**CSC 4760/6760 Big Data Programming**

**Assignment 3**

**Due Date: 11:59 pm, Wednesday, Feb 28, 2018**

Problem 1. (100 points) (Setting up Spark and running the WordCount example)

This assignment aims at letting you learn how to setup Spark on your laptop or desktop. After the installation of Spark, you need to run the WordCount (Python version) example on your computer.

Please follow the instructions provided in the slides “12 Setup Spark on Ubuntu.pptx”. If you have any questions, please talk with the instructor or the TA. We will help you setup Spark on your computer.

The Ubuntu virtual machine image has been updated. In the new version, we have installed Spark. The PySpark is ready to use. Please download it from Google Drive. It is given 20GB disk and 4GB memory in my machine.

**Source Code and Datasets:**

The Python source code is given in the file “WordCount.py”. You need to run it on two datasets:

1) test.txt (display the top-5 most frequent words)

2) peterpan.txt (display the top-30 most frequent words)

The example commands are as follows.

$ spark-submit WordCount.py /home/rob/Assignment3/test.txt 5

$ spark-submit WordCount.py /home/rob/Assignment3/peterpan.txt 30

**Report:**

Please write a report to explain the key steps. Please take the screenshots of the outputs in the terminal for “test.txt” and “peterpan.txt” respectively. Please put them in the report and explain the outputs briefly. You may include the following key steps.

1) Download the updated Virtual Machine Image or Setup Spark by yourself.

2) Download the “WordCount.py” file and two input data files from iCollege.

3) Open a terminal, and run the “WordCount.py” file on “test.txt” and “peterpan.txt” respectively. You need to explain the commands and the outputs.

**Required submission materials:**

a) The report should be a PDF file. Please use a text editor, such as Microsoft Word, to write a report. Please transfer the file into a PDF file and then submit it. The name of the file should be “Assignment3\_LastName.pdf”.