Chapter 1 - notes

John Peach

3/14/2021

```
text <- c("Because I could not stop for Death .",
          "He kindly stopped for me .",
          "The Carriage held but just Ourselves -",
          "and Immortality")

library(dplyr)

text_df <- data.frame(line = 1:4, text = text)

library(tidytext)

text_df %>%
  unnest_tokens(word, text)
```

```
##    line        word
## 1     1     because
## 2     1           i
## 3     1       could
## 4     1         not
## 5     1        stop
## 6     1         for
## 7     1       death
## 8     2          he
## 9     2      kindly
## 10    2     stopped
## 11    2         for
## 12    2          me
## 13    3         the
## 14    3    carriage
## 15    3        held
## 16    3         but
## 17    3        just
## 18    3   ourselves
## 19    4         and
## 20    4 immortality
```

```
library(janeaustenr)
library(dplyr)
library(stringr)

original_books <- austen_books() %>%
  group_by(book) %>%
  mutate(linenumber = row_number(),
```

```
        chapter = cumsum(str_detect(text, regex("^chapter [\\divxlc]",
                                                ignore_case = TRUE)))) %>%
  ungroup()

original_books
```

```
## # A tibble: 73,422 x 4
##    text                    book                linenumber chapter
##    <chr>                   <fct>                    <int>   <int>
##  1 "SENSE AND SENSIBILITY" Sense & Sensibility          1       0
##  2 ""                      Sense & Sensibility          2       0
##  3 "by Jane Austen"        Sense & Sensibility          3       0
##  4 ""                      Sense & Sensibility          4       0
##  5 "(1811)"                Sense & Sensibility          5       0
##  6 ""                      Sense & Sensibility          6       0
##  7 ""                      Sense & Sensibility          7       0
##  8 ""                      Sense & Sensibility          8       0
##  9 ""                      Sense & Sensibility          9       0
## 10 "CHAPTER 1"             Sense & Sensibility         10       1
## # ... with 73,412 more rows
```

```
library(tidytext)
tidy_books <- original_books %>%
  unnest_tokens(word, text)

tidy_books
```

```
## # A tibble: 725,055 x 4
##    book                linenumber chapter word
##    <fct>                    <int>   <int> <chr>
##  1 Sense & Sensibility          1       0 sense
##  2 Sense & Sensibility          1       0 and
##  3 Sense & Sensibility          1       0 sensibility
##  4 Sense & Sensibility          3       0 by
##  5 Sense & Sensibility          3       0 jane
##  6 Sense & Sensibility          3       0 austen
##  7 Sense & Sensibility          5       0 1811
##  8 Sense & Sensibility         10       1 chapter
##  9 Sense & Sensibility         10       1 1
## 10 Sense & Sensibility         13       1 the
## # ... with 725,045 more rows
```

```
data("stop_words")

tidy_books <- tidy_books %>%
  anti_join(stop_words)
```

