**CS7070-Big Data Analytics**

**Spring 2019**

**Programming Project Assignment #1**

**Due Date: Feb 7th, 2019**

In this programming assignment you are expected to write a number of variants of a Map-Reduce program to produce the following outputs. You will have available to you a collection of textbooks on the departmental server on which you will run your programs and generate outputs. The tasks are:

1. (30) Design and execute a MapReduce program to produce the frequencies of all the words in the book collection, retaining only those words whose frequencies are greater than 5. Submit the following:
   1. Your commented code for the Mapper and Reducer
   2. First 50 words and their frequencies from your program’s output file.
   3. Last 50 words and their frequencies from your program’s output file.
2. (30) Design and execute a MapReduce program to produce frequencies of all 2-gram word-pairs in the book collection, retaining only those 2-grams whose frequencies are greater than 5. Submit the following:
   1. Your commented MapReduce program to produce the frequencies of all the 2-grams in all the books.
   2. First 50 2-grams and their frequencies from your program’s output file.
   3. Last 50 2-grams and their frequencies from your program’s output file.
3. (30) Design and execute a MapReduce program to produce the top 100 most frequent words in the book collection. You may need two rounds of Map and Reduce processors. Submit the following:
   1. Your commented code for the Mappers and Reducers.
   2. The list of 100 most frequent words and their frequencies in the book collection.
4. (10) For this submission all parts of your work must be clearly marked and documented. The grader must be able to:
   1. Clearly see your code segments with proper labels
   2. Clearly see a labeled document listing the sequence of commands you used to execute the programs
   3. Clearly see the labeled outputs from the programs.
   4. All codes/documents/outputs must be combined to form a single pdf file. The grader will not be able to unzip the multiple independent files.