# Assignment 2, MGS 655

**Project Description:** Recall, a **trigram** is a sequence of three consecutive words in a sentence. For example, for the sentence "It is raining today" the following are valid trigrams: "It is raining", "is raining today". The order of sequence of words matter i.e. if there are three words $w_1, w_2$ and $w_3$, the bigram ($w_1$, $w_2$, $w_3$) and ($w_2$, $w_1$, $w_3$) are not treated identical. Your goal in this project is to design a map and reduce function for finding trigrams in text.

The following details are relevant for the project:

1. You are required to design and implement the mapper and reducer class along with other associated functions for generating trigrams from text.

2. The program will have to be tested on the SAME data set you used in Project 1. There is, however, a slight variation in data that is recommended for Project 2. Suppose there are 5 text files $f_1, f_2, f_3, f_4, f_5$. You are required to randomly choose any two files and replicate them. Suppose you randomly chose $f_3$ and $f_5$ to be replicated. Your data set should now be $f_1, f_2, f_3, f_4, f_5, f_6, f_7$ where $f_6, f_7$ are simply identical copies of $f_3$ and $f_5$ respectively.

3. Your program is expected to run on the CCR cluster.

4. As in Project 1, you will be required to generate a 2-page report. This, along with the code you submit will be used for the purposes of grading.

5. You are expected to comment on the following:

    - How many "unique" trigrams were identified by your program?
    - Does a change in the number of mappers and reducers affect the run time of your program? For e.g. consider two scenarios: (a) There are 2 mappers and 1 reducer. (b) There are 2 mappers

and 2 reducers. Evaluate the two methods by first comparing the output. Can you comment on how fast either one of them ran? Please present time taken for execution of the program on the CCR cluster.

- Did you have substantially long wait times before your job was processed?