

Document clustering using K -means under Mahout

1. Edit system variables (can be ignored if you already successfully installed Hadoop)

Under your home directory, edit the file `.bash_profile` (for Mac) or `.bashrc` (for Linux) and add the following lines.

```
$vi .bash_profile
export JAVA_HOME=$(/usr/libexec/java_home)
export HADOOP_HOME=/Users/DoubleJ/Software/hadoop-1.0.3 (replace with your hadoop directory)
export HADOOP_CONF_DIR=/Users/DoubleJ/Software/hadoop-1.0.3/conf
```

Then save and exit the file (press ESC and then `:wq + Enter`). To make the edition be effective, use source command.

```
$source .bash_profile
```

OR

```
$source .bashrc
```

2. Download data

Download reuters.tar.gz file and store it in the folder of mahout-work (you can create it first if not exists)

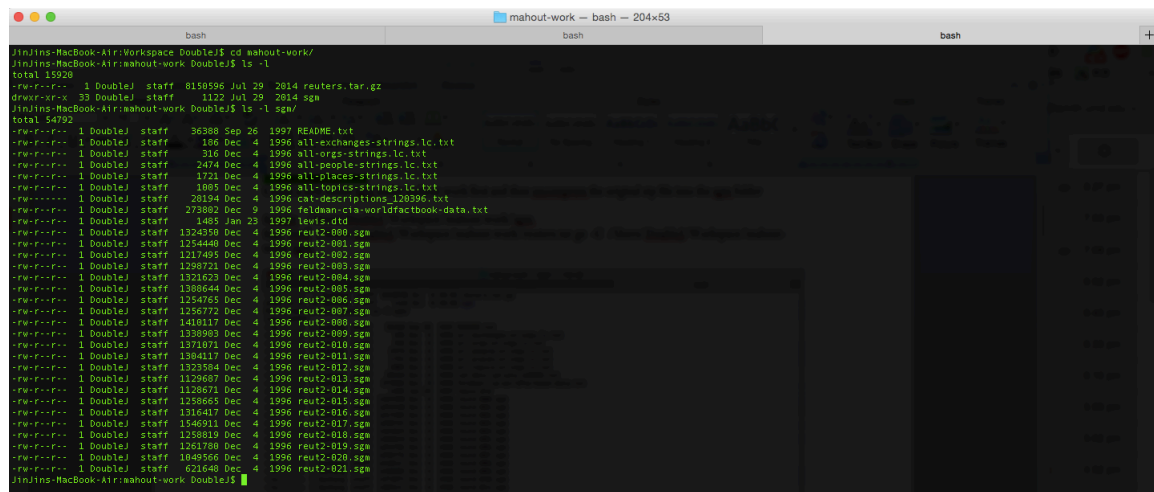
```
$curl http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.tar.gz -o
/Users/DoubleJ/Workspace/mahout-work/reuters.tar.gz
```

Note: The command line above is in one line.

3. Uncompress data

Create a new folder `sgm` under `mahout-work` first and then uncompress the original zip file into the `sgm` folder

```
$mkdir /Users/DoubleJ/Workspace/mahout-work/sgm
$tar xzvf /Users/DoubleJ/Workspace/mahout-work/reuters.tar.gz -C /Users/DoubleJ/Workspace/mahout-
work/sgm
```

