CS 8001 Big Data

HW #2: Word Count on Spark (10 points + 2 bonus points)

Spring 2016

(Due 2/3, Wednesday, midnight)

In this exercise, you will write Python code on Spark that outputs the 20 most common words in the Complete Works of William Shakespeare retrieved from Project Gutenberg and the number of occurrence of each of them. More details are here.

Part I (10 points)

Run Spark locally with one worker thread.

Submit electronically in Blackboard the following:

1) A brief description of your Spark environment, your code structure, and the execution process of the code.
2) Output of your code (20 most common words and their counts).
3) Execution time of your code in seconds.
4) Comparison of the implementation and results of this exercise with those of the HW1
5) Your Python code with appropriate comments.

Part II (2 bonus points)

Run Spark on a cluster in Amazon AWS.

Submit electronically in Blackboard the following:

1) A brief description of your Spark environment and the execution process of the code.
2) Output of your code (20 most common words and their counts) if it's different from the output in Part 1.
3) Execution time of your code in seconds.
4) Comparison of the execution result with that in Part I