# Text Mining Home Work 2

財金二 _ 賴冠霖 _410936051

## Contents

# 1 Introduction

Perform topic modeling for 4 books:

1. The Adventures of Tom Sawyer, by Mark Twain
2. Little Women by Louisa May Alcott
3. The Time Machine by H. G. Wells
4. Sense and Sensibility by Jane Austen

# 2 Download these four books from Project Gutenbergr

```
library(magrittr)
library(tidytext)
library(stringr) #str_detect
library(tidyr)
library(tibble)
library(jiebaR)
library(scales)
library(ggplot2)
library(highcharter)
library(wordcloud2)
library(wordcloud)
library(dplyr)
library(gutenbergr)#get books' data
library(broom)

# 1. Download these four books from Project Gutenbergr

titles <- c("The Adventures of Tom Sawyer", "Little Women",
            "The Time Machine", "Sense and Sensibility")
books <- gutenberg_works(title %in% titles) %>%
  gutenberg_download(meta_fields = "title")
```

# 3 Divide these four books into chapters and treat them as documents.

```
#regex() : regular expression
# 2. Divide these four books into chapters and treat them as documents.

by_chapter <- books %>%
  group_by(title) %>% #4 本書分開
  mutate(chapter = cumsum(str_detect(text, regex("^chapter ", ignore_case = TRUE)))) %>%
  # 字頭有 chapter 抓出來 (str_detect) 的累計次數 (cusum)
  ungroup() %>% # 取消分開 4 本書
  filter(chapter > 0) # 有 Chapter 才要

by_chapter_word <- by_chapter %>%
```

```
  unite(title_chapter, title, chapter) %>% #
  unnest_tokens(word, text) #split into words
```

# 4 Remove stop_words.

```
# 3. Remove stop_words.
word_counts <- by_chapter_word %>%
  anti_join(stop_words) %>% #remove stop_words
  count(title_chapter, word, sort = TRUE) %>% #Count number of words
ungroup()
```

# 5 Count number of words in each document.

```
# 4. Count number of words in each document.
word_counts
```

```
## # A tibble: 84,371 x 3
##    title_chapter                    word      n
##    <chr>                            <chr> <int>
##  1 Little Women_9                   meg      70
##  2 Little Women_8                   jo       61
##  3 Little Women_21                  jo       60
##  4 The Adventures of Tom Sawyer_66 tom      58
##  5 Little Women_12                  jo       57
##  6 Little Women_28                  john     57
##  7 Little Women_32                  jo       57
##  8 Little Women_43                  jo       56
##  9 Little Women_3                   jo       53
## 10 The Adventures of Tom Sawyer_41 tom      53
## # ... with 84,361 more rows
```

# 6 Convert the document into a document-term-matrix.

```
# 5. Convert the document into a document-term-matrix.
chapters_dtm <- word_counts %>%
  cast_dtm(title_chapter, word, n)

chapters_dtm
```

```
## <<DocumentTermMatrix (documents: 167, terms: 15489)>>
## Non-/sparse entries: 84371/2502292
## Sparsity           : 97%
## Maximal term length: 20
## Weighting          : term frequency (tf)
```

## 7 Use the topicmodels package to create a four topic LDA model.

```
# 6. Use the topicmodels package to create a four topic LDA model.
library(topicmodels)
chapters_lda <- LDA(chapters_dtm, k = 4, control = list(seed = 1234))
chapters_lda
```

```
## A LDA_VEM topic model with 4 topics.
```

```
chapters_lda_td <- tidy(chapters_lda)
chapters_lda_td
```

```
## # A tibble: 61,956 x 3
##    topic term      beta
##    <int> <chr>    <dbl>
##  1     1 meg   6.42e-52
##  2     2 meg   1.56e- 2
##  3     3 meg   6.66e- 3
##  4     4 meg   9.85e-25
##  5     1 jo    3.61e-33
##  6     2 jo    1.97e- 2
##  7     3 jo    1.84e- 2
##  8     4 jo    1.85e-16
##  9     1 tom   9.00e-53
## 10     2 tom   6.51e-11
## # ... with 61,946 more rows
```

