CIS*6030 Information Systems

Fall 2022
Instructor: Fangju Wang

Assignment 2 (100%)

For this assignment, you are required to write two Java programs by modifying the WordCount.java program (from hadoop.apache.org), which has been posted on our Moodle page.

Before you write the programs, you should install Hadoop on your computer, then load the A1_data.txt (used in Assignment 1) into the HDFS.

Question 1 (50%)

Modify the WordCount.java program so that the program can find all the strings longer than 25 characters in A1_data.txt, and output each of those strings and its frequency. The output can be something like

```
agriculturalcommunication 1
agriculturalcommunications 1
compassionateconsideration 1
......
```

question 2 (50%)

Modify the WordCount.java program so that the program can count the total number of the strings in A1_data.txt, and output the total number in the form of key-value pair, like

```
Total number: 468823.
```

Please submit your work as a tar file. In your submission, please include a readme file, telling how to compile and execute your programs, where the input and output directories are, and how to check the program outputs. Don't submit A1_data.txt with your files.

Due time: 23:59, Monday October 17, 2022. Submit your work to Moodle. **NO late assignment will be accepted.**