Strings Lab

David Gerard

2019-02-25

Learning Objectives

- Practice `stringr`, `dplyr`, and `ggplot2`.

Exercises

For this exercise, we will consider the works of Jane Austen as stored in the janeaustenr package.

```
library(janeaustenr)
bookdf <- austen_books()
```

`bookdf` (which we created above) is a data frame that contains a line of text and the book from which that line belongs.

1. Populate `bookdf` with line numbers (so at the start of each book, the line number begins at 1).
2. Add a column to `bookdf` called `new_chapter` that is `TRUE` if the line begins a new chapter and `FALSE` otherwise.
3. Read about the `cumsum()` function. Try it out on the following vector:

   ```
   c(FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, TRUE, TRUE, FALSE)
   ```

   What do you think `cumsum()` does when evaluated with a logical vector?
4. Use the `cumsum()` function (as well as other functions) to find the chapter number of each line of text. Add this as a new column to `bookdf` called `chapter`.
5. Apply this code to get one word per row in the data frame `janedf`:

   ```
   library(tidytext)
   bookdf %>%
     unnest_tokens(word, text) ->
     janedf
   ```

6. Use stringr and regular expressions to create shortened titles for the books that contain just the first characters of each word. For example, `"Sense & Sensibility"` should change to `"S&S"` while `"Emma"` should change to just `"E"`. Add these shortened titles to the `janedf` data frame.

   Hints: I used `str_replace_all()` for this question. Try making the regex changes on this data frame then use joining.

   ```
   book_title <- tibble(book = levels(janedf$book))
   ```

7. Is there an association between word length and book? First, calculate the proportion of words of each length for each book.

   Now use `geom_line()` to plot the word length against proportion of words, color coding by book.

8. From the `bookdf` data frame, create a data frame with two columns, `book` and `text`. There should be only six rows, and each element in `text` should contain the entire text from the book in `book`.

9. Create a function that will take as input a string and return another string where the name of any Bennet sister mentioned is preceded with "mecha". The Bennet sisters are Elizabeth (Eliza or Lizzy), Mary, Kitty (Catherine), Lydia, and Jane.

   For example, in my implementation, `mechabennet()`, I have the following outputs:

```
text <- "Elizabeth passed quietly out of the room, Jane and Kitty followed"
mechabennet(text)
```

```
## [1] "mecha-Elizabeth passed quietly out of the room, mecha-Jane and mecha-Kitty followed"
```