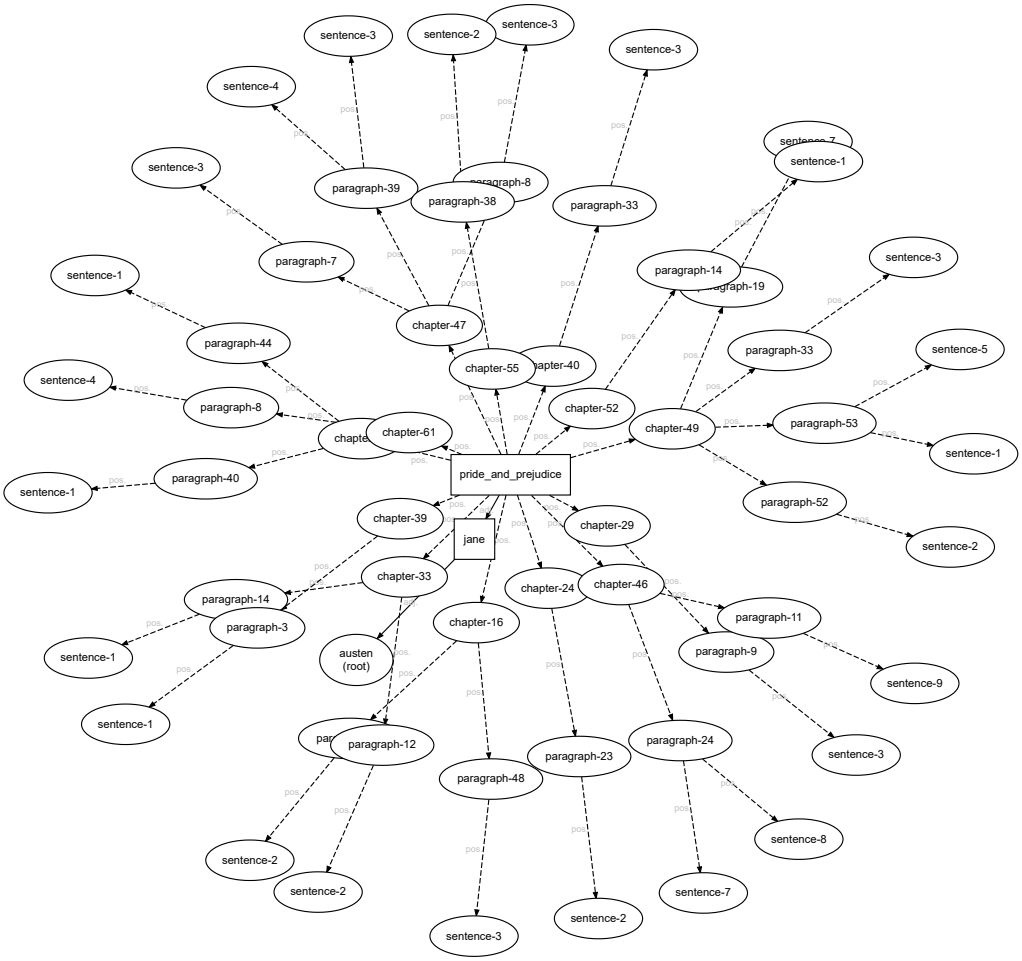
```
## Returned 1 thru 25 of 25 results
```

```
money_df <- tnum.objectsToDf(qmoney)
```

2.2 Explortary Data Analysis

2.2.1 Tree Plot

```
#Tree plot of money

# plot_tree_money <- tnum.makePhraseGraphFromPathList(tnum.getAttrFromList(qmoney,"subject"))
plot_tree_money<-tnum.makePhraseGraphFromPathList(money_df$subject)
tnum.plotGraph(plot_tree_money)
```



The tree plot shows that money seems not to be a word that occurs many times in the book. Only 13 chapters contain the word “money”.

