HUDM 5026 - Introduction to Data Analysis and Graphics in R
HW 08 – Strings and Regex

Instructions.

- Use R Markdown to create an html document with the homework tasks.
- You are encouraged to discuss problems with classmates, but all work you submit must be your own.
- As always, any plots should have appropriate axis and overall labels.

For this HW you will work with the text from the famous novel, *Alice's Adventures in Wonderland*, by Lewis Carroll. Navigate to Project Gutenberg to view a text file version of the novel here `https://www.gutenberg.org/files/11/11-0.txt`. Read the file into R using `read_file()` from package **readr**, which is part of the tidyverse.

```
aiw <- read_file(file = "https://www.gutenberg.org/files/11/11-0.txt")
```

Task 1 *How many characters does `aiw` contain? Is it a vector? What is the length of the vector?*

Task 2 *Examine the first 3000 characters of the text of the book with `str_sub()`.*

Task 3 *Note that when imported, the text file brought along some formatting in the way of `"\n"` and `"\r"` strings. Find and replace these white space using `str_replace_all()` and save the output as `aiw2`.*

Task 4 *Project Gutenberg has added some text at the beginning and appended some legal boilerplate text to the end. Find the true beginning and ending of the text and save only the book text (i.e., drop off the boilerplate). Save this as `aiw2`.*

Task 5 *Note that `aiw2` is a single long character string. We would like to separate the words so that each is it's own element of a character vector. So, once Task 2 is accomplished, use `str_split()` to do that and rename the result to `aiw3`. This will produce output in a list, but we want a vector, so use `unlist()` on `aiw3` to turn it into a long character vector; again, save the output as `aiw3` here.*

Task 6 *What proportion of words in the book contain at least one uppercase letter?*

Task 7 *What proportion of words in the book use some form of punctuation?*

Task 8 *Alice in Wonderland is one of the most influential books for children written over the past 200 years. As an example, ever wonder why Nintendo's Super Mario grows when he eats a mushroom? Although Shigeru Miyamoto (Mario's creator) has more recently denied a direct influence, he has spoken in the past about how he was thinking about Alice in Wonderland while creating the video game. How many times does the word "mushroom" occur in the text, regardless of letter case? Locate each instance of the word and print out enough of the surrounding text to display the context for each instance.*

Task 9 *Use a function we have discussed to replace all instances of the word 'mushroom' with the word 'apple'.*