Problem solving with MapReduce

(30 points) Problem 1: Word Count

Create a Java implementation of the Word Count example discussed in class. Recall that you need to implement the Map and Reduce functions.

Some Suggestions:

1. Hadoop uses data types from the org.apache.hadoop.io package instead of the standard Java data types (String, Int, etc.). Check Hadoop's API at: http://hadoop.apache.org/docs/r2.4.0/api/
2. Use smaller file examples for testing your implementation.
3. Run your fully debugged version on a large text. Here is an example, Pride and Prejudice, by Jane Austen: http://www.gutenberg.org/files/1342/1342-h/1342-h.htm

Additional testing datasets are the Bible and the complete works of Shakespeare available at http://lintool.github.io/Cloud9/docs/exercises/bigrams.html

Deliverables (on Pride and Prejudice):

- Run your implementation with 5 reducers. You should see five *part-r-XXXXX* files, because we have 5 reducers. Now copy the data from HDFS onto the local disk:
  - E.g, hadoop fs -get wc/part-r-*
- You should be able to examine the contents of the file using normal shell commands.
- Give the total number of unique terms.
- Give the fifth to last term (need to sort the list) and its number of occurrences.
- Give the first term and its number of occurrences.

(45 points) Problem 2: Count Words with Letters

Modify your Word Count implementation so that only words consisting entirely of letters are counted. The word must match the following Java regular expression: "[A-Za-z]+"

Deliverables (on Pride and Prejudice):

- Use 5 reducers. (We you can experience with different number of reducers.)
- Give the total number of unique terms.
- Give the fifth to last term (need to sort the list) and its number of occurrences.
- Give the first term and its number of occurrences.

(15 points) Problem 3: The file sample.dat has two blocks A and B, explain in HDFS how this file is written to a Hadoop cluster with one namenode and three datanodes in a default configuration.

(10 points) Problem 5: In Quorum Consensus algorithm, Reads go to a read quorum of size R and writes go to a write quorum of size W. For a group of 5 replicas, explain and compare the following three possibilities:

1) R=5 and W=1

2) R=1 and W=5

3) R=3 and W=3