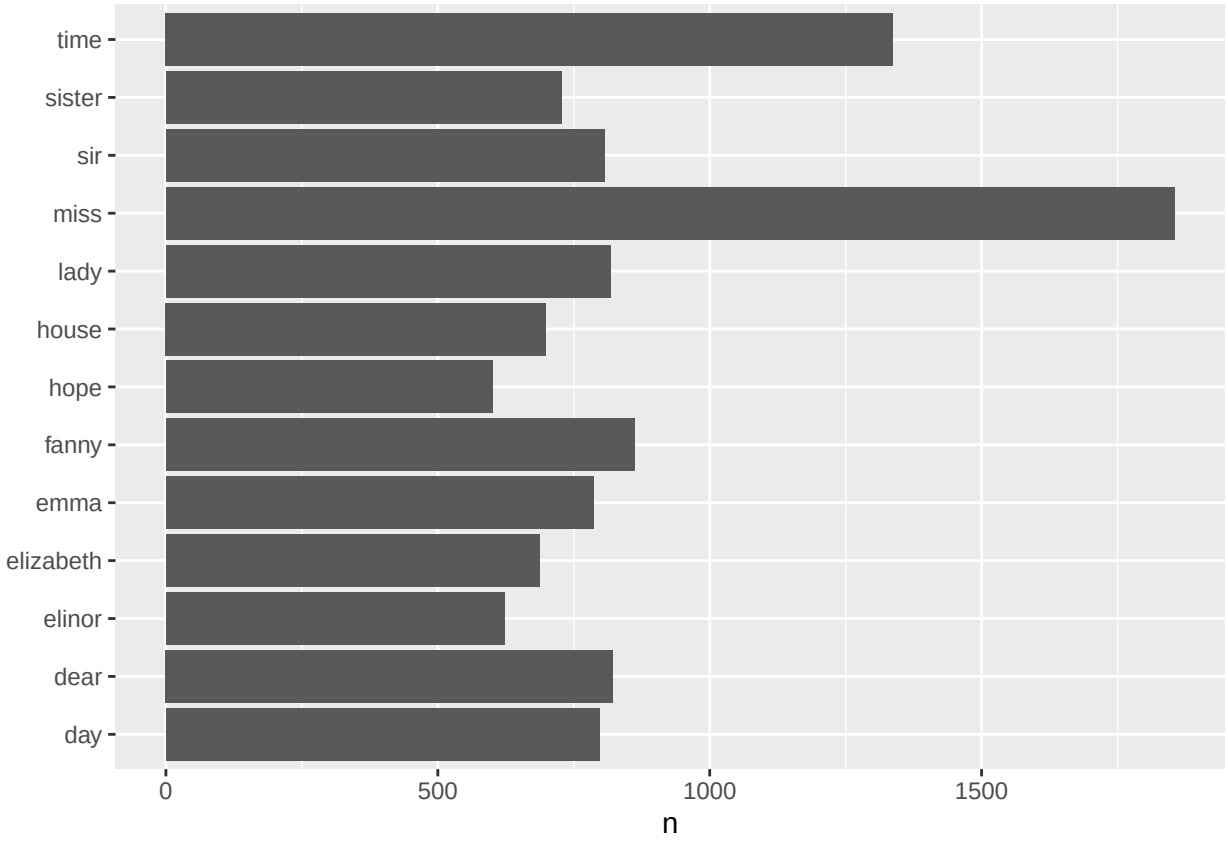
```
## Joining, by = "word"
```

```
tidy_books %>%
  count(word, sort = TRUE)
```

```
## # A tibble: 13,914 x 2
##    word       n
##    <chr>  <int>
##  1 miss    1855
```

```
##  2 time    1337
##  3 fanny    862
##  4 dear     822
##  5 lady     817
##  6 sir      806
##  7 day      797
##  8 emma     787
##  9 sister   727
## 10 house    699
## # ... with 13,904 more rows
```

```
library(ggplot2)
tidy_books %>%
  count(word, sort = TRUE) %>%
  dplyr::filter(n > 600) %>%
  ggplot(aes(word, n)) +
    geom_col() +
    xlab(NULL) +
    coord_flip()
```




```
library(gutenbergr)
hgwells <- gutenberg_download(c(35, 36, 5230, 159))
```

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

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

```
tidy_hgwells <- hgwells %>%
  unnest_tokens(word, text) %>%
```

```
  anti_join(stop_words)
```

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

```
tidy_hgwells %>%
  count(word, sort = TRUE)
```

```
## # A tibble: 11,830 x 2
##    word       n
##    <chr>  <int>
##  1 time     461
##  2 people   302
##  3 door     260
##  4 heard    249
##  5 black    232
##  6 stood    229
##  7 white    224
##  8 hand     218
##  9 kemp     213
## 10 eyes     210
## # ... with 11,820 more rows
```

```
bronte <- gutenberg_download(c(1260, 768, 969, 9182, 767))

tidy_bronte <- bronte %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

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

```
tidy_bronte %>%
  count(word, sort = TRUE)
```

```
## # A tibble: 23,303 x 2
##    word       n
##    <chr>  <int>
##  1 time    1064
##  2 miss     854
##  3 day      826
##  4 hand     767
##  5 eyes     713
##  6 don't    666
##  7 night    648
##  8 heart    638
##  9 looked   601
## 10 door     591
## # ... with 23,293 more rows
```

```
library(tidyr)
frequency <- bind_rows(mutate(tidy_bronte, author = "Bronte Sisters"),
                       mutate(tidy_hgwells, author = "H. G. Wells"),
                       mutate(tidy_books, author = "Jane Austen")) %>%
  mutate(word = str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author) %>%
  mutate(proprotion = n / sum(n)) %>%
```

```
  select(-n) %>%
  spread(author, proprotion) %>%
  gather(author, proportion, `Bronte Sisters`:`H. G. Wells`)
```

