Web Mining

Assignment 1: Hadoop, Pig and Hive (25 points, 5% weightage)

Date Posted: Aug 15, 2013
Date of submission: Aug 22, 2013. 9pm.

Goal: To introduce Hadoop and related technologies to students through simple examples.

Requirements: You must have access to some Hadoop cluster or install a 1-node Hadoop cluster on your own machine.

Submission Instructions: Create a directory with the name "<rollno>_as1". Within that create 4 directories q1,q2,q3,q4 for the four questions below. Zip "<rollno>_as1" folder to get <rollno>_as1.zip and upload it. Please do NOT upload the original datasets. Actually do not upload any other extra files, except those which have been asked for below.

Questions

1. Write MapReduce Java code for inverted index generation in Hadoop. [10 points]
   Deliverables include the following.
   a. A short description of the logic. Name the file as README.txt
   b. Mapper code
   c. Reducer code
   d. For the supplied dataset, return the inverted index file. 1 line per word. Format of the line is as follows. Word: docid1, docid2, docid3, … Name the file as invertedIndex.txt

   Put all of these in directory q1.

   Notes:
   a. Lowercase all the documents. Do not perform any other form of cleanup (don't remove punctuations or numbers etc). Do not perform any stemming.
   b. Use the file names as document ids.

   Dataset description: documentCollection.zip. This is a collection from the famous 20newsgroups dataset. It contains 16331 documents.

2. Write a Pig script to find similar patents from a patent-citation dataset. Two patents are considered similar if at least 5 patents cite both the patents. [5 points]
   a. A short description of the logic. Name the file as README.txt
   b. Pig script
   c. Similar patents file which lists similar patents. File name: simPatents.txt. File format should be as follows. Each line contains patentID1, patentsID2, co-citation count (all 3 values separated by a space).

Put all of these in directory q2.

Dataset description: `cit-Patents.zip` is a patent citation dataset obtained from http://snap.stanford.edu/data/cit-Patents.html Please read the webpage for more details.

3. Write a Hive script to find the frequency of books published each year. [5 points]
   a. A short description of the logic. Name the file as README.txt
   b. Hive script.
   c. Frequency file. 1 year per line. Filename: yearFreq.txt. Format of the line. Year:Frequency.

   Put all of these in directory q3.

   Dataset description: `BX-Books.zip is a book details dataset. Refer` http://www.informatik.uni-freiburg.de/~cziegler/BX/ for more details.

4. What do the following mean in the context of Hadoop (explain in one sentence). HDFS, HBase, Zookeeper, Chukwa, Scribe, Mahout, Cassandra , Dumbo, Spark , Sqoop, Jaql, Oozie, Giraph, Hcatlog, Fuse-DS [5 points]

   Put this as a file called README.txt in directory q4.