Chapter 6

John Peach

4/1/2021

```
library(topicmodels)
data("AssociatedPress")
AssociatedPress
```

```
## <<DocumentTermMatrix (documents: 2246, terms: 10473)>>
## Non-/sparse entries: 302031/23220327
## Sparsity           : 99%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

```
ap_lda <- LDA(AssociatedPress, k = 2, control = list(seed = 1234))
ap_lda
```

```
## A LDA_VEM topic model with 2 topics.
```

```
library(tidytext)
ap_topics <- tidy(ap_lda, matrix = "beta")
ap_topics
```

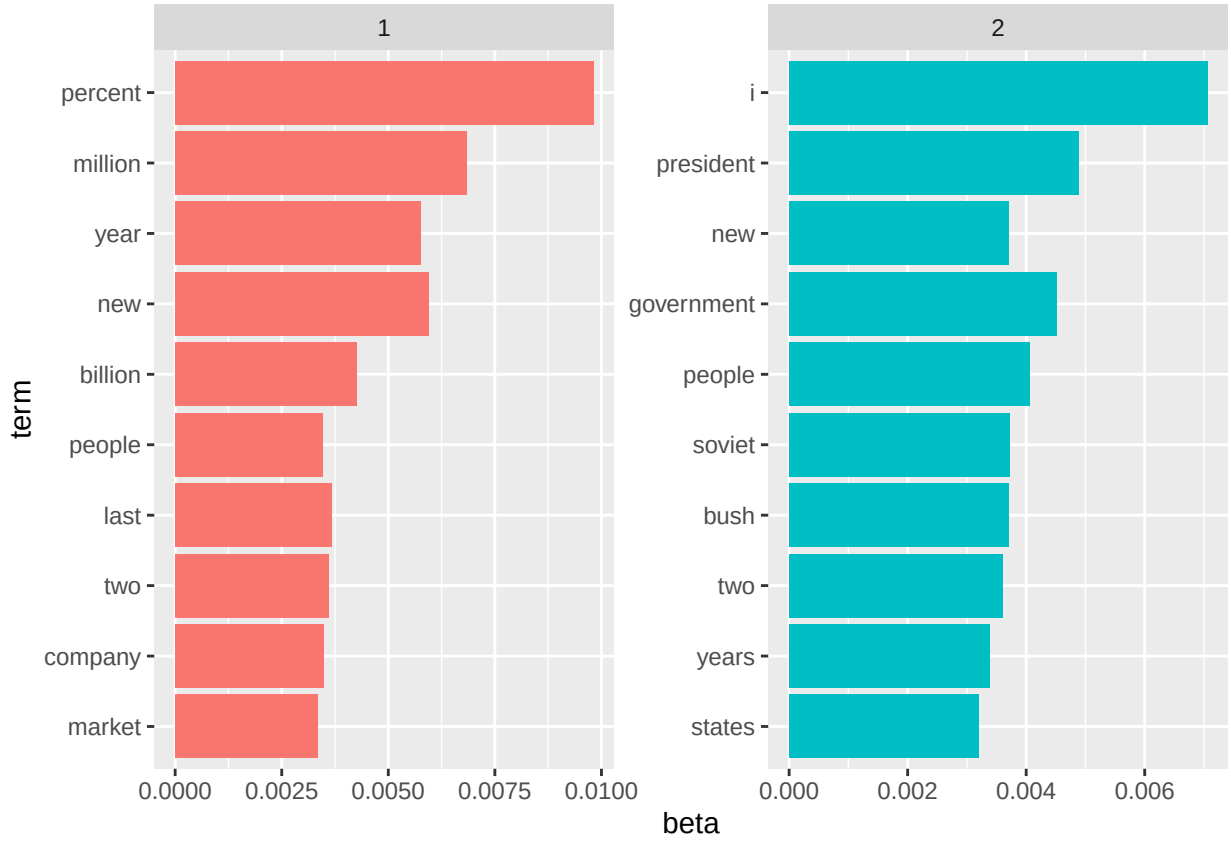
```
## # A tibble: 20,946 x 3
##    topic term           beta
##    <int> <chr>         <dbl>
##  1     1 aaron      1.69e-12
##  2     2 aaron      3.90e- 5
##  3     1 abandon    2.65e- 5
##  4     2 abandon    3.99e- 5
##  5     1 abandoned  1.39e- 4
##  6     2 abandoned  5.88e- 5
##  7     1 abandoning 2.45e-33
##  8     2 abandoning 2.34e- 5
##  9     1 abbott     2.13e- 6
## 10     2 abbott     2.97e- 5
## # ... with 20,936 more rows
```

