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 same format of previous weeks’ assignments.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For this assignment, please use the same 1,322 text data files that were used for the Week 4 assignment.

Q1) Using the ‘tm’ package, perform the following three data preprocessing:

1. Remove all numeric data.
2. Remove all special characters and punctuations.
3. Apply the Porter’s stemming.

**\*\* Note**: You don’t need to perform additional or customized preprocessing. Just use the ‘tm’ package to perform the above three preprocessings.

Q2) Using the preprocessed data, build a document-term-matrix to store *term-frequency* (*tf*). Each cell of this matrix should present the raw frequency of each term that appears in each document.

1. Report (i) the highest term frequency value, (ii) its corresponding term, and (iii) its corresponding document’s file name. (Attach a screen shot in MS document)

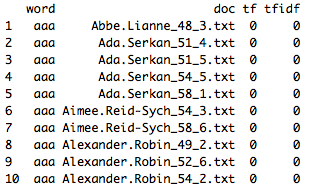
Q3) Build a document-term-matrix with the preprocessed data. Each cell of this matrix should present the *tf-idf* value.

1. Report (i) the highest *tf-idf* value, (ii) and its corresponding term, and (iii) its corresponding document’s file name. (Attach a screen shot in MS document)

Q4) Build a data frame that has four columns like: (1) terms, (2) document name, (3) *tf* values, and (4) *tf-idf* values. (Note: The total number of observations in the data frame should be equal to the multiplied value of the total number of rows and the total number of columns, both of which can be obtained from the document-term-matrix that you built in Q2 or Q3).

1. Report the top 10 observations of the data frame (Attach a screen shot in MS document).
2. Report the bottom 10 observations of the data frame (Attach a screen shot in MS document).

\* Note: If done correctly, the final data frame should look like Figure 1 below.



(Figure 1: Sample data frame)

Note: Q5 and Q6 need to use aggregate() function to get *sum* and *mean* values from the data frame that you have built so far. Watch this video tutorial to learn the aggregate() function: <https://www.youtube.com/watch?v=zmiC7X9fUmo>. While this 8 minutes video tutorial might be enough to answer Q5 and Q6, if necessary, please do additional web search to study aggregate() function.

Q5) Aggregate the above data frame (built from Q4) to get summed *tf* score and *tf-idf* score per each document. (As you will aggregate the data frame in Q4 according to each document, the resulting data frame should have 1,322 row and 3 columns where each column represents document name, summed *tf* score, and summed *tf-idf* score). To this 1,322 by 3 data frame, bind additional three columns of (a) author name, (b) volume number, and (c) issue number. Once done with the binding, the final data column might have the dimension of 1,322 by 6.

1. Report the top 10 observations of the resulting data frame (Attach a screen shot in MS document).
2. Report the bottom 10 observations of the resulting data frame (Attach a screen shot in MS document).

Q6) As a result of aggregation in Q5, you might have a data frame having 1,322 by 6 dimension. Now, aggregate the data frame again to get mean *tf* score and mean *tf-idf* score per each author.

1. Report the top 10 observations of the resulting data frame (Attach a screen shot in MS document).
2. Report the bottom 10 observations of the resulting data frame (Attach a screen shot in MS document).

Q7) To the data frame that you built in Q6, bind additional two columns: One column to represent the total number of publications that each author published, and another column to categorize the authors into one of three different groups in order to indicate their publication productivity. Categorize the author group as ‘1’, if they made a total of one publication. Categorize the author group as ‘2’, if they made a total of 2 or 3 publications. Categorize the author group as ‘3’, if they made 4 or more publications.

1. Report the top 10 observations of the resulting data frame (Attach a screen shot in MS document).
2. Report the bottom 10 observations of the resulting data frame (Attach a screen shot in MS document).

\*Note: The resulting data frame might have 5 columns like: “authors”, “tf-mean”, “tfidf-mean”, “total number of publications”, and “publication productivity”.