BERMMC010 Programming - Assignment 2

In this assignment we will be working with textual data. You can retrieve this data using a given function in **assignment-2.R**. This function will download all books written by Jane Austen. This function will attempt to install the gutenbergr package to do this. In case this does not work, you may have to manually install it.
In most cases you can answer the question by providing the code that will give the answer. If a textual answer is required, add it as a comment to the code. You will be graded on correctness as well as proper style (check https://style.tidyverse.org/).

Part 1 – Scaffolding

1. Download the assignment-2.R file from Canvas, put it in the same project as before.

2. Create a new branch called **assignment-2** specifically for this assignment and work on that branch for the remainder of the assignment. Commit and push the **assignment-2.R** and **austen_word_freqs.Rds** files into this new branch.

3. While doing the assignment, try to commit every time you do a question from the assignment. It's not the end of the world if you forget sometimes, but it is good to practice with small and regular commits.

4. Each question will ask you to create a function. Add those functions to the ones already in the file.

5. Use RStudio's option to insert Roxygen comments to add documentation to each function.

Part 2 – Data Analysis

1. Create a function called **tidy_df** with two input parameters, the first a data frame, the second a column prefix. This function should automatically tidy the data frame by gathering all columns that start with the given column prefix.

   For example, if a data frame has column var1, var2, and var3, this function should add a column "**variable**" containing either "var1", "var2", or "var3" and a column "**value**" that contains the value that previously was in the "var1", "var2", or "var3" column.

   Be sure to leave columns that do not start with the given prefix as is.

2. Create a function called **extract_possible_names** that uses a regular expression to extract all words that start with a capital letter. Use the function **get_jane_austen_data** to get a data frame of textual data.

   Your function should return a data frame with one row per extracted name and

a. A column “id”, a unique identifier for this name
b. A column “text_id” which is the “id” from the original data frame to link back
c. A column “name” which is the extracted name

The **tidy_df** function you created before might come in handy here…

3. Unfortunately, your function also extracts words that are capitalized because they are the first word of the sentence. Create a function called **filter_names** to filter out those words. For this task, you can use a data frame of word frequencies, given in **austen_word_freqs.Rds**. Keep only names that appear capitalized at least 75% of the time.

4. Which book by Jane Austen contains the highest number of unique names? And which book contains the most occurrences of names? Use your list of names from the previous assignment. Answer this question by creating a function called **count_names_per_book** which returns a data frame with a column containing the book titles (**title**), a column with the number of unique names per book (**unique_names**), and a column with total number of name occurrences per book (**name_occurrences**).

Part 3 – Submission

Find the URL of your final commit in the **assignment-2** branch on GitHub. Submit this URL on Canvas.