HW3

2019020336 배현주

10/13/2019

```
library(janeaustenr)
library(tidytext)
library(dplyr)
library(tidyr)
library(stringr)
library(scales)
library(ggplot2)
library(corrplot)
library(tidyverse)
library(textdata)
austen <- austen_books()
```

1. Conduct preprocessing including tokenization (using unnest_tokens) and removing stopwords (using data(stop_words)).

```
austen_pre <- austen %>%
  group_by(book) %>%
  mutate(linenumber = row_number(),
         chapter = cumsum(str_detect(text,
                                     regex("^chapter [\\divslc]",
                                           ignore_case = TRUE)))) %>%
  ungroup() %>%
  unnest_tokens(word, text) %>%
  mutate(word = str_extract(word, "[a-z']+")) %>%
  filter(!is.na(word)) %>%
  anti_join(stop_words)

Joining, by = "word"

head(austen_pre)

# A tibble: 6 x 4
  book                linenumber chapter word
  <fct>                    <int>   <int> <chr>
1 Sense & Sensibility          1       0 sense
2 Sense & Sensibility          1       0 sensibility
3 Sense & Sensibility          3       0 jane
4 Sense & Sensibility          3       0 austen
```

```
5 Sense & Sensibility          10       1 chapter
6 Sense & Sensibility          13       1 family
```

2. Caculate the term-document maxtrix whose column is novel (Document), row is word, and value is word frequency.

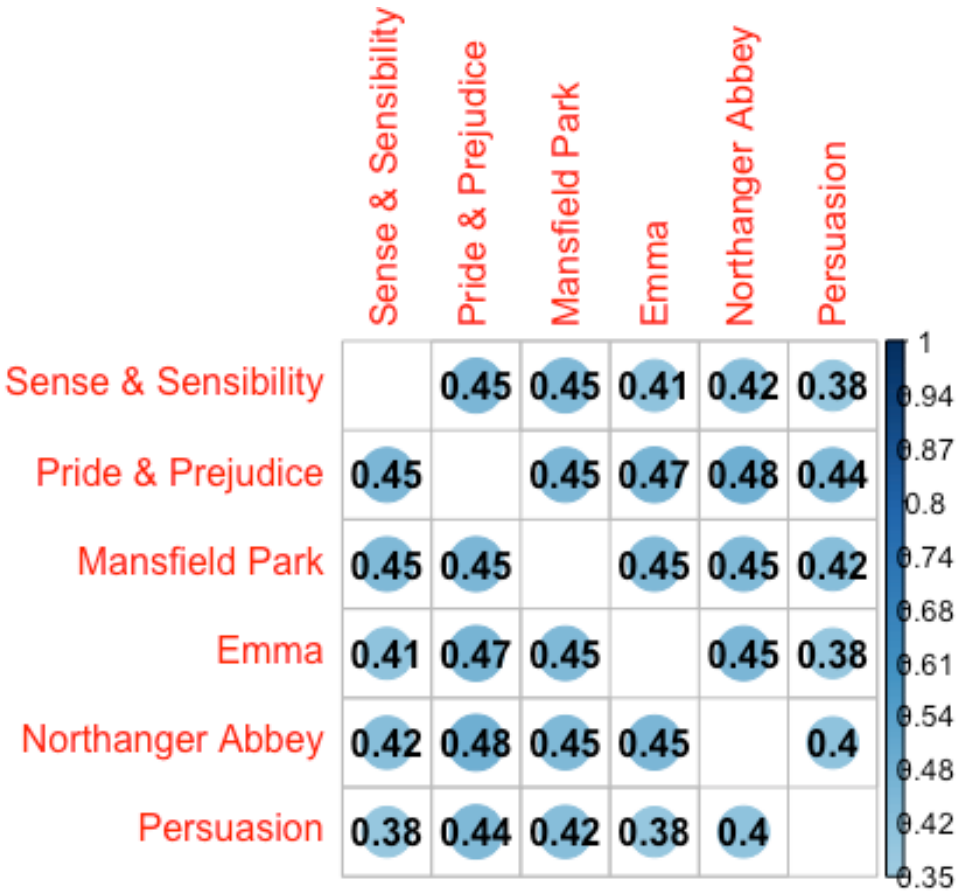
```
austen_freq <- austen_pre %>%
  group_by(book) %>%
  count(word, sort=TRUE) %>%
  spread(key="book", value="n", fill=0)
head(austen_freq)
```

```
# A tibble: 6 x 7
  word  `Sense & Sensib… `Pride & Prejud… `Mansfield Park`  Emma
  <chr>            <dbl>            <dbl>             <dbl> <dbl>
1 a'n't                0                0                 0     0
2 aban…                1                0                 0     0
3 abas…                0                0                 1     0
4 abate                0                0                 2     0
5 abat…                1                1                 1     0
6 abat…                0                0                 1     0
# … with 2 more variables: `Northanger Abbey` <dbl>, Persuasion <dbl>
```

3. Given the term-document maxtrix, each novel is represented as a vector (which is sparse). Find two-most similar and different novels. Justify your answers.

