Text analizes classs

Kristof Menyhert

2018-02-22

Coding example:

In this exercise I am presenting an example what can you do with the gutenbergR package combining with the tidytext and the tidyverse package. Then do some basic piloting with ggplot.

With the help of the tools mentioned above we can see how words in two books are related to each other. In this way we can have some insights how the two books are relate to each other, notice some important words for each book.

In the following lines you can find an example how to deal do this:

Load some packages:

```
library(gutenbergr)
library(tidytext)
library(tidyverse)
library(ggplot2)
library(scales)
```

Presenting how to get the text of a given book:

Get Mark Twain's Tom Sawyer:

```
mark_twain <- gutenberg_works(author == "Twain, Mark", str_detect(title, "The Adventures of Tom Sawyer"
```

We should look for the id:

```
head(mark_twain[, c("title", "gutenberg_id")])
```

```
## # A tibble: 6 x 2
##   title                                   gutenberg_id
##   <chr>                                          <int>
## 1 The Adventures of Tom Sawyer                      74
## 2 The Adventures of Tom Sawyer, Part 1.           7193
## 3 The Adventures of Tom Sawyer, Part 2.           7194
## 4 The Adventures of Tom Sawyer, Part 3.           7195
## 5 The Adventures of Tom Sawyer, Part 4.           7196
## 6 The Adventures of Tom Sawyer, Part 5.           7197
```

In this case the id for this book is 74.

Find another book:

In this case I choose Agatha Christi's The Secret Adversary:

```
agatha_christie <- gutenberg_works(str_detect(author, "Agatha"))

head(agatha_christie)
```

```
## # A tibble: 4 x 8
##   gutenberg_id title    author  gutenberg_autho~ language gutenberg_books~
##          <int> <chr>    <chr>              <int> <chr>    <chr>
## 1          863 The Mys~ Christ~               451 en       Detective Ficti~
## 2         1155 The Sec~ Christ~               451 en       Detective Ficti~
## 3         6945 Marguer~ Armour~              2269 en       <NA>
## 4        18145 Lady Ro~ Armour~              2269 en       <NA>
## # ... with 2 more variables: rights <chr>, has_text <lgl>
```

Download both of the book:

```
books <- gutenberg_download(c(74, 1155), meta_fields = "title")
```

Use the unnest tokens function to split the data to words:

```
words <- books %>% unnest_tokens(word, text) %>% anti_join(stop_words)

head(words)
```

```
## # A tibble: 6 x 3
##   gutenberg_id title                          word
##          <int> <chr>                          <chr>
## 1           74 The Adventures of Tom Sawyer adventures
## 2           74 The Adventures of Tom Sawyer tom
## 3           74 The Adventures of Tom Sawyer sawyer
## 4           74 The Adventures of Tom Sawyer mark
## 5           74 The Adventures of Tom Sawyer twain
## 6           74 The Adventures of Tom Sawyer samuel
```

Count the words how many times they appear in a book by title and sort them:

```
word_counts <- words %>%
  count(title, word, sort = TRUE)

head(word_counts, 10)
```