## 8 Find the top 5 terms within each topic.

```
# 7. Find the top 5 terms within each topic.
top_terms <- chapters_lda_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_terms
```

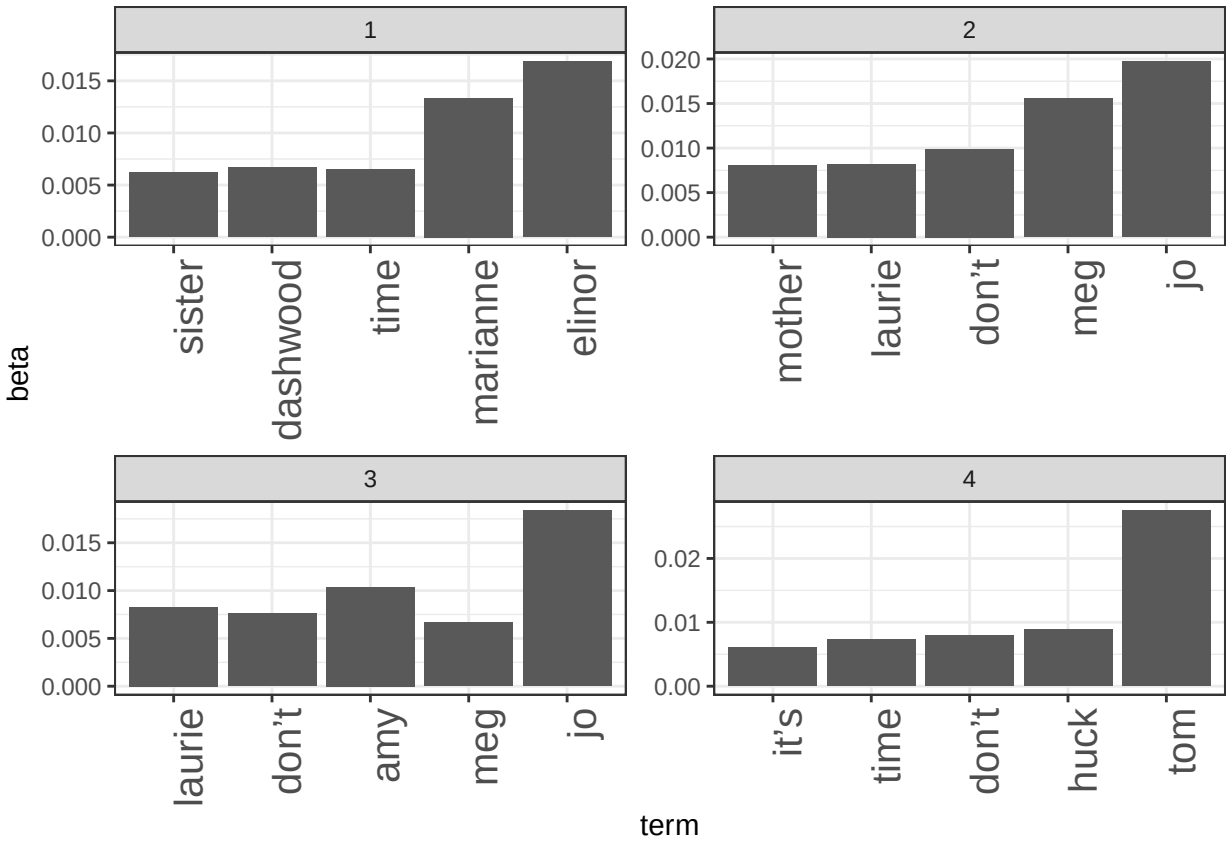
```
## # A tibble: 20 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 elinor   0.0169
##  2     1 marianne 0.0133
##  3     1 dashwood 0.00667
##  4     1 time     0.00648
##  5     1 sister   0.00621
##  6     2 jo       0.0197
##  7     2 meg      0.0156
```

```
##  8     2 don' t    0.00989
##  9     2 laurie   0.00822
## 10     2 mother   0.00801
## 11     3 jo       0.0184
## 12     3 amy      0.0104
## 13     3 laurie   0.00829
## 14     3 don' t    0.00757
## 15     3 meg      0.00666
## 16     4 tom      0.0275
## 17     4 huck     0.00886
## 18     4 don' t    0.00798
## 19     4 time     0.00727
## 20     4 it' s     0.00607
```

