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

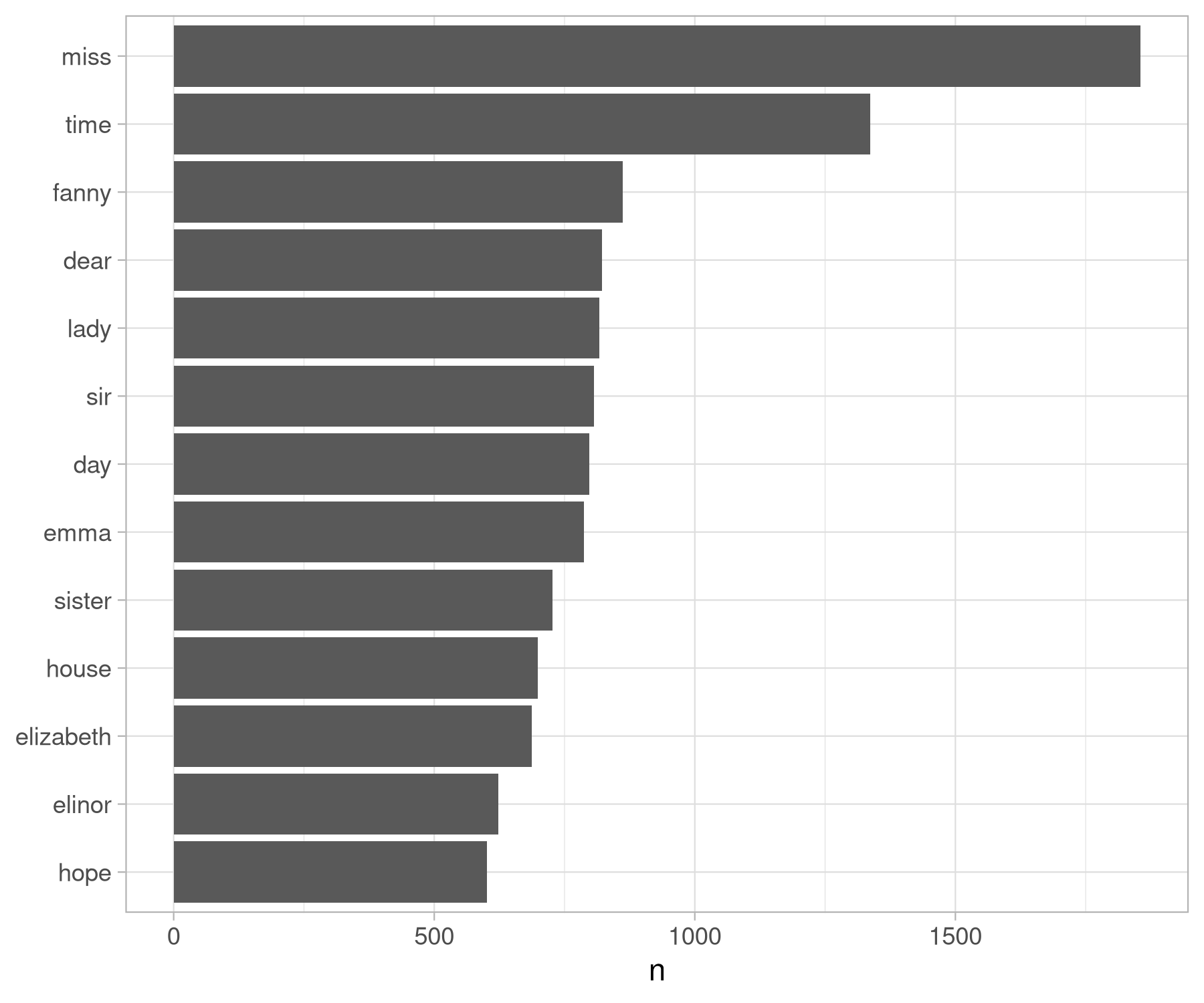
## # A tibble: 725,055 x 4  
## book linenumber chapter word   
## <fct> <int> <int> <chr>   
## 1 Sense & Sensibility 1 0 sense   
## 2 Sense & Sensibility 1 0 and   
## 3 Sense & Sensibility 1 0 sensibility  
## 4 Sense & Sensibility 3 0 by   
## 5 Sense & Sensibility 3 0 jane   
## 6 Sense & Sensibility 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 most common words in Jane Austen’s novels

# the same but with gutenberg download  
## To learn more about gutenbergr, check out the [package's tutorial at rOpenSci](https://ropensci.org/tutorials/gutenbergr\_tutorial.html), where it is one of rOpenSci's packages for data access.

library(gutenbergr)  
  
## Own example: Alice  
  
alice <- gutenberg\_download(11)  
  
tidy\_alice <- alice %>%  
 unnest\_tokens(word, text) %>%  
 anti\_join(stop\_words)  
  
tidy\_alice %>%  
 count(word, sort = TRUE)  
  
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()