2.2.2 Sentiment Score

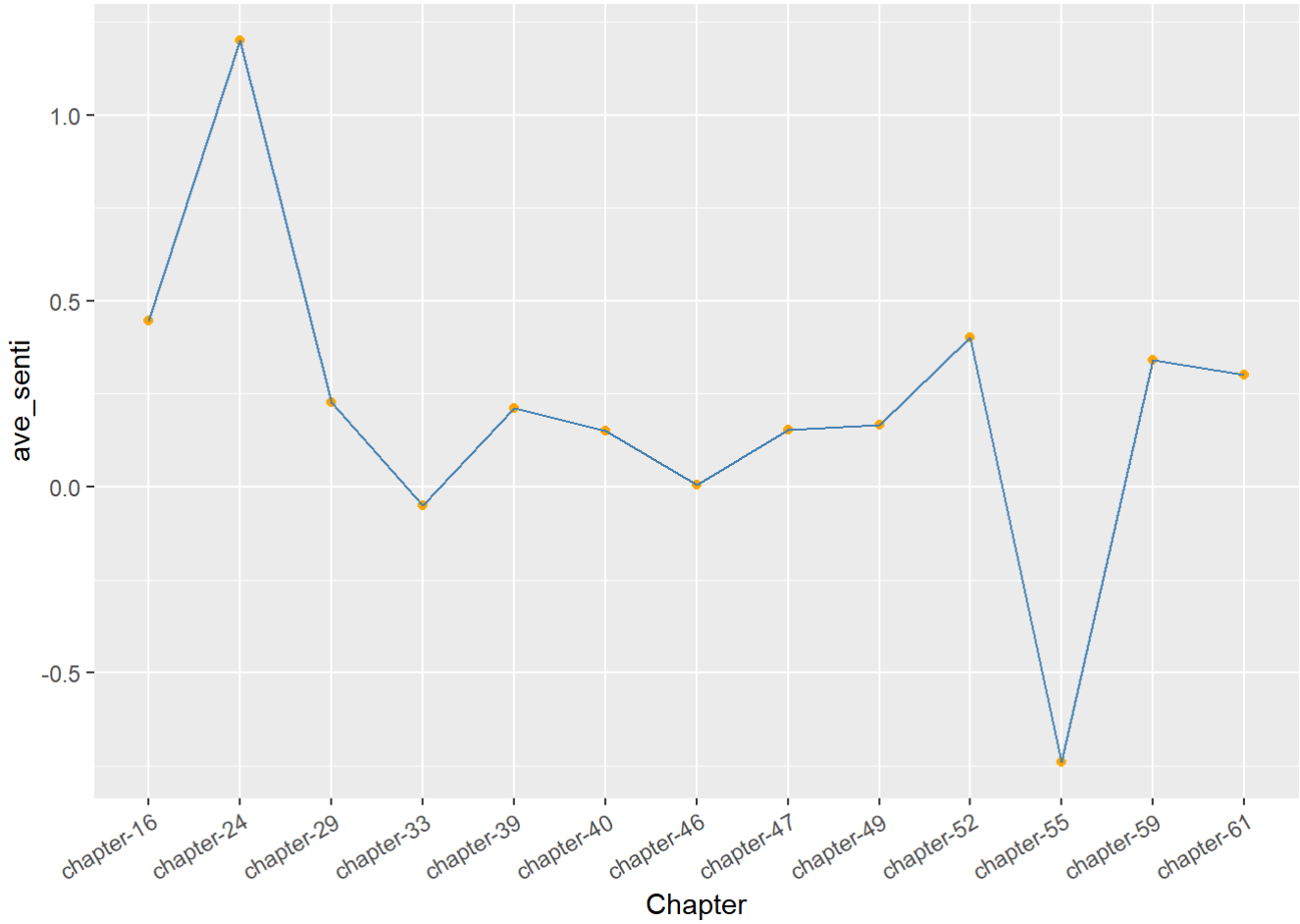
Moreover, we would also like to know what is the sentiment when the word money appears in the sentence. And we will use the `sentiment_by` function in `sentimentr` package to compute the sentiment score of each sentence.

```
money_df2 <- money_df %>%
  separate(subject,c("book","chapter","paragraph","sentence"),sep="/") %>%
  select(chapter,paragraph,sentence,string.value) %>%
  rename(text=string.value) %>%
  # Delete stop words
  filter(!text %in% stop_words$word)

# Delete the "" in the text
money_df2$text <- gsub("\"","",money_df2$text)
```

```
# Insert an empty column-senti_score
money_df2$senti_score <- rep(NA,times=nrow(money_df2))

# Compute sentiment score
for (i in 1:nrow(money_df2)) {
  sentences <- get_sentences(money_df2$text[i])
  money_df2$senti_score[i] <- sentiment_by(sentences)$ave_sentiment
}
```

After computing the sentiment score, we will draw a line chart to visualize the trend of sentiment score across the chapters that the word money appears.

```
money_senti <- money_df2 %>% group_by(chapter) %>%
  summarise(ave_senti=mean(senti_score)) %>%
  arrange(desc(ave_senti))

# Plot the sentiment score
ggplot(data = money_senti,aes(x=chapter,y=ave_senti,group=1)) +
  geom_point(col="orange")+
  geom_line(col="steelblue")+
  xlab("Chapter")  +theme(axis.text.x = element_text(angle = 30, hjust = 1))
```




```
  ggtitle("Sentiment Trend of Money-appearing Chapter")
```

```
## $title
## [1] "Sentiment Trend of Money-appearing Chapter"
##
## attr(,"class")
## [1] "labels"
```