**ISMG 6470 Text Data Analytics - Assignment.**

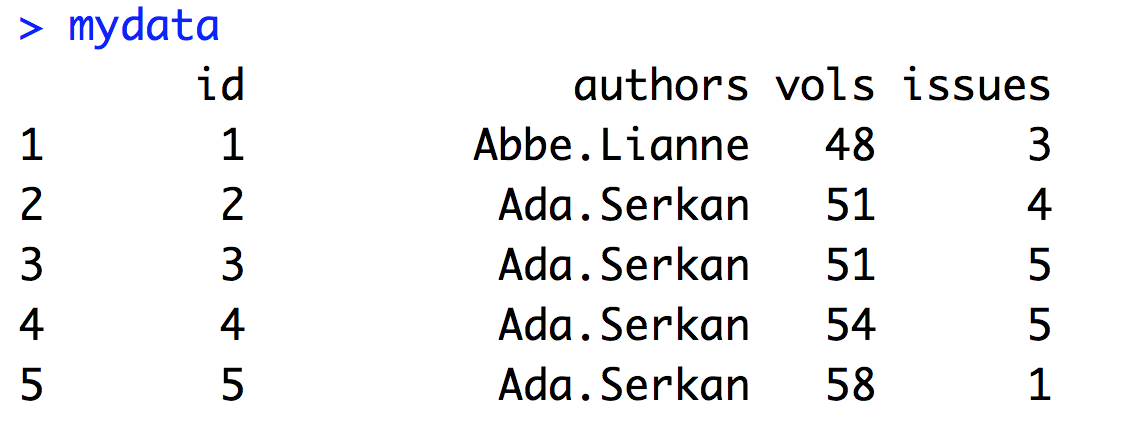
For the below questions, submit two separate answers: (1) MS Word file that answers all the questions, and (2) R code file that you used to answer the questions. The MS Word file and R code file should be submitted in the below format:

1. MS Word file:
   * 1. At the top, include the course ID, course name, student ID number, student name, and submission date.
     2. In the body, do not remove the questions. Instead, under each question, briefly describe your answer. Per each answer, provide (i) brief description on your answer (ideally in less than 2 lines), and (ii) R codes that might be used to answer the question.
2. R code file:
   1. At the top of the answer-code, make comments (using # appropriately) for your name, your student ID number, submission date.
   2. Make appropriate comments to describe what each line of code is to perform.

This assignment uses total 1,323 text files. Please download the zipped assignment file from Canvas. The zipped data file (named EnglishAbstract.zip) comprises total 1,323 text files. Each text contains an abstract of a journal paper, and each file name is structured like: LastName.FirstName\_Volume\_Issue. Each file name contains three pieces of information. For example, the file name, Abbe.Lianne\_48\_3.txt indicates that the journal paper was written by the author, “Lianne Abbe” (i.e., LastName.FirstName), and published in Volume 48 and Issue 3. To complete this assignment, you need to first build a corpus by pre-processing all text documents. Please report answers as requested below.

Q1) First, build a corpus with all the 1,323 text data. And then,

1. read all the file names into RStudio,
2. create four separate objects, each of which represents “id number,” “name,” “volume,” and “issues” respectively.
3. create a data frame by using those four separate objects where each object represents a separate column. (So, the data frame should have 1,323 rows and 4 columns in a below format).



Q2) Report a table that shows how many journal papers (i.e., total number of “issues”) have been produced per each volume.

Q3) Report who is the most prolific author (i.e., who published the highest number of journal papers)?

Q4) Report the followings by using the corpus, which is NOT pre-processed yet.

1. Aggregating all the abstracts, report the mean number of characters, minimum number of characters, maximum number of characters, and standard deviation of characters.
2. Aggregating all the abstracts, report the mean number of words, minimum number of words, maximum number of words, and standard deviation of words.

So, total 8 numbers should be reported.

Q5) As described below, perform the 4 steps pre-processing for the corpus that you built above.

* Remove all numbers
* Remove all punctuations and special characters
* Remove all stop words (by using SMART dictionary)
* Apply the stemming (by using Porter’s stemmer)

And then, report the followings by using the pre-processed corpus.

1. Aggregating all the abstracts, report the mean number of characters, minimum number of characters, maximum number of characters, and standard deviation of characters.
2. Aggregating all the abstracts, report the mean number of words, minimum number of words, maximum number of words, and standard deviation of words.

So, total 8 numbers should be reported.