- correlation matrix

```
austen_corr <- austen_freq %>%
  column_to_rownames("word") %>%
  as.matrix() %>%
  cor()
corrplot(austen_corr, cl.lim = c(0.35, 1),
         addCoef.col = "black", diag = FALSE)
```



By calculating the correlation matrix of the word frequency in each books, the most similar books are 'Pride & Prejudice' and 'Northanger Abbey'. On the other hand, the most different books are 'Sense & Sensibility' and 'Persuasion'.

- sentiment analysis plot (nrc)

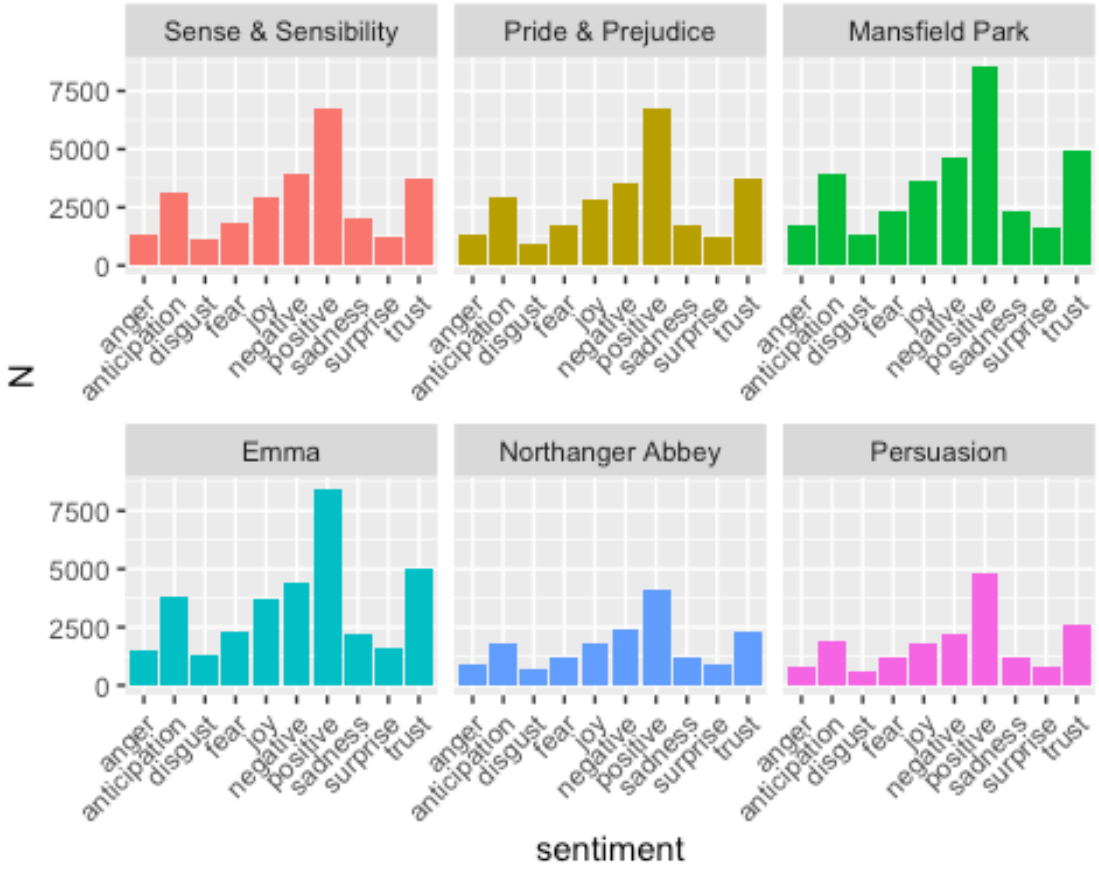
```
austen_nrc <- austen_pre %>%
  inner_join(get_sentiments("nrc")) %>%
  count(book, index = linenumber %% 80, sentiment) %>%
  group_by(book, sentiment) %>%
  summarise(N = sum(n))

austen_nrc %>%
  group_by(book) %>%
  filter(N == min(N) | N == max(N))

# A tibble: 12 x 3
# Groups:   book [6]
   book                sentiment     N
   <fct>               <chr>     <int>
 1 Sense & Sensibility disgust    1160
 2 Sense & Sensibility positive   6739
 3 Pride & Prejudice   disgust     966
```

```
 4 Pride & Prejudice   positive   6792
 5 Mansfield Park      disgust    1310
 6 Mansfield Park      positive   8542
 7 Emma                disgust    1310
 8 Emma                positive   8468
 9 Northanger Abbey    disgust     697
10 Northanger Abbey    positive   4094
11 Persuasion          disgust     641
12 Persuasion          positive   4814
```

```
ggplot(austen_nrc, aes(sentiment, N, fill=book)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ book, ncol = 3, scales = "free_x") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



By using “nrc” sentiment lexicon, it is possible to draw the distribution of sentiments in books. ‘Disgust’ & ‘Positive’ are common lowest & highest counts in books. By comparing the distribution of each books (usually based on the peak points), “Mansfield Park” and “Emma” are most similar books. On the other hand, “Mansfield Park” and “Persuasion” are most different books.