# 9 Visualize the word count for each topic.

```
# 8. Visualize the word count for each topic.
library(ggplot2)
theme_set(theme_bw())

top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ topic, scales = "free") +
  theme(axis.text.x = element_text(size = 15, angle = 90, hjust = 1))
```




## 10 Find out which topics are associated with each document and check if we could put the chapters back together in the correct books.

```
# 9. Find out which topics are associated with each document
#and check if we could put the chapters back together in the correct books.
chapters_lda_gamma <- tidy(chapters_lda, matrix = "gamma")
chapters_lda_gamma
```

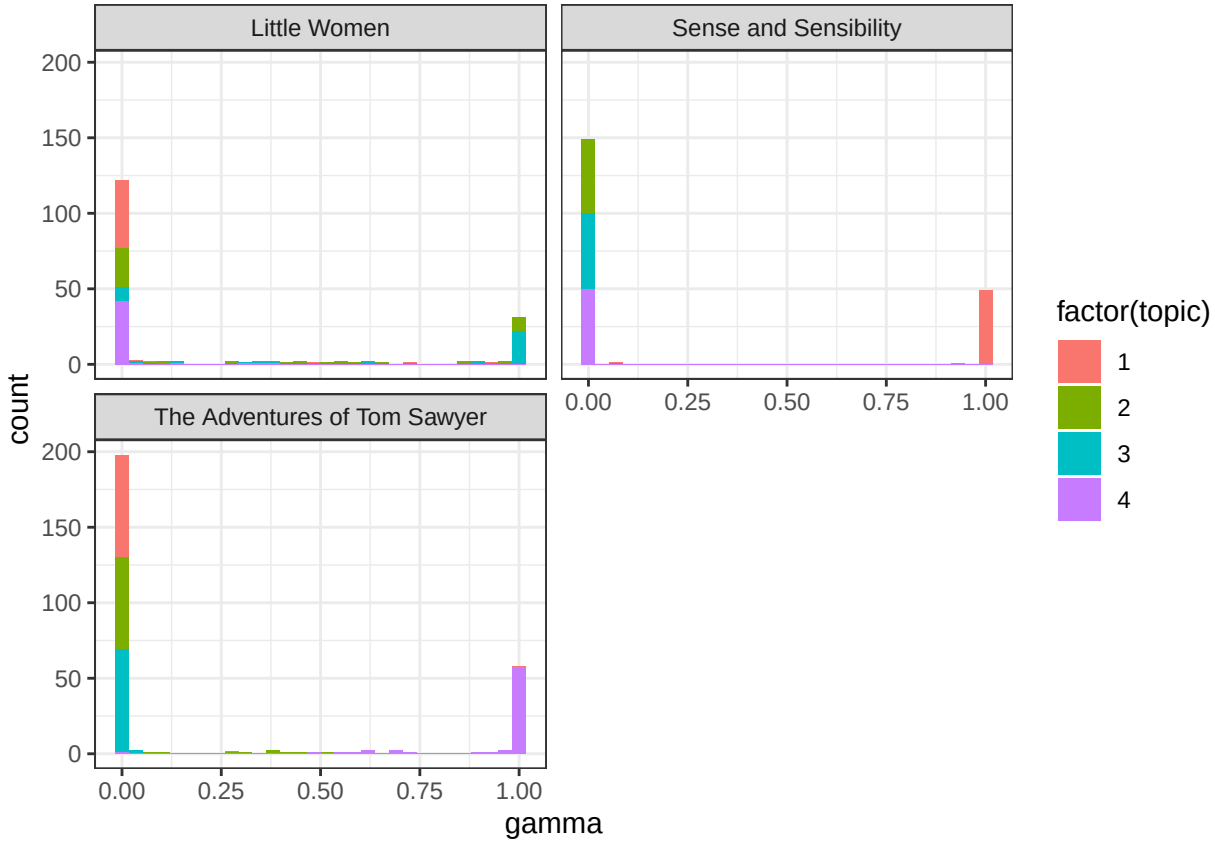
```
## # A tibble: 668 x 3
##    document                          topic      gamma
##    <chr>                             <int>      <dbl>
##  1 Little Women_9                        1 0.00000899
##  2 Little Women_8                        1 0.0000136
##  3 Little Women_21                       1 0.0000130
##  4 The Adventures of Tom Sawyer_66       1 0.0000187
##  5 Little Women_12                       1 0.00000722
##  6 Little Women_28                       1 0.00000976
##  7 Little Women_32                       1 0.0000133
##  8 Little Women_43                       1 0.00000962
##  9 Little Women_3                        1 0.0000143
## 10 The Adventures of Tom Sawyer_41       1 0.0000145
```

```
## # ... with 658 more rows
```

```
chapters_lda_gamma <- chapters_lda_gamma %>%
  separate(document, c("title", "chapter"), sep = "_", convert = TRUE)
chapters_lda_gamma
```

```
## # A tibble: 668 x 4
##    title                          chapter topic      gamma
##    <chr>                            <int> <int>      <dbl>
##  1 Little Women                         9     1 0.00000899
##  2 Little Women                         8     1 0.0000136
##  3 Little Women                        21     1 0.0000130
##  4 The Adventures of Tom Sawyer        66     1 0.0000187
##  5 Little Women                        12     1 0.00000722
##  6 Little Women                        28     1 0.00000976
##  7 Little Women                        32     1 0.0000133
##  8 Little Women                        43     1 0.00000962
##  9 Little Women                         3     1 0.0000143
## 10 The Adventures of Tom Sawyer        41     1 0.0000145
## # ... with 658 more rows
```

```
ggplot(chapters_lda_gamma, aes(gamma, fill = factor(topic))) +
  geom_histogram() +
  facet_wrap(~ title, nrow = 2)
```



```
chapter_classifications <- chapters_lda_gamma %>%
  group_by(title, chapter) %>%
  top_n(1, gamma) %>%
  ungroup() %>%
  arrange(gamma)

chapter_classifications
```

```
