**CS7070-Big Data Analytics**

**Spring 2019**

**Programming Project Assignment #2 and #3**

**Due Dates: Q1 and Q2: March 4th, 2019 (9PM); Q3 and Q4: March 8th, 2019 (9PM)**

In this programming assignment, you are expected to use and modify the MapReduce programs for computing the TFIDF for terms in a set of documents. The attached zip file contains programs for computing TFIDF as per the description given by Lalit in the class last week. It also includes a text file that should be used for testing your programs. The tasks to be accomplished by you are:

1. (20) Execute all phases of the TFIDF program and submit the following items:
   1. TFIDF for all terms in each document, sorted alphabetically by words’ letters, and formatted for easy readability.
   2. List of top fifteen words from each document having the highest TFIDF value. This selection of top 15 words may be done manually or by a sequential program.
2. (20) Modify the programs to remove from consideration all those words that occur only once in each document. Repeat the tasks of Q1 above. Comment on any changes in the results of part (b).
3. (30) Now consider a “Term” to mean a 2-gram (two words occurring sequentially) in a document. Modify the programs given to you to compute the TFIDF for each Term (2-gram). Submit the following items:
   1. List of top 15 “Terms” (2-grams) for each document, having the highest TFIDF values. The task of selecting the top 15 terms does not need to be done by the MapReduce program.
   2. Which output – obtained in 1(b) or in 2(a) – better characterizes the documents? Give reasons for your answers.
4. (20) Once your program is working for the above two parts, run the programs on a larger collection of documents (to be provided to you next week) and repeat the above two tasks. Discuss the results for 1(b), 2(a), and 2(b) in the context of the new set of documents.
5. (5+5) Well organized and clearly understandable presentation of results in the submission.