Instructions For Running Map Reduce Program

**How to log in by SSH terminal?**

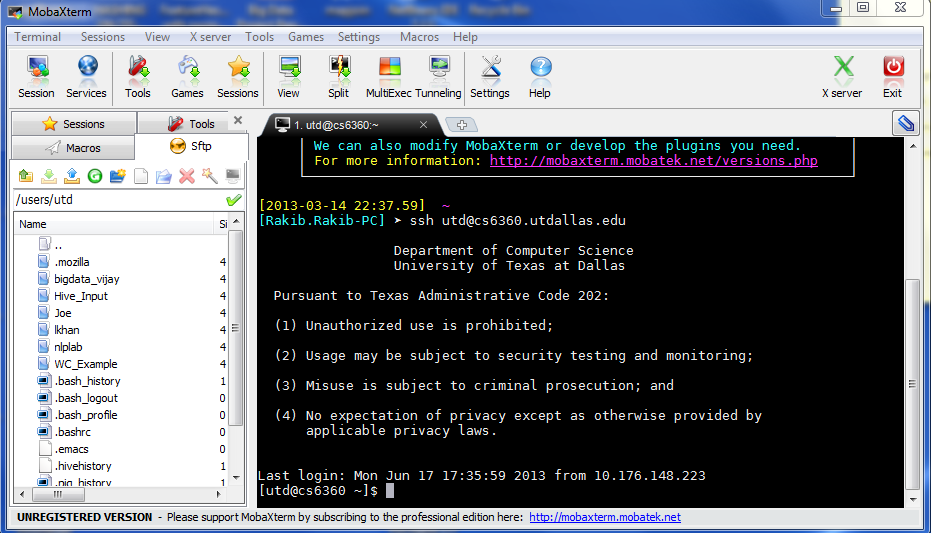
Please download an SSH client on your Windows laptops before you arrive to the workshop.  Here are the download URL's:  
  
      <http://mobaxterm.mobatek.net/MobaXterm_Setup_6.3.msi>  
      <http://www.hlt.utdallas.edu/MobaXterm_Setup_6.3.msi>

Log in **cs6360.utdallas.edu** with

Username: **utd**

Password: **hadoop.**

Run **MobaXterm** and type **utd@cs6360.utdallas.edu**

****

After log in verify your current directory by **pwd** command on ssh terminal. **pwd** shows your current directory **/users/utd**

Make sure that you have the **hadoop-core-1.0.4.jar** file in **/users/utd**. you can list all files and folder in **/users/utd** by **ls**command on terminal.

**Create a Folder by mkdir command with your name(e.g. Joe) inside /users/utd . Invoke your directory by issuing cd command on terminal .**

**In this document examples are shown using folder name 'Joe'. Please use your own name of folder in place of 'Joe' inside the commands.**

**Hands On Exercise on Mahout**

**Clustering**

**Dataset**:

All the instances are located in the HDFS path */reviews-extracted*. Please don’t copy to any other location. Data has been extracted from the imdb62 dataset \*. These movie reviews were crawled from www.imdb.com in May 2009.

Each line in imdb62 is in the following tab-separated format:

(userId) (itemId ) (title+content)

where:

- userId is the user's ID, as used in IMDb (e.g., for userId=33913 the user's review page is http://www.imdb.com/user/ur0033913/comments)

- itemId is the item's ID. For this section, IDs have been taken starting from 1 to N where N is the number of instances.

- title is the review's title

- content is the review's content

For this section, 5000 instances have been used from this dataset. Each instance has been rewritten to a separate file (document) with a slight modification to suite Mahout Requirements.

**Requirements:**

Apply k-means clustering algorithm in Mahout considering the following:

* Go to mahout bin folder

**cd /usr/local/mahout-0.7/bin/**

* Build the sequence file from text file in a given input directory.

**./mahout seqdirectory -c UTF-8 -i /reviews-extracted/ -o /Joe/kmeans**

* For building the dictionary, eliminate any token that appears in 80% of the docs. Apply TF-IDF in Mahout to generate the vectors.

**./mahout seq2sparse -i /Joe/kmeans/ -o /Joe/kmeans-vector/ -ow -chunk 100 -x 80**

* Generate 60 clusters. For the similarity measure, use the Cosine Distance Measure.

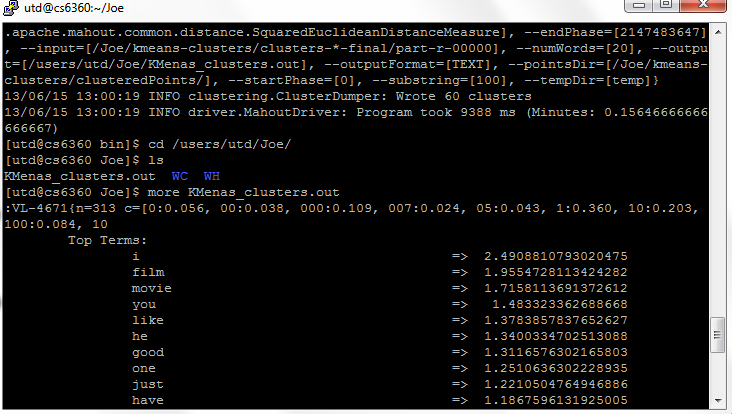
**./mahout kmeans -i /Joe/kmeans-vector/tfidf-vectors/ -c /Joe/kmeans-centroids -cl -o /Joe/kmeans-clusters -k 60 -ow -x 10 -dm org.apache.mahout.common.distance.CosineDistanceMeasure**

* Show the clusters and clustered points.