```
library(ggplot2)
library(dplyr)

ap_top_terms <- ap_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

ap_top_terms %>%
  mutate(term = reorder(term, beta)) %>%
```

```
  ggplot(aes(term, beta, fill = factor(topic))) +
    geom_col(show.legend = FALSE) +
    facet_wrap(~ topic, scales = "free") +
    coord_flip()
```




```
library(tidyr)

beta_spread <- ap_topics %>%
  mutate(topic = paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  dplyr::filter(topic1 > 0.001 | topic2 > 0.001) %>%
  mutate(log_ratio = log2(topic2 / topic1))

beta_spread
```

```
## # A tibble: 198 x 4
##    term              topic1      topic2 log_ratio
##    <chr>              <dbl>       <dbl>     <dbl>
##  1 administration 0.000431  0.00138         1.68
##  2 ago            0.00107   0.000842       -0.339
##  3 agreement      0.000671  0.00104         0.630
##  4 aid            0.0000476 0.00105         4.46
##  5 air            0.00214   0.000297       -2.85
##  6 american       0.00203   0.00168        -0.270
##  7 analysts       0.00109   0.000000578   -10.9
##  8 area           0.00137   0.000231       -2.57
##  9 army           0.000262  0.00105         2.00
```

```
## 10 asked          0.000189  0.00156         3.05
## # ... with 188 more rows
```

```
ap_documents <- tidy(ap_lda, matrix = "gamma")
ap_documents
```

```
## # A tibble: 4,492 x 3
##    document topic    gamma
##       <int> <int>    <dbl>
##  1        1     1 0.248
##  2        2     1 0.362
##  3        3     1 0.527
##  4        4     1 0.357
##  5        5     1 0.181
##  6        6     1 0.000588
##  7        7     1 0.773
##  8        8     1 0.00445
##  9        9     1 0.967
## 10       10     1 0.147
## # ... with 4,482 more rows
```

```
tidy(AssociatedPress) %>%
  dplyr::filter(document == 6) %>%
  arrange(desc(count))
```

```
## # A tibble: 287 x 3
##    document term           count
##       <int> <chr>          <dbl>
##  1        6 noriega           16
##  2        6 panama            12
##  3        6 jackson            6
##  4        6 powell             6
##  5        6 administration     5
##  6        6 economic           5
##  7        6 general            5
##  8        6 i                  5
##  9        6 panamanian         5
## 10        6 american           4
## # ... with 277 more rows
```

```
titles <- c("Twenty Thousand Leagues under the Sea",
            "The War of the Worlds",
            "Pride and Prejudice",
            "Great Expectations")
library(gutenbergr)
books <- gutenberg_works(title %in% titles) %>%
  gutenberg_download(meta_fields = "title")
```

```
## Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest
```

```
## Using mirror http://aleph.gutenberg.org
```

```
library(stringr)
reg <- regex("^chapter", ignore_case = TRUE)
by_chapter <- books %>%
  group_by(title) %>%
  mutate(chapter = cumsum(str_detect(text, reg))) %>%
```

```
  ungroup() %>%
  dplyr::filter(chapter > 0) %>%
  unite(document, title, chapter)

by_chapter_word <- by_chapter %>%
  unnest_tokens(word, text)

word_counts <- by_chapter_word %>%
  anti_join(stop_words) %>%
  count(document, word, sort = TRUE) %>%
  ungroup()
```

```
## Joining, by = "word"
```

```
word_counts
```

```
## # A tibble: 79,267 x 3
##    document                 word        n
##    <chr>                    <chr>   <int>
##  1 Great Expectations_57    joe        88
##  2 Great Expectations_7     joe        70
##  3 Great Expectations_17    biddy      63
##  4 Great Expectations_27    joe        58
##  5 Great Expectations_38    estella    58
##  6 Great Expectations_2     joe        56
##  7 Great Expectations_23    pocket     53
##  8 Great Expectations_15    joe        50
##  9 Great Expectations_18    joe        50
## 10 The War of the Worlds_16 brother    50
## # ... with 79,257 more rows
```

