Distributed Systems: Lab Assignment 1 on Hadoop and MapReduce

Complete the installation of Hadoop on your machine as discussed in class. If you encounter any problems with the installation, please contact the TA, as discussed in my previous email. [Note: The installation part of this assignment will NOT contribute to any grades.]

Consider a large document. You can download any non-copyrighted large document as input for this assignment.

Given the above large document as input, you need to determine the frequency of occurrence of each word in that document using the concept of MapReduce.

Show the frequency of occurrence of the words in the input document by using a suitable visualization approach e.g., word cloud, histograms. (You can use Excel or R to do such visualization.)

- You can simulate a Hadoop cluster on a single machine. Each node in the cluster is allocated one part of the document. Suppose you are simulating 3 nodes in the cluster. You can divide the input document into three parts and then allocate part 1 to node 1, part 2 to node 2 and so on.
- You can use any programming language, but must incorporate the concept of MapReduce i.e., dividing the document into parts and distributing to multiple simulated nodes on one machine.
- Do the above for at least three documents of increasing sizes. Just ensure that you are doing this assignment with large documents as input. This would give you an idea of scalability and why you need to parallelize.
- Deadline for submission of the assignment: March 7, 2018, 3 pm IST
- You need to submit only your codes. Optionally, you can also submit a brief 1-page document describing the work done in your assignment.
- Grading criteria is based on quality of code, effort, visualization of the results, correctness etc.
- This lab assignment will contribute to 10% of your overall course grade.