Assignment 3 - Spark

1. Launch the Spark shell.
2. Make a parallel collection of `Array(1, 2, 3, 4, 5)` and sum up all its elements.
3. Create an RDD named `pagecounts` from the given input file *hamlet*.
4. Get the first 10 lines of *hamlet* (i.e., first 10 records of `pagecounts`).
5. Make a more readable print of the step 4.
6. Count the total records in the data set `pagecounts`, and confirm its correctness by comparing the result with the Bash `wc` command: `wc -l hamlet`.
7. Monitor the jobs through the web interface.
8. Filter the data set `pagecounts` and return the items that have the word *this*.
9. Cache the new data set in memory, to avoid reading from disks.
10. Find the lines with the most number of words.
11. Count the total number words.
12. Count the number of unique words.
13. Count the number of each word.
14. Save the data set in a text file.
15. Collect the word counts in the shell.
16. Write a standalone application in Spark to count the total number of words in the input file *hamlet*.
17. Follow the instructions in the given source codes and run the word count applications on MapReduce and Stratosphere.