```
chapters_dtm <- word_counts %>%
  cast_dtm(document, word, n)

chapters_dtm
```

```
## <<DocumentTermMatrix (documents: 133, terms: 16685)>>
## Non-/sparse entries: 79267/2139838
## Sparsity           : 96%
## Maximal term length: 19
## Weighting          : term frequency (tf)
```

```
chapters_lda <- LDA(chapters_dtm, k = 4, control = list(seed = 1234))
chapters_lda
```

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

```
chapter_topics <- tidy(chapters_lda, matrix = "beta")
chapter_topics
```

```
## # A tibble: 66,740 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 joe     1.28e- 2
##  2     2 joe     1.96e-23
##  3     3 joe     8.27e-51
##  4     4 joe     2.93e-17
```

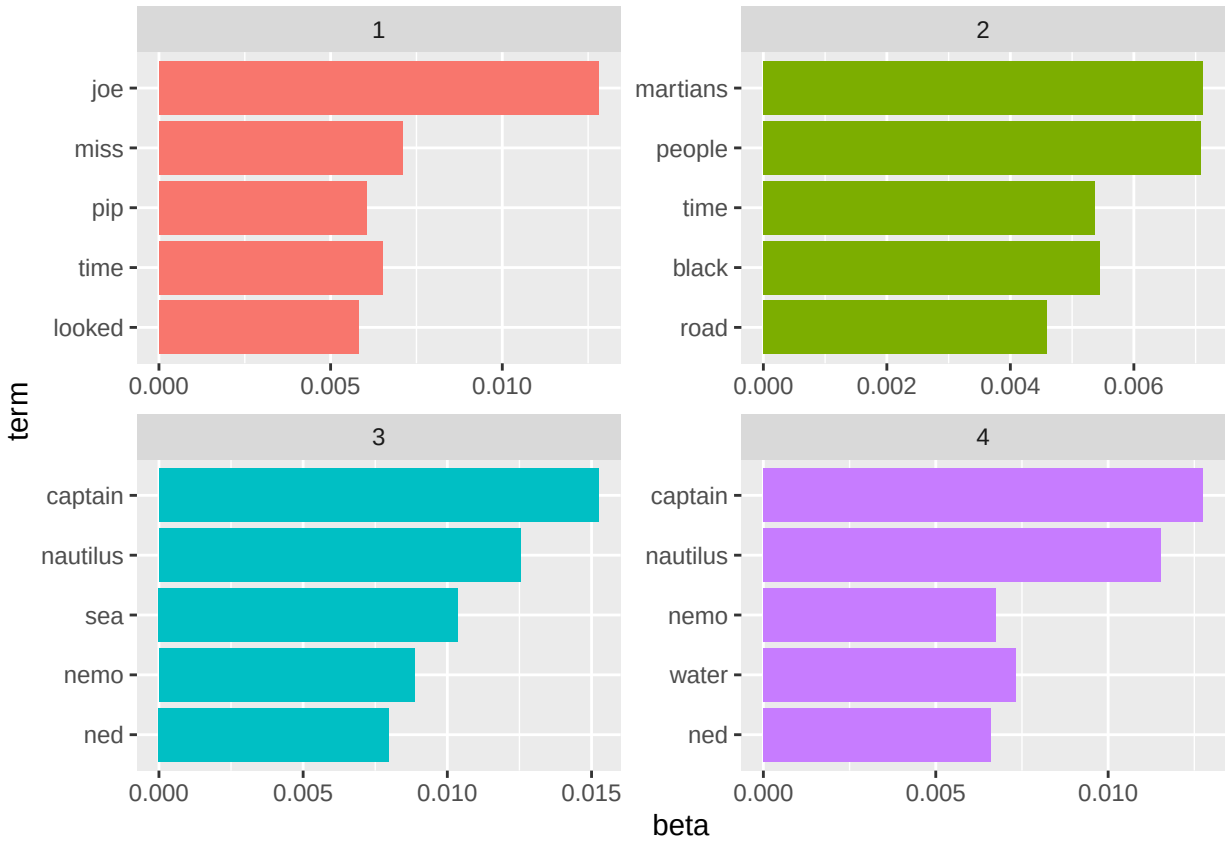
```
##  5     1 biddy   4.22e- 3
##  6     2 biddy   5.38e-21
##  7     3 biddy   6.75e-60
##  8     4 biddy   1.07e-20
##  9     1 estella 4.39e- 3
## 10     2 estella 6.43e-25
## # ... with 66,730 more rows
```