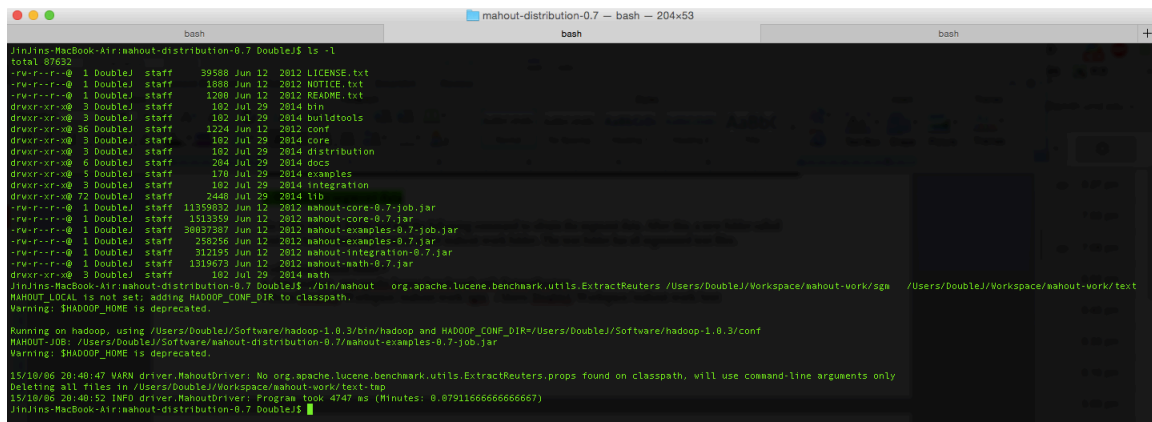


```
bash
JinJins-MacBook-Air:Workspace DoubleJ$ cd mahout-work/
JinJins-MacBook-Air:mahout-work DoubleJ$ ls -l
total 15920
-rw-r--r--  1 DoubleJ  staff   8150596 Jul 29  2014 reuters.tar.gz
drwxr-xr-x  33 DoubleJ  staff    1122 Jul 29  2014 sgm
JinJins-MacBook-Air:mahout-work DoubleJ$ ls -l sgm/
total 54792
-rw-r--r--  1 DoubleJ  staff    36388 Sep 26  1997 README.txt
-rw-r--r--  1 DoubleJ  staff     106 Dec  4  1996 all-exchanges-strings.lc.txt
-rw-r--r--  1 DoubleJ  staff     316 Dec  4  1996 all-orgs-strings.lc.txt
-rw-r--r--  1 DoubleJ  staff     2474 Dec  4  1996 all-people-strings.lc.txt
-rw-r--r--  1 DoubleJ  staff     1721 Dec  4  1996 all-places-strings.lc.txt
-rw-r--r--  1 DoubleJ  staff     1885 Dec  4  1996 all-topics-strings.lc.txt
-rw-r--r--  1 DoubleJ  staff    28194 Dec  4  1996 cat-descriptions-120396.txt
-rw-r--r--  1 DoubleJ  staff    273802 Dec  9  1996 feldman-cia-worldfactbook-data.txt
-rw-r--r--  1 DoubleJ  staff     1485 Jan 23  1997 lewis.dtd
-rw-r--r--  1 DoubleJ  staff    1324350 Dec  4  1996 reut2-000.sgm
-rw-r--r--  1 DoubleJ  staff    1254440 Dec  4  1996 reut2-001.sgm
-rw-r--r--  1 DoubleJ  staff    1217495 Dec  4  1996 reut2-002.sgm
-rw-r--r--  1 DoubleJ  staff    1298721 Dec  4  1996 reut2-003.sgm
-rw-r--r--  1 DoubleJ  staff    1321622 Dec  4  1996 reut2-004.sgm
-rw-r--r--  1 DoubleJ  staff    1388644 Dec  4  1996 reut2-005.sgm
-rw-r--r--  1 DoubleJ  staff    1254765 Dec  4  1996 reut2-006.sgm
-rw-r--r--  1 DoubleJ  staff    1256772 Dec  4  1996 reut2-007.sgm
-rw-r--r--  1 DoubleJ  staff    1410117 Dec  4  1996 reut2-008.sgm
-rw-r--r--  1 DoubleJ  staff    1338993 Dec  4  1996 reut2-009.sgm
-rw-r--r--  1 DoubleJ  staff    1371071 Dec  4  1996 reut2-010.sgm
-rw-r--r--  1 DoubleJ  staff    1384117 Dec  4  1996 reut2-011.sgm
-rw-r--r--  1 DoubleJ  staff    1323584 Dec  4  1996 reut2-012.sgm
-rw-r--r--  1 DoubleJ  staff    1129687 Dec  4  1996 reut2-013.sgm
-rw-r--r--  1 DoubleJ  staff    1236671 Dec  4  1996 reut2-014.sgm
-rw-r--r--  1 DoubleJ  staff    1258665 Dec  4  1996 reut2-015.sgm
-rw-r--r--  1 DoubleJ  staff    1316417 Dec  4  1996 reut2-016.sgm
-rw-r--r--  1 DoubleJ  staff    1346911 Dec  4  1996 reut2-017.sgm
-rw-r--r--  1 DoubleJ  staff    1258819 Dec  4  1996 reut2-018.sgm
-rw-r--r--  1 DoubleJ  staff    1261788 Dec  4  1996 reut2-019.sgm
-rw-r--r--  1 DoubleJ  staff    1048566 Dec  4  1996 reut2-020.sgm
-rw-r--r--  1 DoubleJ  staff     621648 Dec  4  1996 reut2-021.sgm
JinJins-MacBook-Air:mahout-work DoubleJ$
```