## # A tibble: 167 x 4
##    title                        chapter topic gamma
##    <chr>                          <int> <int> <dbl>
##  1 Little Women                      45     3 0.404
##  2 Little Women                       9     3 0.504
##  3 The Adventures of Tom Sawyer      43     2 0.508
##  4 The Adventures of Tom Sawyer      48     4 0.549
##  5 Little Women                      32     3 0.554
##  6 Little Women                      24     2 0.561
##  7 Little Women                      27     3 0.578
##  8 The Adventures of Tom Sawyer      51     4 0.586
##  9 The Adventures of Tom Sawyer       9     4 0.604
## 10 Little Women                      34     3 0.605
## # ... with 157 more rows
```

```
book_topics <- chapter_classifications %>%
  count(title, topic) %>%
  top_n(1, n) %>%
  ungroup() %>%
  transmute(consensus = title, topic)

book_topics
```

```
## # A tibble: 1 x 2
##   consensus                   topic
##   <chr>                       <int>
## 1 The Adventures of Tom Sawyer     4
```

```
#See which chapters were misidentified
chapter_classifications %>%
  inner_join(book_topics, by = "topic") %>%
  count(title, consensus)
```

```
## # A tibble: 1 x 3
##   title                       consensus                       n
##   <chr>                       <chr>                       <int>
## 1 The Adventures of Tom Sawyer The Adventures of Tom Sawyer    68
```

```
assignments <- augment(chapters_lda, data = chapters_dtm)

assignments <- assignments %>%
  separate(document, c("title", "chapter"), sep = "_", convert = TRUE) %>%
  inner_join(book_topics, by = c(".topic" = "topic"))

assignments
```

```
## # A tibble: 18,025 x 6
##    title                        chapter term  count .topic consensus
##    <chr>                          <int> <chr> <dbl>  <dbl> <chr>
##  1 The Adventures of Tom Sawyer      66 tom      58      4 The Adventures of To~
##  2 The Adventures of Tom Sawyer      41 tom      53      4 The Adventures of To~
##  3 The Adventures of Tom Sawyer      53 tom      44      4 The Adventures of To~
##  4 The Adventures of Tom Sawyer      68 tom      40      4 The Adventures of To~
##  5 The Adventures of Tom Sawyer      45 tom      36      4 The Adventures of To~
##  6 The Adventures of Tom Sawyer      36 tom      31      4 The Adventures of To~
##  7 The Adventures of Tom Sawyer      70 tom      31      4 The Adventures of To~
##  8 The Adventures of Tom Sawyer      39 tom      28      4 The Adventures of To~
##  9 The Adventures of Tom Sawyer      42 tom      27      4 The Adventures of To~
## 10 The Adventures of Tom Sawyer      51 tom      23      4 The Adventures of To~
## # ... with 18,015 more rows
```

```
assignments %>%
  count(title, consensus, wt = count) %>%
  spread(consensus, n, fill = 0)
```

```