```
top_terms <- chapter_topics %>%
  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 joe      0.0128
##  2     1 miss     0.00709
##  3     1 time     0.00651
##  4     1 pip      0.00603
##  5     1 looked   0.00580
##  6     2 martians 0.00712
##  7     2 people   0.00708
##  8     2 black    0.00545
##  9     2 time     0.00537
## 10     2 road     0.00459
## 11     3 captain  0.0152
## 12     3 nautilus 0.0125
## 13     3 sea      0.0104
## 14     3 nemo     0.00885
## 15     3 ned      0.00798
## 16     4 captain  0.0128
## 17     4 nautilus 0.0115
## 18     4 water    0.00732
## 19     4 nemo     0.00674
## 20     4 ned      0.00659
```

```
library(ggplot2)

top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
    geom_col(show.legend = FALSE) +
    facet_wrap(~ topic, scales = "free") +
    coord_flip()
```




```
chapters_gamma <- tidy(chapters_lda, matrix = "gamma")
chapters_gamma
```

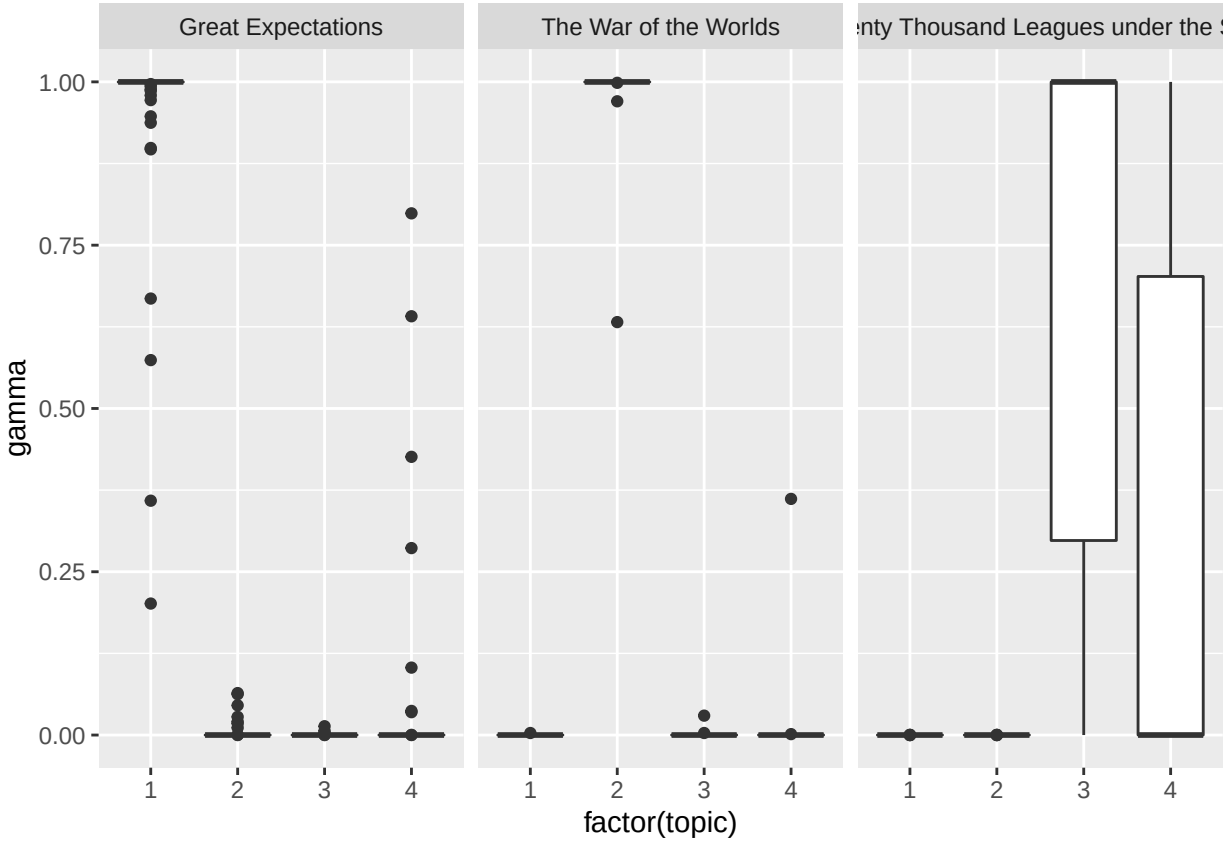
```
## # A tibble: 532 x 3
##    document                 topic     gamma
##    <chr>                    <int>     <dbl>
##  1 Great Expectations_57        1 1.00
##  2 Great Expectations_7         1 1.00
##  3 Great Expectations_17        1 1.00
##  4 Great Expectations_27        1 1.00
##  5 Great Expectations_38        1 1.00
##  6 Great Expectations_2         1 1.00
##  7 Great Expectations_23        1 0.574
##  8 Great Expectations_15        1 1.00
##  9 Great Expectations_18        1 1.00
## 10 The War of the Worlds_16     1 0.0000128
## # ... with 522 more rows
```

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

