Project 3: MapReduce Bigrams
CSE 291 Spring 2016

Assigned: Tuesday, 17 May
Due: Tuesday, 31 May

Overview

This project is designed to give you a basic familiarity with the MapReduce paradigm via the Hadoop framework. We give you Dockerfiles to enable you to build and start a Hadoop cluster that is pre-loaded with the Hadoop word-count exampland text data files.

Your first task is to use the Docker console to start up a Hadoop cluster consisting of one Hadoop/HDFS master node and four(4) Hadoop/HDFS client nodes.

Your second task is to run the WordCount example, as is, without modification, and confirm that your cluster is working.

Your third task is to modify the WordCount example to create a BigramExample, which counts the number of bigrams within the corpus. At the conclusion of its exectution, it should output three pieces of information, one per line: (1)the total number of bigrams, (2)i the most common bigram, and (3) the number of bigrams required to add up to 10% of all bigrams.

Collaboration

You may work with up to one other person on this project.

Counting Bigrams

For the purpose of this assignment, don't get fancy. We know it it the end of the quarter. Just use static variables within the mapper to keep track of the current word and the previous word. Output a count of one for each pair, and then combine and reduce from there. This gives you a sorted histogram of bigrams.

Generating the Output

Once you have the histogram of the bigrams, it is straight-forward to determinethe most frequent, or even top-N, bigrams. If you reduce the histogram of bigrams to a count of bigrams, it will bs straight-froward to determine how many of the top-N are needed to reach at least 10% of the total.