4. Obtain the segment data

Go to the mahout folder and run the following command to obtain the segment data. After this, a new folder called text will be automatically created under mahout-work folder. The text folder has all segmented text files.

```
$ cd <your mahout folder>
$ ./bin/mahout org.apache.lucene.benchmark.utils.ExtractReuters
/Users/DoubleJ/Workspace/mahout-work/sgm /Users/DoubleJ/Workspace/mahout-work/text
```



```
bash
mahout-distribution-0.7 - bash - 204x53

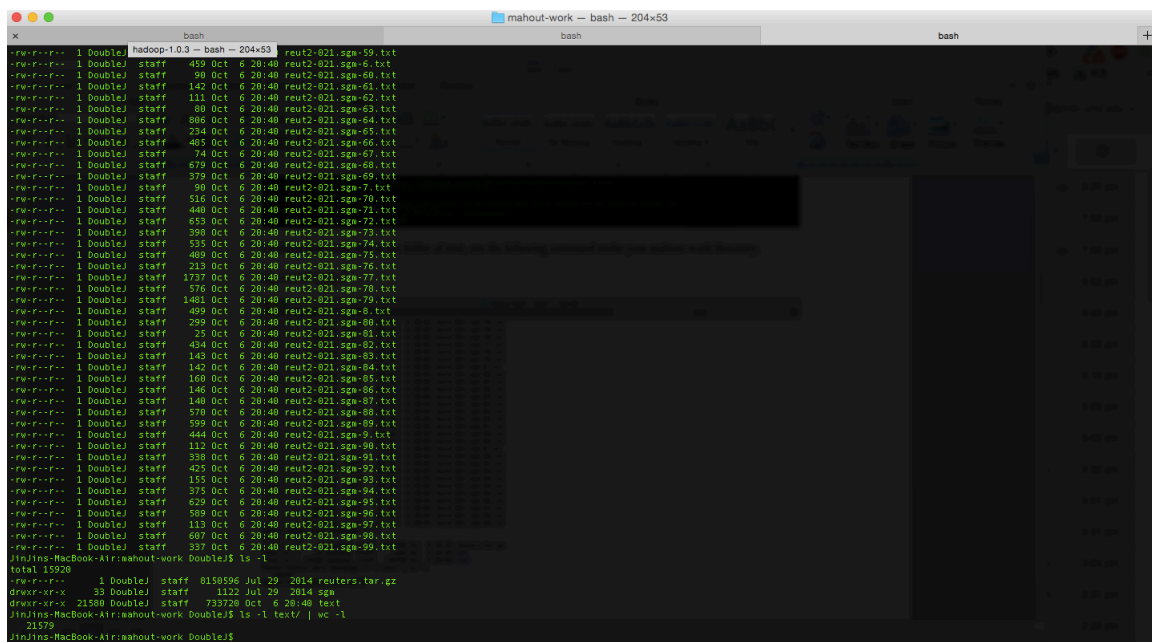
JinJins-MacBook-Air:mahout-distribution-0.7 DoubleJ$ ls -l
total 87632
-rw-r--r--@ 1 DoubleJ staff 39588 Jun 12 2012 LICENSE.txt
-rw-r--r--@ 1 DoubleJ staff 1888 Jun 12 2012 NOTICE.txt
-rw-r--r--@ 1 DoubleJ staff 1280 Jun 12 2012 README.txt
drwxr-xr-x@ 3 DoubleJ staff 102 Jul 29 2014 bin
drwxr-xr-x@ 3 DoubleJ staff 102 Jul 29 2014 buildtools
drwxr-xr-x@ 36 DoubleJ staff 1224 Jun 12 2012 conf
drwxr-xr-x@ 3 DoubleJ staff 102 Jul 29 2014 core
drwxr-xr-x@ 3 DoubleJ staff 102 Jul 29 2014 distribution
drwxr-xr-x@ 6 DoubleJ staff 204 Jul 29 2014 docs
drwxr-xr-x@ 5 DoubleJ staff 170 Jul 29 2014 examples
drwxr-xr-x@ 3 DoubleJ staff 102 Jul 29 2014 integration
drwxr-xr-x@ 72 DoubleJ staff 2448 Jul 29 2014 lib
-rw-r--r--@ 1 DoubleJ staff 1135932 Jun 12 2012 mahout-core-0.7-job.jar