```
## # A tibble: 532 x 4
##    title                chapter topic     gamma
##    <chr>                  <int> <int>     <dbl>
##  1 Great Expectations        57     1 1.00
##  2 Great Expectations         7     1 1.00
##  3 Great Expectations        17     1 1.00
```

```
##  4 Great Expectations         27     1 1.00
##  5 Great Expectations         38     1 1.00
##  6 Great Expectations          2     1 1.00
##  7 Great Expectations         23     1 0.574
##  8 Great Expectations         15     1 1.00
##  9 Great Expectations         18     1 1.00
## 10 The War of the Worlds      16     1 0.0000128
## # ... with 522 more rows
```

```
chapters_gamma %>%
  mutate(title = reorder(title, gamma * topic)) %>%
  ggplot(aes(factor(topic), gamma)) +
    geom_boxplot() +
    facet_wrap(~ title)
```




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

chapter_classifications
```

```
## # A tibble: 133 x 4
##    title              chapter topic gamma
##    <chr>                <int> <int> <dbl>
##  1 Great Expectations      57     1 1.00
##  2 Great Expectations       7     1 1.00
##  3 Great Expectations      17     1 1.00
```

```
##  4 Great Expectations      27     1 1.00
##  5 Great Expectations      38     1 1.00
##  6 Great Expectations       2     1 1.00
##  7 Great Expectations      23     1 0.574
##  8 Great Expectations      15     1 1.00
##  9 Great Expectations      18     1 1.00
## 10 Great Expectations       9     1 1.00
## # ... with 123 more rows
```

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

chapter_classifications %>%
  inner_join(book_topics, by = "topic") %>%
  dplyr::filter(title != consensus)
```

```
## # A tibble: 0 x 5
## # ... with 5 variables: title <chr>, chapter <int>, topic <int>, gamma <dbl>,
## #   consensus <chr>
```

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

```
## # A tibble: 79,267 x 4
##    document              term  count .topic
##    <chr>                 <chr> <dbl>  <dbl>
##  1 Great Expectations_57 joe      88      1
##  2 Great Expectations_7  joe      70      1
##  3 Great Expectations_17 joe       5      1
##  4 Great Expectations_27 joe      58      1
##  5 Great Expectations_2  joe      56      1
##  6 Great Expectations_23 joe       1      1
##  7 Great Expectations_15 joe      50      1
##  8 Great Expectations_18 joe      50      1
##  9 Great Expectations_9  joe      44      1
## 10 Great Expectations_13 joe      40      1
## # ... with 79,257 more rows
```

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

```
## # A tibble: 69,494 x 6
##    title              chapter term  count .topic consensus
##    <chr>                <int> <chr> <dbl>  <dbl> <chr>
##  1 Great Expectations      57 joe      88      1 Great Expectations
##  2 Great Expectations       7 joe      70      1 Great Expectations
##  3 Great Expectations      17 joe       5      1 Great Expectations
##  4 Great Expectations      27 joe      58      1 Great Expectations
##  5 Great Expectations       2 joe      56      1 Great Expectations
##  6 Great Expectations      23 joe       1      1 Great Expectations
```