## # A tibble: 2 x 2
##   title                        `The Adventures of Tom Sawyer`
##   <chr>                                                 <dbl>
## 1 Little Women                                           1158
## 2 The Adventures of Tom Sawyer                          25008
```

```
wrong_words <- assignments %>%
  filter(title != consensus)

wrong_words
```

```
## # A tibble: 986 x 6
##    title        chapter term  count .topic consensus
##    <chr>          <int> <chr> <dbl>  <dbl> <chr>
##  1 Little Women      34 boys      4      4 The Adventures of Tom Sawyer
##  2 Little Women       7 boys      1      4 The Adventures of Tom Sawyer
##  3 Little Women      45 boys      2      4 The Adventures of Tom Sawyer
##  4 Little Women      45 it' s     4      4 The Adventures of Tom Sawyer
##  5 Little Women      34 boy       2      4 The Adventures of Tom Sawyer
##  6 Little Women      27 boy       1      4 The Adventures of Tom Sawyer
##  7 Little Women      45 boy       4      4 The Adventures of Tom Sawyer
##  8 Little Women      45 i' ll     1      4 The Adventures of Tom Sawyer
##  9 Little Women      45 aunt      4      4 The Adventures of Tom Sawyer
## 10 Little Women      34 book      2      4 The Adventures of Tom Sawyer
## # ... with 976 more rows
```

```
wrong_words %>%
  count(title, consensus, term, wt = count) %>%
  ungroup() %>%
  arrange(desc(n))
```

```
## # A tibble: 881 x 4
##    title        consensus                  term       n
```

```
##    <chr>        <chr>                       <chr>  <dbl>
##  1 Little Women The Adventures of Tom Sawyer book      10
##  2 Little Women The Adventures of Tom Sawyer school     9
##  3 Little Women The Adventures of Tom Sawyer boy        7
##  4 Little Women The Adventures of Tom Sawyer boys       7
##  5 Little Women The Adventures of Tom Sawyer legs       7
##  6 Little Women The Adventures of Tom Sawyer eye        6
##  7 Little Women The Adventures of Tom Sawyer grew       6
##  8 Little Women The Adventures of Tom Sawyer prize      6
##  9 Little Women The Adventures of Tom Sawyer tale       6
## 10 Little Women The Adventures of Tom Sawyer worth      6
## # ... with 871 more rows
```

```
word_counts %>%
  filter(word == "book")
```

```
## # A tibble: 44 x 3
##    title_chapter                    word      n
##    <chr>                            <chr> <int>
##  1 The Adventures of Tom Sawyer_55 book     16
##  2 Little Women_27                 book      8
##  3 Little Women_8                  book      6
##  4 Little Women_12                 book      5
##  5 Little Women_33                 book      5
##  6 The Adventures of Tom Sawyer_39 book      5
##  7 Little Women_2                  book      4
##  8 Little Women_28                 book      4
##  9 The Adventures of Tom Sawyer_43 book      4
## 10 Little Women_13                 book      3
## # ... with 34 more rows
```