-rw-r--r--@ 1 DoubleJ staff 1513359 Jun 12 2012 mahout-core-0.7.jar
-rw-r--r--@ 1 DoubleJ staff 3083287 Jun 12 2012 mahout-examples-0.7-job.jar
-rw-r--r--@ 1 DoubleJ staff 259256 Jun 12 2012 mahout-examples-0.7.jar
-rw-r--r--@ 1 DoubleJ staff 312195 Jun 12 2012 mahout-integration-0.7.jar
-rw-r--r--@ 1 DoubleJ staff 1319673 Jun 12 2012 mahout-math-0.7.jar
drwxr-xr-x@ 3 DoubleJ staff 102 Jul 29 2014 mahout
JinJins-MacBook-Air:mahout-distribution-0.7 DoubleJ$ ./bin/mahout org.apache.lucene.benchmark.utils.ExtractReuters /Users/DoubleJ/Workspace/mahout-work/sgm /Users/DoubleJ/Workspace/mahout-work/text
MAHOUT_LOCAL is not set: adding HADOOP_CONF_DIR to classpath.
Warning: $HADOOP_HOME is deprecated.

Running on hadoop, using /Users/DoubleJ/Software/hadoop-1.0.3/bin/hadoop and HADOOP_CONF_DIR=/Users/DoubleJ/Software/hadoop-1.0.3/conf
MAHOUT_JOB=/Users/DoubleJ/Software/mahout-distribution-0.7/mahout-examples-0.7-job.jar
Warning: $HADOOP_HOME is deprecated.

15/10/06 20:48:47 WARN driver.MahoutDriver: No org.apache.lucene.benchmark.utils.ExtractReuters.props found on classpath, will use command-line arguments only
Deleting all files in /Users/DoubleJ/Workspace/mahout-work/text-tmp
15/10/06 20:48:52 INFO driver.MahoutDriver: Program took 4747 ms (Minutes: 0.07911666666666667)
JinJins-MacBook-Air:mahout-distribution-0.7 DoubleJ$
```