```
##  7 Great Expectations      15 joe      50      1 Great Expectations
##  8 Great Expectations      18 joe      50      1 Great Expectations
##  9 Great Expectations       9 joe      44      1 Great Expectations
## 10 Great Expectations      13 joe      40      1 Great Expectations
## # ... with 69,484 more rows
```

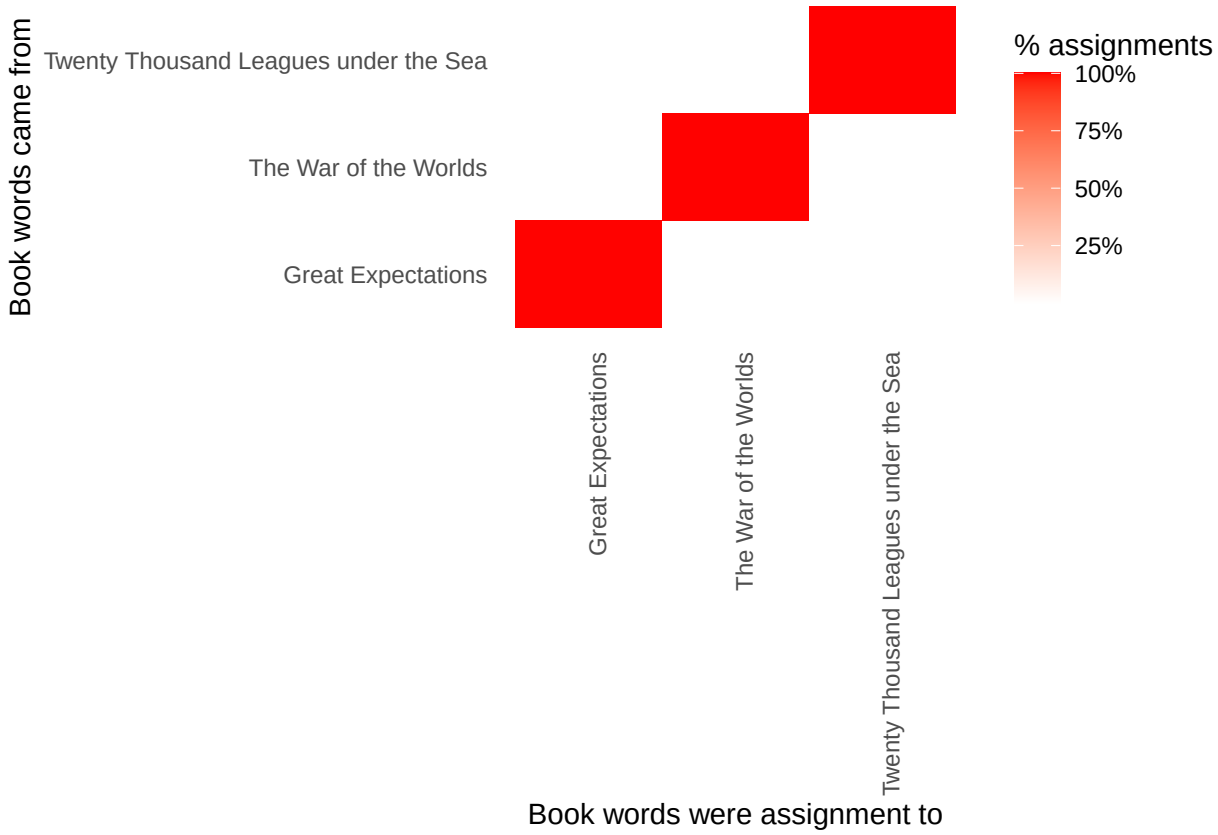
```
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
assignments %>%
  count(title, consensus, wt = count) %>%
  group_by(title) %>%
  mutate(percentage = n / sum(n)) %>%
  ggplot(aes(consensus, title, fill = percentage)) +
    geom_tile() +
    scale_fill_gradient2(high = 'red', label = percent_format()) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1),
          panel.grid = element_blank()) +
    labs(x = "Book words were assignment to",
         y = "Book words came from",
         fill = "% assignments")
```




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

```
## # A tibble: 67 x 6
##    title              chapter term     count .topic consensus
##    <chr>                <int> <chr>    <dbl>  <dbl> <chr>
##  1 Great Expectatio~       32 captain      1      3 Twenty Thousand Leagues unde~
##  2 The War of the W~       17 captain      5      3 Twenty Thousand Leagues unde~
##  3 Great Expectatio~       39 land         1      3 Twenty Thousand Leagues unde~
##  4 The War of the W~       17 land         1      3 Twenty Thousand Leagues unde~
##  5 The War of the W~       17 vessel       1      3 Twenty Thousand Leagues unde~
##  6 Great Expectatio~       16 houses       1      2 The War of the Worlds
##  7 The War of the W~       17 ocean        1      3 Twenty Thousand Leagues unde~
##  8 Great Expectatio~       14 saloon       1      3 Twenty Thousand Leagues unde~
##  9 The War of the W~       17 march        1      3 Twenty Thousand Leagues unde~
## 10 Great Expectatio~       59 hurrying     1      2 The War of the Worlds
## # ... with 57 more rows
```

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

