**COSC 5010 Cloud Computing and Big Data**

**Homework-1**

**Due Day: 02/20/2015**

Please implement a Mapper/Reducer using Hadoop. The input text, is available at the file input.txt.

For each word in the input text, please compute the occurrences of other words appear in every line that it appeared. For example, assume that we have the following line as an input text.

*I had seen little of Holmes lately. My marriage had drifted us*

With the word *I*,

*• had* occurs twice*.*

*• seen* occurs once*.*

*• little* occurs once*.*

*•* and so on.

With the word *had*,

• *I* occurs once.

• *had* occurs once. (not zero times!)

• and so on.

For this assignment, you are going to compute the occurrences of every word for all lines, not just for a single line.

You must follow these guidelines to obtain correct result.

• Convert every string into lower-case.

• Get rid of all punctuation, i.e., any character other than a to z and space. It will be helpful to use a regular expression to replace them with empty strings. Of course, this will result in some non-sense situations (im, youre, and etc.), but you do not have to worry about this.

• Do not count the subject word. In other words, if you are computing the occurrence of words for a word *the*, you should NOT count the same *the*. Note, however, that you should count other *the*’s appear in the same line if there are any!

**Please write a readme file showing how to run your code. Please Zip all files (including the source code) into a file then upload it on the course website**