To check how many files in the folder of text, use the following command under your mahout-work directory.

```
$ls -l text/ | wc -l
```



```
bash
mahout-work - bash - 204x53

x
JinJins-MacBook-Air:mahout-work DoubleJ$ ls -l
total 15928
-rw-r--r--@ 1 DoubleJ staff 8158596 Jul 29 2014 reuters.tar.gz
drwxr-xr-x 35 DoubleJ staff 1122 Jul 29 2014 sgm
drwxr-xr-x 21588 DoubleJ staff 733798 Oct 6 2014 text
JinJins-MacBook-Air:mahout-work DoubleJ$ ls -l text/ | wc -l
21579
JinJins-MacBook-Air:mahout-work DoubleJ$
```

5. Generate the sequence directory, which contains all texts

5.1 Send the segment data to the Hadoop file system

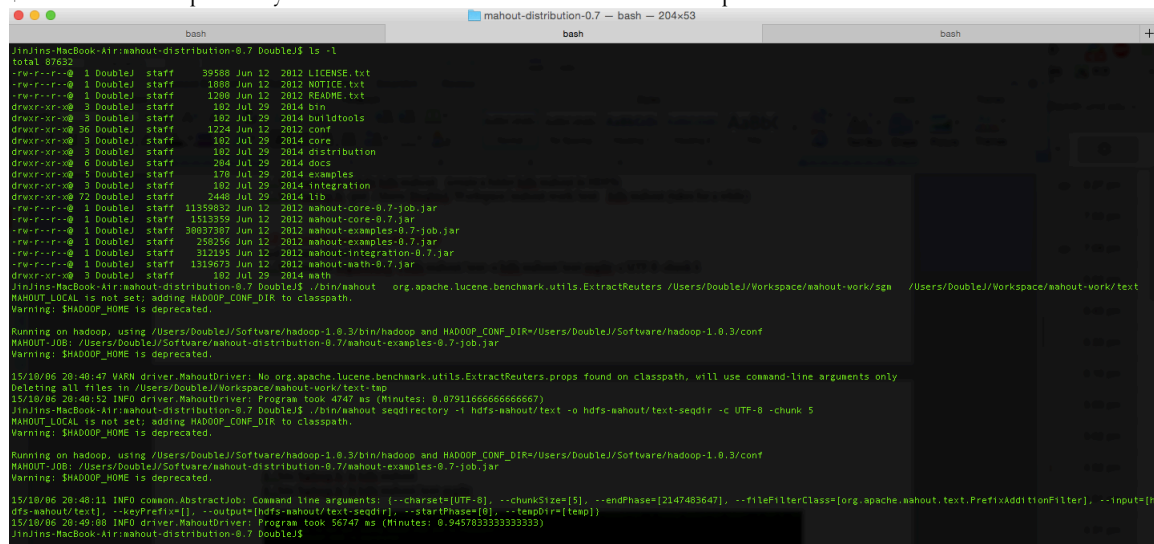
```
$cd $HADOOP_HOME
```

OR if you do not set HADOOP_HOME environment variable in the .bashrc or .bash_profile

```
$cd <your hadoop folder>
$./bin/start-all.sh
$./bin/hadoop fs -mkdir hdfs-mahout (create a folder hdfs-mahout in HDFS)
$./bin/hadoop fs -put /Users/DoubleJ/Workspace/mahout-work/text hdfs-mahout (takes for a while)
```