```
## # A tibble: 65 x 4
##    title                consensus                              term            n
##    <chr>                <chr>                                  <chr>       <dbl>
##  1 The War of the Worlds Twenty Thousand Leagues under the Sea captain         5
```

```
##  2 Great Expectations    The War of the Worlds                crept           3
##  3 Great Expectations    The War of the Worlds                dense           2
##  4 Great Expectations    The War of the Worlds                active          1
##  5 Great Expectations    The War of the Worlds                authorities     1
##  6 Great Expectations    The War of the Worlds                avenue          1
##  7 Great Expectations    The War of the Worlds                beach           1
##  8 Great Expectations    The War of the Worlds                bend            1
##  9 Great Expectations    The War of the Worlds                blundered       1
## 10 Great Expectations    The War of the Worlds                cheering        1
## # ... with 55 more rows
```

```
word_counts %>%
  dplyr::filter(word == "flopson")
```

```
## # A tibble: 3 x 3
##   document              word        n
##   <chr>                 <chr>   <int>
## 1 Great Expectations_22 flopson    10
## 2 Great Expectations_23 flopson     7
## 3 Great Expectations_33 flopson     1
```

```
library(mallet)
```

```
## Loading required package: rJava
```

```
collapsed <- by_chapter_word %>%
  anti_join(stop_words, by = "word") %>%
  mutate(word = str_replace(word, "'", "")) %>%
  group_by(document) %>%
  summarise(text = paste(word, collapse = " "))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
file.create(empty_file <- tempfile())
```

```
## [1] TRUE
```

```
docs <- mallet.import(collapsed$document, collapsed$text, empty_file)

mallet_model <- MalletLDA(num.topics = 4)
mallet_model$loadDocuments(docs)
mallet_model$train(100)
```

```
tidy(mallet_model)
```

```
## Warning: `tbl_df()` is deprecated as of dplyr 1.0.0.
## Please use `tibble::as_tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## # A tibble: 64,760 x 3
##    topic term           beta
##    <int> <chr>         <dbl>
##  1     1 chapter 0.000827
##  2     2 chapter 0.00279
##  3     3 chapter 0.00101
##  4     4 chapter 0.000000299
##  5     1 father  0.000000359
```

```
##  6     2 father  0.0000336
##  7     3 father  0.000000325
##  8     4 father  0.00210
##  9     1 s       0.000000359
## 10     2 s       0.000000332
## # ... with 64,750 more rows
```

```
tidy(mallet_model, matrix = "gamma")
```

```
## # A tibble: 532 x 3
##    document              topic  gamma
##    <chr>                 <int>  <dbl>
##  1 Great Expectations_1      1 0.162
##  2 Great Expectations_10     1 0.135
##  3 Great Expectations_11     1 0.142
##  4 Great Expectations_12     1 0.135
##  5 Great Expectations_13     1 0.0736
##  6 Great Expectations_14     1 0.135
##  7 Great Expectations_15     1 0.150
##  8 Great Expectations_16     1 0.182
##  9 Great Expectations_17     1 0.115
## 10 Great Expectations_18     1 0.111
## # ... with 522 more rows
```

```
term_counts <- rename(word_counts, term = word)
augment(mallet_model, term_counts)
```

```
## # A tibble: 79,267 x 4
##    document                term        n .topic
##    <chr>                   <chr>   <int>  <int>
##  1 Great Expectations_57   joe        88      4
##  2 Great Expectations_7    joe        70      4
##  3 Great Expectations_17   biddy      63      4
##  4 Great Expectations_27   joe        58      4
##  5 Great Expectations_38   estella    58      4
##  6 Great Expectations_2    joe        56      4
##  7 Great Expectations_23   pocket     53      4
##  8 Great Expectations_15   joe        50      4
##  9 Great Expectations_18   joe        50      4
## 10 The War of the Worlds_16 brother    50      1
## # ... with 79,257 more rows
```