CPSC 4330 Big Data Analytics

In-class Exercise 8 – Hadoop Streaming

Task 1:
Write a shell script that will print the N most frequent terms in a subset of the corpus: the Shakespeare plays. The Shakespeare plays are all files in 'textcorpora' (you can download 'textcorpora' from Canvas) which filename begins with 'shakespeare-'.

- Your script should emit N records to the terminal of the form <term><space><count>. The parameter N will be an argument to your script.
- Your solution must use the programs that have been provided to you (posted on Canvas). See below.
    - the WordCount MapReduce application WordCount.java
    - the script compile-map-reduce
    - the script run-map-reduce
- Your script must "clean after itself" – by deleting all files it created.

Task 2:

The file Hadoop_2k.log (posted on Canvas) contains log entries from a Hadoop run sampled for ten minutes (Oct 18, 2015, between 6:01PM and 6:10PM). Each log entry has a severity level, which is one of INFO, WARN, ERROR, and FATAL.

Use Hadoop Streaming to process these log lines to record, for each minute, the number of log entries in each category.

Your script, named process-log-file, will copy the input file to HDFS, then run Hadoop streaming to produce a file with one line for each minute in the log; each line should have these 6 fields:

1. Minute number (a number between 1 and 10)
2. Total number of log entries for that minute
3. Number of log entries for that minute with severity INFO
4. Number of log entries for that minute with severity WARN
5. Number of log entries for that minute with severity ERROR
6. Number of log entries for that minute with severity FATAL

Your script will then print the lines of the output file to standard output, in ascending order of minute number.

Use the run-hadoop-streaming script (posted on Canvas) to structure your solution: create the directory log-processing containing your mapper and reducer, then your script process-log-files will call run-hadoop-streaming to kick off the log processing job, and display (write to standard output) the properly sorted output.