Stat 610 Homework 1

Due Thursday, September 15, 11:59pm

Background

For this assignment, you will use your text processing skills to extract information from some novels by Charles Dickens. We will try to find how many named characters there are in each book using regular expressions.

We will assume that any named character in a Dickens novel will be referred to at least once by their honorific (https://en.wikipedia.org/wiki/English_honorifics), and so we can restrict ourselves to names of that form. We will further restrict ourselves to names starting with Mr., Mrs., Dr., Ms., and Miss. After the honorific, a name contains one or more words, each of which is capitalized. Names can also include abbreviations, so the regular expression you use for finding names should also include names like Mr. E. S. Baker.

The data you will be using are books that I downloaded from Project Gutenberg. I've done a little reformatting for you, and I have removed the backmatter. The primary difference between the Gutenberg files and our files are that the files on the website have one paragraph per line instead of having line breaks within the paragraphs. Each book is a text file, which you can look at in a text editor before you get started.

Assignment

Your assignment is as follows:

- Download and unzip the `books.zip` data file from jfukuyama.github.io/teaching/stat610/assignments/books.zip, and put the contents in a folder called `books` in whatever folder you're using as a working directory for this assignment. What this means is that if you run the command `list.files('books')`, the output should be

```
 [1] "1289-0.txt"  "1400-0.txt"  "564-0.txt"   "580-0.txt"   "653-0.txt"
 [6] "675-0.txt"   "678-0.txt"   "700-0.txt"   "766-0.txt"   "786-0.txt"
[11] "821-0.txt"   "882-0.txt"   "883-0.txt"   "917-0.txt"   "963-0.txt"
[16] "967-0.txt"   "968-0.txt"   "98-0.txt"    "pg1023.txt"  "pg19337.txt"
[21] "pg730.txt"
```

- Create a data frame (empty for now) with columns `title`, `n_words`, `n_chars`, and `n_individuals`. It is best to pre-allocate it to be the correct size, so it should have 21 rows (the same as the number of books we're looking at).
- For each book, find the title, count the number of words, and count the number of characters (using the `nchar` function), and save that information in the data frame you created.

  Hint: To read in the text corresponding to one of the books, you can use the `readr` package. Use `install.packages("readr")` to install it. Then you can use the `read_file` command

to read the text file in as a string to R.

```
library(readr)
b = read_file("books/1289-0.txt")
```

I would recommend using the `list.files` command above along with a for loop to read in the files.

- Create a regular expression that matches names of the sort described above.
- For each book, use your regular expression to find all the names of the characters. Find the number of unique names (the `unique` function might be useful here) and put it in the `n_individuals` column of the data frame you created at the beginning.
- Print out the data frame you created. It should now be filled with information on the titles of the books, the length (measured in words and characters), and the number of named individuals.
- For each book, print out the set of unique names that you found using your regular expression.

Submission parameters

You should submit an Rmd file and the corresponding pdf on canvas. The files should contain both the code you ran and the answers to the problems.