ISM 4403 Homework Week 7

### **Tasks:**

The goal of this assignment is to output each step’s intermediate data in a suitable format. The files must be named in the following format. The book field may be omitted for summary data.

Format template:

homework7.{BOOK}.{STEP}.{YYYYMMDD.HHmmS}.txt

Example:

homework7.TheMucker.1.20190811.145605.txt

1. Use R to download the collective works of Edgar Rice Burroughs as text documents from Project Gutenberg (<https://www.gutenberg.org/ebooks/author/48>) making sure to name the files using the format mentioned above.
   1. Hint: I recommend using the curl library to do this.
   2. Hint: You may also you Gutenbergr, but we want to see how you save the files to the computer, which still requires the use of Curlr**. check**
2. Use R to lowercase all text in the documents lower casing all words.
   1. Hint, a to lowercase function or library can perform this task.
3. Use R to munge all data in the text documents removing all punctuation.
   1. Hint munge means to clean the data.
   2. I recommend splitting each word in each book at this point using a function such as strsplit. (split on new lines)
4. Determine the frequency of each word in each book.
   1. Use a function such as “count”, or Tidy structures.
5. Determine all words that occur in more than one book and generate a list containing unique words for each book and their frequency in the book.
   1. Use a function or structure such as Tidy or a function such as diff.

Rubric

20 points for downloading all works from project Gutenberg. It is recommended that you place a link for each book into a text file and use this to import the data.

20 points removing all punctuation and writing it to an intermediate file for each book.

20 points lowercasing the documents for loading the punctuation free file and lowercasing each word.

20 points for determining the frequency of each word in the books and storing it in an intermediate file.

20 points determine all words that occur in more than one book and generate a list containing unique words for each book and their frequency in the book. Then storing this in an intermediate file. This file should combine the results of all the books into a single file.