**Week 7 – Install Mahout and Test**

**Assumptions**

1. 1.           Hadoop HDFS stand-alone system built and running
2. 2.           Remote ssh access is available to the system
3. 3.           JPS shows hadoop services are running
4. 4.           Hadoop user is logged into the Hadoop/HDFS virtual machine
5. 5.           Maven 3.3.3 is installed on your virtual machine
6. 6.           Current Spark version installed on the system for Spark supported tes

**Introduction**

**Apache Mahout** is a project of the [Apache Software Foundation](https://en.wikipedia.org/wiki/Apache_Software_Foundation) to produce [free](https://en.wikipedia.org/wiki/Free_software) implementations of [distributed](https://en.wikipedia.org/wiki/Distributed_computing) or otherwise [scalable](https://en.wikipedia.org/wiki/Scalability) [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms focused primarily in the areas of [collaborative filtering](https://en.wikipedia.org/wiki/Collaborative_filtering), clustering and classification. Many of the implementations use the [Apache Hadoop](https://en.wikipedia.org/wiki/Apache_Hadoop) platform.[1][2] Mahout also provides Java libraries for common maths operations (focused on linear algebra and statistics) and primitive Java collections. Mahout is a work in progress; the number of implemented algorithms has grown quickly,[3] but various algorithms are still missing.

While Mahout's core algorithms for [clustering](https://en.wikipedia.org/wiki/Cluster_analysis), classification and batch based collaborative filtering are implemented on top of Apache Hadoop using the [map/reduce](https://en.wikipedia.org/wiki/Map/reduce)paradigm, it does not restrict contributions to Hadoop-based implementations. Contributions that run on a single node or on a non-Hadoop cluster are also welcomed. For example, the 'Taste' collaborative-filtering recommender component of Mahout was originally a separate project and can run stand-alone without Hadoop. Integration with initiatives such as the Pregel-like Giraph are actively under discussion.[[*when?*](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Dates_and_numbers#Chronological_items)] (Wikipedia, 2015)

**Installation**

Execute the following commands -

* wget <http://apache.arvixe.com/mahout/0.11.0/apache-mahout-distribution-0.11.0.tar.gz>
* tar xvzf apache-mahout-distribution-0.11.0.tar.gz
* mv apache-mahout-distribution-0.11.0 apache-mahout-0.11.0
* nano .bashrc  (vi or vim are also acceptable alternatives to file editors on linux)
  + o   Add/update the following lines to your bash profile script
  + o   export MAHOUT\_HOME=/home/hadoop/apache-mahout-0.11.0
  + o   export MAHOUT\_LOCAL=true  # for running on a standalone hadoop machine
  + o   # unset MAHOUT\_LOCAL FOR RUNNING ON A CLUSTER
  + o   export PATH=$PATH:$MAHOUT\_HOME/bin
  + o   Save the file and exit to the command prompt
  + source .bashrc
  + cd apache-mahout-0.11.0

**Test – Basic Functionality**

* Change to the examples directory in your $MAHOUT\_HOME directory
* ./classify-wikipedia.sh
* Select 2 for the BinaryCBayes algorithm

You should see output similar to the following indicating that the mahout engine is starting

ok. You chose 2 and we'll use BinaryCBayes

creating work directory at /tmp/mahout-work-wiki

/home/hadoop/apache-mahout-0.11.0/examples/bin

+ echo 'Preparing wikipedia data'

Preparing wikipedia data

+ rm -rf /tmp/mahout-work-wiki/wiki

+ mkdir /tmp/mahout-work-wiki/wiki

+ '[' xBinaryCBayes == xCBayes ']'

+ '[' xBinaryCBayes == xBinaryCBayes ']'

+ cp /home/hadoop/apache-mahout-0.11.0/examples/bin/resources/country2.txt /tmp/mahout-work-wiki/country.txt

+ chmod 666 /tmp/mahout-work-wiki/country.txt

+ '[' /home/hadoop/hadoop '!=' '' ']'

+ '[' true == '' ']'

+ echo 'Creating sequence files from wikiXML'

Creating sequence files from wikiXML

+ /home/hadoop/apache-mahout-0.11.0/bin/mahout seqwiki -c /tmp/mahout-work-wiki/country.txt -i /tmp/mahout-work-wiki/wikixml/enwiki-latest-pages-articles.xml -o /tmp/mahout-work-wiki/wikipediainput

You will see considerable output between the final output below.  This the mapreduce job(s) running and various mahout processes.

15/10/04 01:37:38 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/output

15/10/04 01:37:39 INFO CodecPool: Got brand-new compressor [.deflate]

15/10/04 01:37:39 INFO CodecPool: Got brand-new decompressor [.deflate]

15/10/04 01:37:42 INFO TestNaiveBayesDriver: Complementary Results:

=======================================================

Summary

-------------------------------------------------------

Correctly Classified Instances          :       2088         86.3167%

Incorrectly Classified Instances        :        331          13.6833%

Total Classified Instances              :       2419

=======================================================

Confusion Matrix

-------------------------------------------------------

a                b                <--Classified as

510           14              |  524     a     = united kingdom

317           1578         |  1895  b     = united states

=======================================================

Statistics

-------------------------------------------------------

Kappa                                       0.6647

Accuracy                                   86.3167%

Reliability                                   60.2%

Reliability (standard deviation)            0.5261

Weighted precision                          0.9101

Weighted recall                             0.8632

Weighted F1 score                           0.8726

15/10/04 01:37:42 INFO MahoutDriver: Program took 4202 ms (Minutes: 0.07003333333333334)

[hadoop@localhost bin]$