**./mahout clusterdump -d /Joe/kmeans-vector/dictionary.file-0 -dt sequencefile -i /Joe/kmeans-clusters/clusters-\*-final/part-r-00000 -n 20 -b 100 -o /users/utd/Joe/KMenas\_clusters.out -p /Joe/kmeans-clusters/clusteredPoints/**

**./mahout seqdumper -i /Joe/kmeans-clusters/clusteredPoints/part-m-00000**

You will see a file **KMenas\_clusters.out** in **/users/utd/Joe**.

****

**Classification**

**Dataset**:

The same dataset mentioned in Clustering example will be used for this question as well. The sequence files have already been generated and are located in the HDFS location /*reviews-seq*. Please don’t copy to any other location.

**Requirements**:

Apply Naïve Bayes Classification algorithm in Mahout considering the following:

* Go to mahout bin folder

**cd /usr/local/mahout-0.7/bin/**

* For building the dictionary, eliminate any token that appears in 60% of the docs. Apply TF-IDF in Mahout to generate the vectors.

**./mahout seq2sparse -i /reviews-seq -o /Joe/nb-vectors -ow -x 60**

* Use 70% for Training the model and 30% for testing.

**./mahout split -i /Joe/nb-vectors/tfidf-vectors --trainingOutput /Joe/nb-train-vectors --testOutput /Joe/nb-test-vectors --randomSelectionPct 30 --overwrite --sequenceFiles -xm sequential**

**./mahout trainnb -i /Joe/nb-train-vectors -el -li labelindex -o /Joe/nb-model -ow -c**

* Test the *Training* set to show the Correctly Classified Instances and the show Confusion Matrix.

**clear**

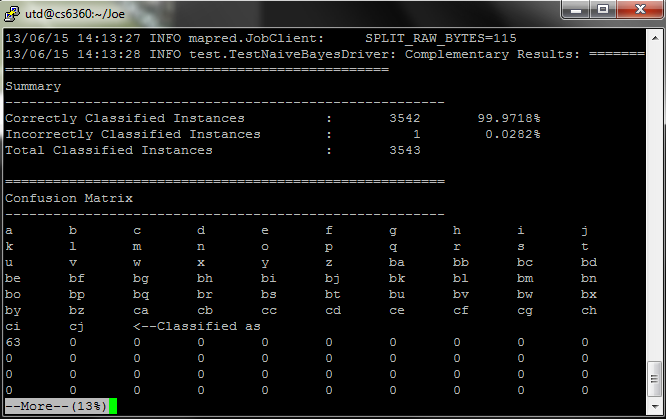
**./mahout testnb -i /Joe/nb-train-vectors -m /Joe/nb-model -l labelindex -ow -o /Joe/nb-tainingtest -c 2> /users/utd/Joe/Classifier\_TrainingModelTest.txt**

* Test the *Testing* set to show the Correctly Classified Instances and the show Confusion Matrix.

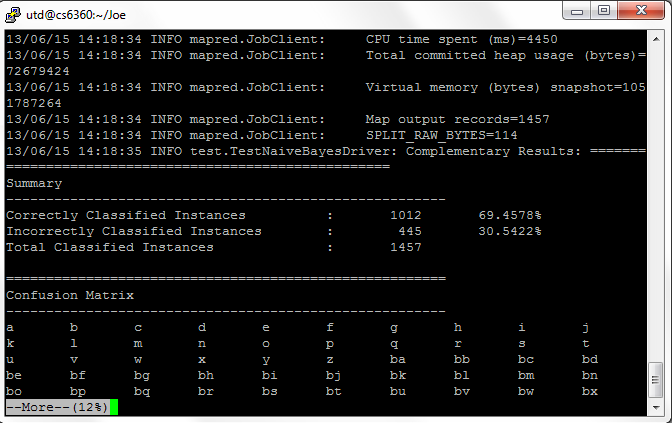
**clear**

**./mahout testnb -i /Joe/nb-test-vectors -m /Joe/nb-model -l labelindex -ow -o /Joe/nb-testingtest -c 2> /users/utd/Joe/Classifier\_TestingModelTest.txt**

Open the **/users/utd/Joe/Classifier\_TrainingModelTest.txt (**On terminal use **more /users/utd/Joe/Classifier\_TrainingModelTest.txt)**

****

Open the **/users/utd/Joe/Classifier\_TestingModelTest.txt (**On terminal use **more /users/utd/Joe/Classifier\_TestingModelTest.txt)**

****