Programming Project – TF-IDF Computation Using Map-Reduce

Group members: Sonali Sharma

Anumita Srivastava

In this Project, we have implemented the TF-IDF computation using the map-reduce paradigm. The input was taken from the map-reduce-corpus.tar.gz that contains 556 files containing Imdb movie reviews. This is a small subset of files from the Imdb data set.

The program runs in 3 steps:

* Job 1: In the first step, we will count how many words are present in the documents in the corpus, that we have merged to form one file that serves as input to the map part.
* The input in this stage would be the document formed upon merging and the output at the end would be a word and the number of times it appears in the document.
* Job 2: in this step, the total number of words will be counted in the document. The input to mapper would be <word,n> and the output would be <word, n/N> where n/N is the comparison made of each word’s occurrence with the total number of words.
* Job 3: In the final stage, we will calculate the TF-IDF. The mapper would get the input from Job 2, which would be <word, n/N>. The output from mapper for any term t1, would be <t1,n/N>
* The reducer would then calculate the TDIDF as n/N \* log(D/d) where D = total number of documents in corpus. In our case D= 556; and d = number of documents in corpus where the term appears
* The final output is shown as ( (t1@doc1), (d/556), TFIDF).

Steps of Execution:

1. We implemented the project on Cloudera CDH5 with VM Ware Fusion.

2. Steps for implementation on cluster are given as:

1) use linux scp command to send your project data to nike. e.g. scp <source> <username@destination:path>

2) log in into nike

3) scp the above data to your vm using:- scp <source> <username@vmname:directory>

4) send this data to hadoop-master.

5) Open a new terminal window and login into nike then login into vm

6) cd into hadoop-cluster-docker directory and execute ./start-container.sh

Run the following commands on your vm

1) sudo su -

2) service docker start

3) docker images

4) docker ps

5) run sudo usermod -aG docker <username>

6) logout of vm and log in again

7) docker cp <source> <container-name>:<path>

8) Use commands like hadoop jar <jar-file-name> <args>