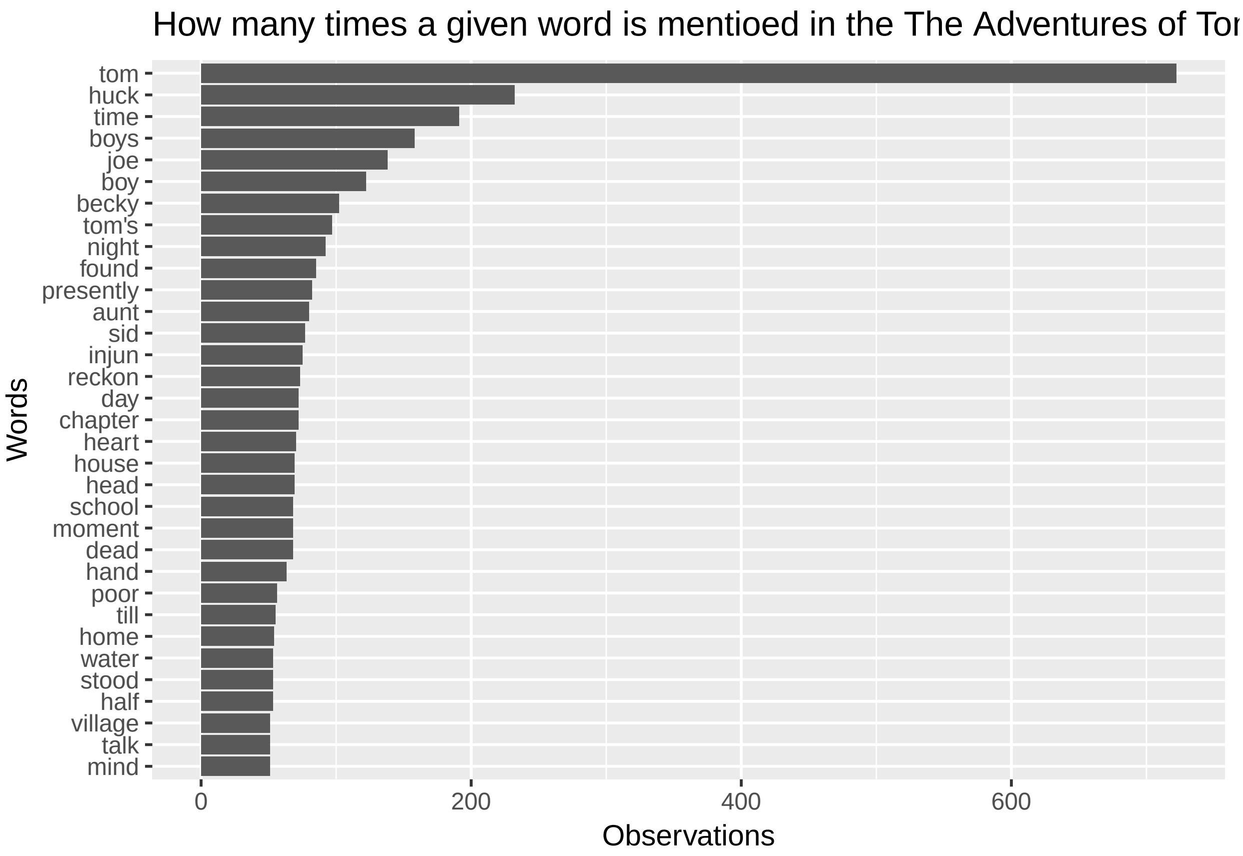
```
## # A tibble: 10 x 3
##    title                        word         n
##    <chr>                        <chr>    <int>
##  1 The Adventures of Tom Sawyer tom        722
##  2 The Secret Adversary         tuppence   585
##  3 The Secret Adversary         tommy      546
##  4 The Secret Adversary         julius     303
##  5 The Secret Adversary         sir        243
##  6 The Adventures of Tom Sawyer huck       232
##  7 The Adventures of Tom Sawyer time       191
##  8 The Secret Adversary         don't      190
##  9 The Adventures of Tom Sawyer boys       158
## 10 The Secret Adversary         james      156
```

Plotting:

Which are the most frequent word in The Adventures of Tom Sawyer by Mark Twain?

See the plot below:

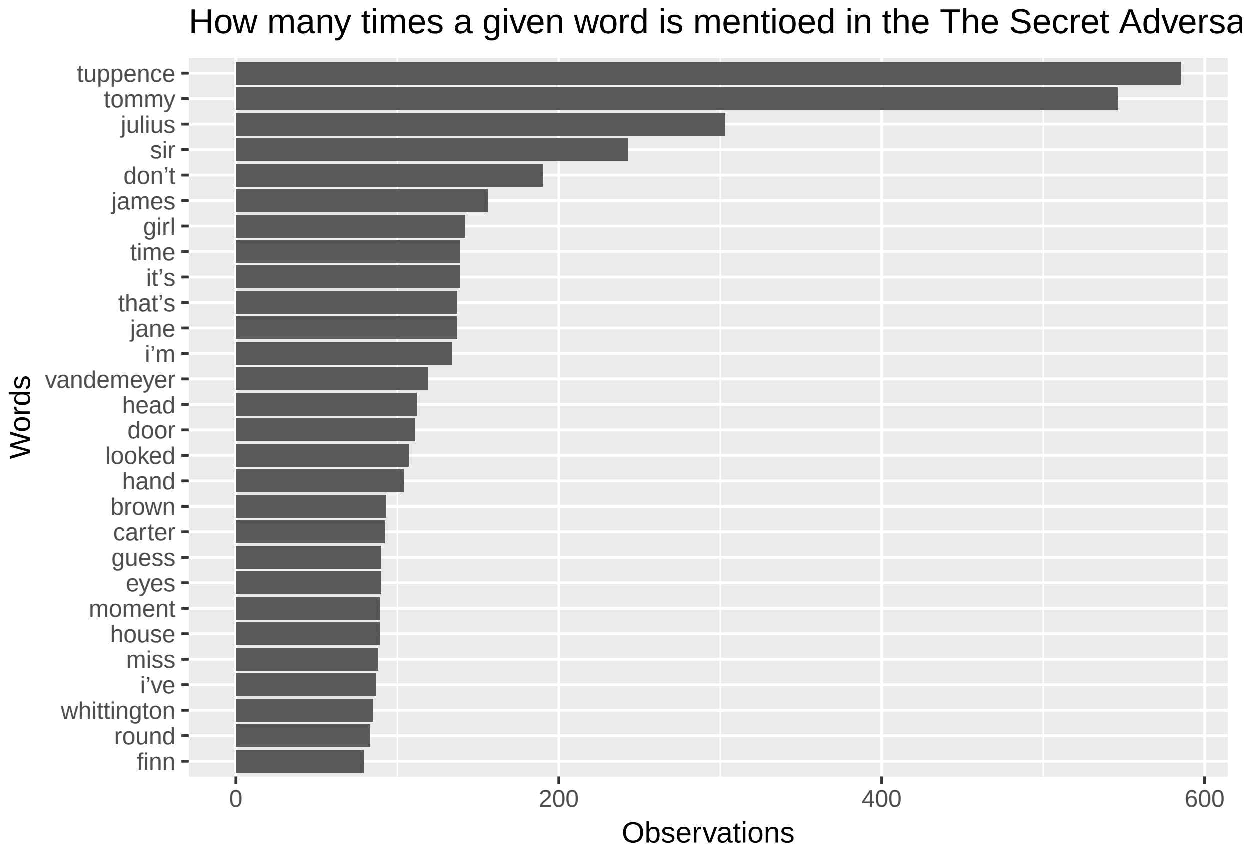
```
word_counts %>%
  filter(n > 50, title == "The Adventures of Tom Sawyer") %>%
  mutate(word = reorder(word,n)) %>%
  ggplot(aes(word, n)) + geom_col() +
  labs(title = "How many times a given word is mentioed in the The Adventures of Tom Sawyer by Mark Twai
  coord_flip()
```



Which are the most frequent word in The Secret Adversary by Agatha Christie?

See the plot below:

```
word_counts %>%
  filter(n > 75, title == "The Secret Adversary") %>%
  mutate(word = reorder(word,n)) %>%
  ggplot(aes(word, n)) + geom_col() +
  labs(title = "How many times a given word is mentioed in the The Secret Adversary by Agatha Christie"
coord_flip()
```



Another type of plot:

We should calculate the frequencies of words by book:

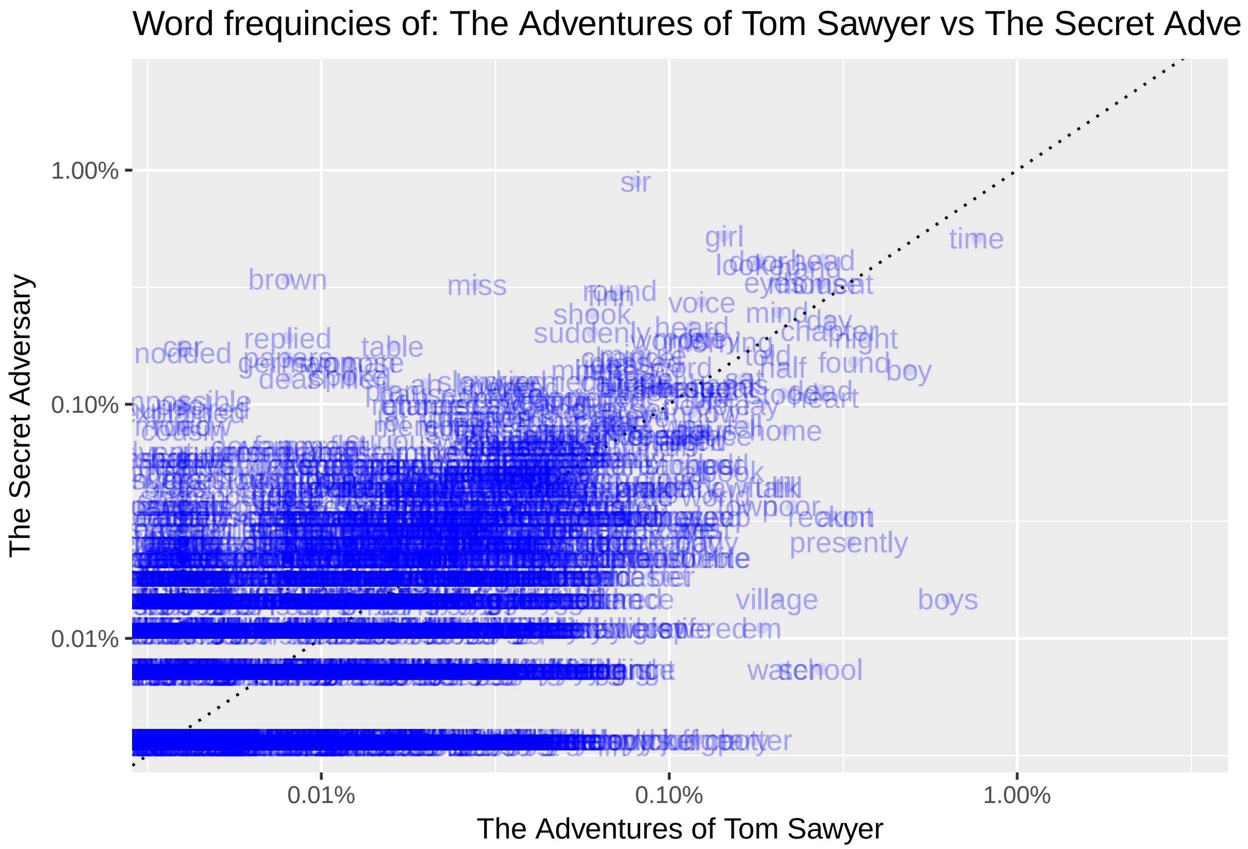
```
word_counts2 <- word_counts %>%
  group_by(title) %>%
  mutate(proportion = n / sum(n)) %>%
  select(-n) %>%
  spread(title, proportion)

head(word_counts2, 10)
```

```
## # A tibble: 10 x 3
##    word       `The Adventures of Tom Sawyer` `The Secret Adversary`
##    <chr>                               <dbl>                  <dbl>
##  1 _a                             NA                      0.0000369
##  2 _ad                            NA                      0.0000369
##  3 _ain't_                         0.000120              NA
##  4 _all_                           0.0000801             NA
##  5 _always_                        0.0000400             NA
##  6 _anatomy_                       0.0000400             NA
##  7 _and                           NA                      0.0000738
##  8 _any_                           0.0000400              0.0000369
##  9 _any_body                       0.0000801             NA
## 10 _anything_                      0.0000400             NA
```

Then, we can inspect the results by piloting the word frequencies together:

```
ggplot(word_counts2, aes(x = word_counts2[,2], y = word_counts2[,3])) + geom_abline(slope =1, linetype=
```



Brown, miss, sir , round, etc. are frequently used in The Secret Adversary then in The Adventure of Tom Sawyer, while boys, school, water, village are more frequent in Mark Twain's book.