CSC 4760/6760 DSCI 4760 Big Data Programming

Assignment 2

Due Date: 10PM ET, 10/30/2022

Problem 1. (100 points) - **Setting up Spark on Linux and running the WordCount example in Python using a Jupyter Notebook**.

Please follow the instructions provided in the Apache Spark Tutorial Videos in iCollege.

Source Code and Datasets:

The Python source code is given in the video titled “4- Spark Word Count in Python”. You need to run it on two datasets:

1) test.txt (This example is provided in the video)

2) peterpan.txt

Report:

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

1) How you setup Apache Spark in your machine and verified the setup.
2) How you ran wordcount in PySpark (using a Jupyter Notebook) on the two datasets provided.

Required submission materials:

a) The report should be a PDF file. The name of the file should be “Assignment2_LastName.pdf”.