

homework_1.R

mojmir

2020-12-10

```
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)
```

```
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) %>%
  filter(n > 600) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip()
```

```
# the same but with gutenbergr download
```

```
## To learn more about gutenbergr, check out the [package's tutorial at rOpenSci](https://ropensci.org/)
```

```
library(gutenbergr)
```

```
## Own example: Alice
```

```
alice <- gutenbergr_download(11)
```

```
tidy_alice <- alice %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

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

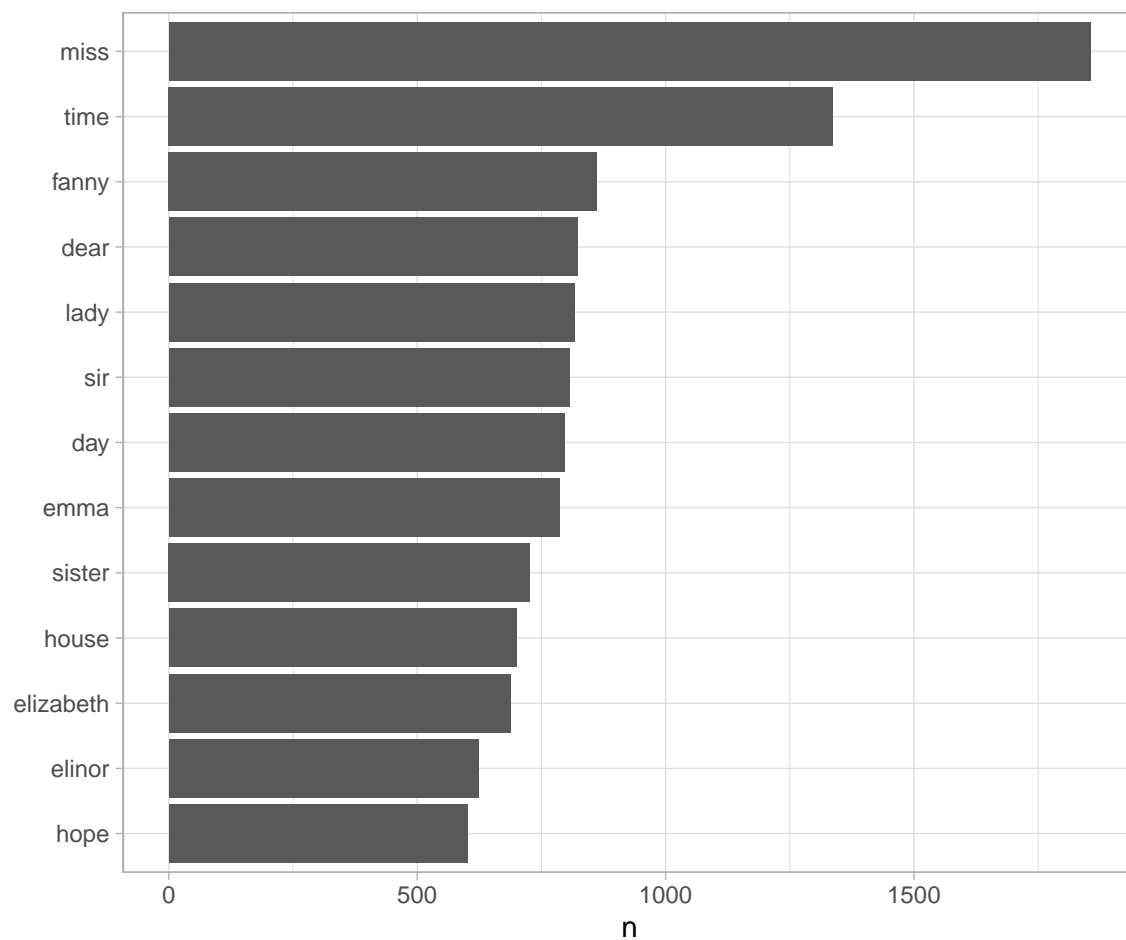


Figure 1: The most common words in Jane Austen's novels

```
## # A tibble: 2,322 x 2
##   word      n
##   <chr>   <int>
## 1 alice   386
## 2 time    71
## 3 queen   68
## 4 king    61
## 5 don't   60
## 6 it's    57
## 7 mock    57
## 8 i'm     56
## 9 turtle  56
## 10 gryphon 55
## # ... with 2,312 more rows
```

```
library(ggplot2)
```

```
tidy_alice %>%
  count(word, sort = TRUE) %>%
  filter(n > 50) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip() +
  ggtitle("Alice: most common words")
```