```
library(ggplot2)
library(scales)
```

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

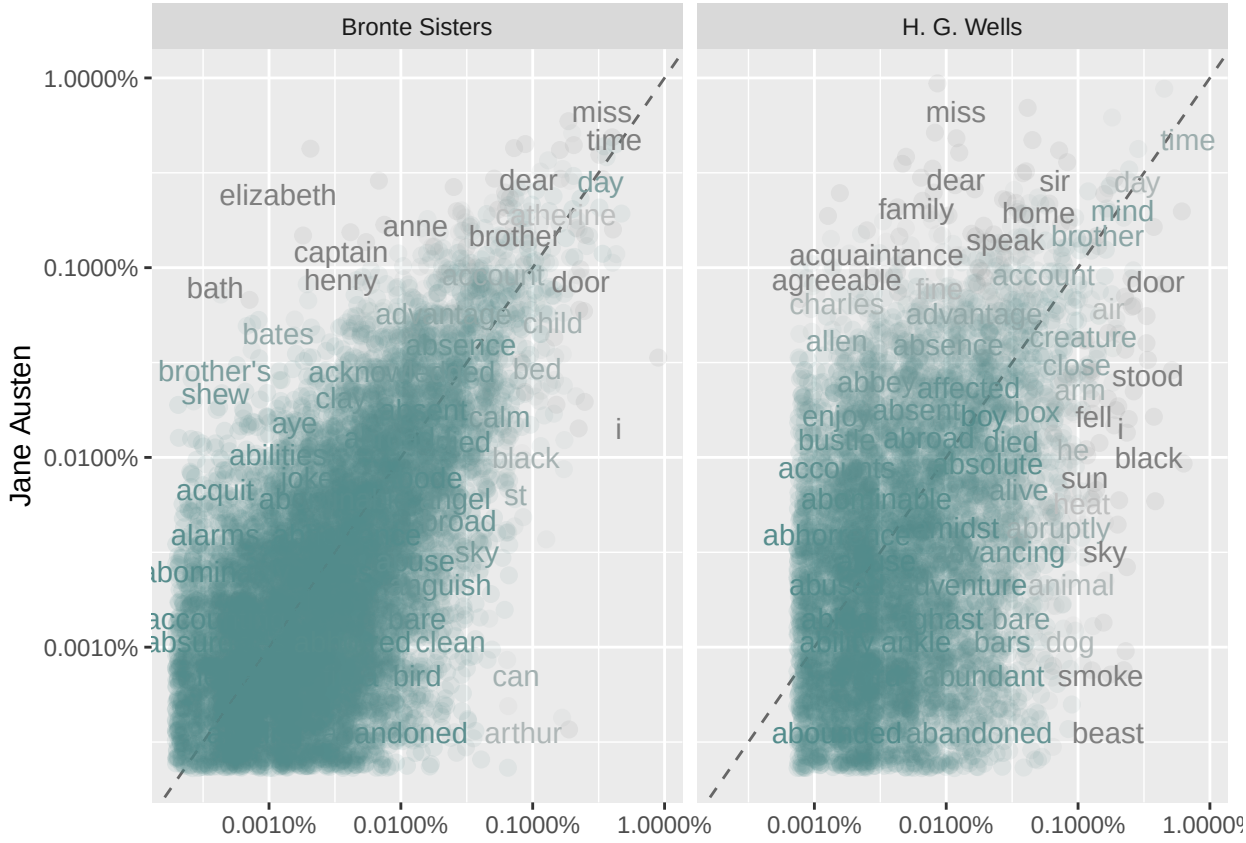
```
## The following object is masked from 'package:purrr':
##
##     discard
```

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

```
ggplot(frequency, aes(x = proportion, y = `Jane Austen`,
  colour = abs(`Jane Austen` - proportion))) +
  geom_abline(colour = 'gray40', lty = 2) +
  geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  scale_color_gradient(limits = c(0, 0.001),
                       low = 'darkslategray4', high = 'gray75') +
  facet_wrap(~author, ncol = 2) +
  theme(legend.position = 'none') +
  labs(y = "Jane Austen", x = NULL)
```

```
## Warning: Removed 40857 rows containing missing values (geom_point).
```

```
## Warning: Removed 40859 rows containing missing values (geom_text).
```




```
cor.test(data = frequency[frequency$author == "Bronte Sisters", ],
         ~ proportion + `Jane Austen`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and Jane Austen
## t = 111.09, df = 10345, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7286567 0.7462329
## sample estimates:
##       cor
## 0.7375697
```

```
cor.test(data = frequency[frequency$author == "H. G. Wells", ],
         ~ proportion + `Jane Austen`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and Jane Austen
## t = 36.083, df = 6046, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3999815 0.4414612
## sample estimates:
##       cor
```

```
## 0.4209414
```