5.2 Create the sequence directory

```
$cd <your mahout folder>
$./bin/mahout seqdirectory -i hdfs-mahout/text -o hdfs-mahout/text-seqdir -c UTF-8 -chunk 5
```



```
bash
mahout-distribution-0.7 -- bash -- 204x53

jln@jlns-MacBook-Air:~/mahout-distribution-0.7$ ls -l
total 87632
-rw-r--r--@ 1 DoubleJ staff 39588 Jun 12 2012 LICENSE.txt
-rw-r--r--@ 1 DoubleJ staff 1888 Jun 12 2012 NOTICE.txt
-rw-r--r--@ 1 DoubleJ staff 1280 Jun 12 2012 README.txt
drwxr-xr-x@ 3 DoubleJ staff 182 Jul 29 2014 bin
drwxr-xr-x@ 3 DoubleJ staff 182 Jul 29 2014 buildtools
drwxr-xr-x@ 36 DoubleJ staff 1224 Jun 12 2012 conf
drwxr-xr-x@ 3 DoubleJ staff 182 Jul 29 2014 core
drwxr-xr-x@ 3 DoubleJ staff 182 Jul 29 2014 distribution
drwxr-xr-x@ 6 DoubleJ staff 284 Jul 29 2014 docs
drwxr-xr-x@ 5 DoubleJ staff 178 Jul 29 2014 examples
drwxr-xr-x@ 3 DoubleJ staff 182 Jul 29 2014 integration
drwxr-xr-x@ 72 DoubleJ staff 2448 Jul 29 2014 lib
-rw-r--r--@ 1 DoubleJ staff 11359832 Jun 12 2012 mahout-core-0.7-job.jar
-rw-r--r--@ 1 DoubleJ staff 1513359 Jun 12 2012 mahout-core-0.7.jar
-rw-r--r--@ 1 DoubleJ staff 38837897 Jun 12 2012 mahout-examples-0.7-job.jar
-rw-r--r--@ 1 DoubleJ staff 259256 Jun 12 2012 mahout-examples-0.7.jar
-rw-r--r--@ 1 DoubleJ staff 312195 Jun 12 2012 mahout-integration-0.7.jar
-rw-r--r--@ 1 DoubleJ staff 1319673 Jun 12 2012 mahout-math-0.7.jar
drwxr-xr-x@ 3 DoubleJ staff 182 Jul 29 2014 math
jln@jlns-MacBook-Air:~/mahout-distribution-0.7$ ./bin/mahout org.apache.lucene.benchmark.utils.ExtractReuters /Users/DoubleJ/Workspace/mahout-work/sgm /Users/DoubleJ/Workspace/mahout-work/text
MAHOUT_LOCAL is not set; adding MAHOUT_CONF_DIR to classpath.
Warning: $MAHOUT_HOME is deprecated.

Running on hadoop, using /Users/DoubleJ/Software/hadoop-1.0.3/bin/hadoop and MAHOUT_CONF_DIR=/Users/DoubleJ/Software/hadoop-1.0.3/conf
MAHOUT_JOB=/Users/DoubleJ/Software/mahout-distribution-0.7/mahout-examples-0.7-job.jar
Warning: $MAHOUT_HOME is deprecated.

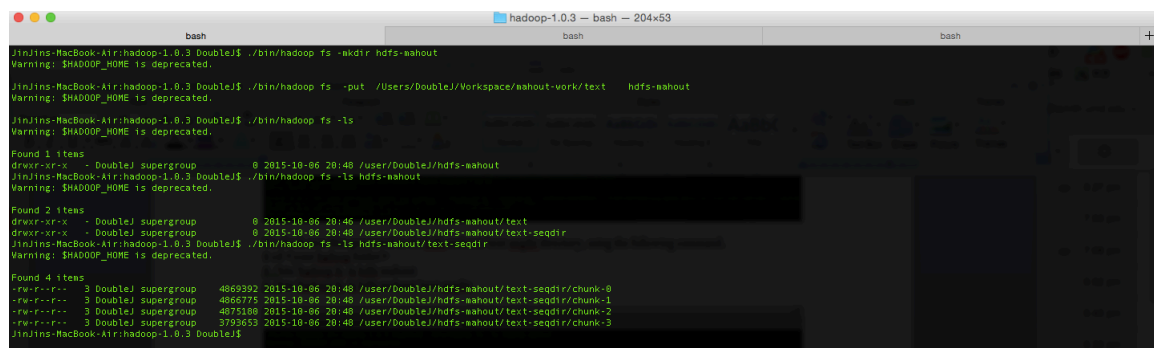
15/10/06 20:48:47 WARN driver.MahoutDriver: No org.apache.lucene.benchmark.utils.ExtractReuters.props found on classpath, will use command-line arguments only
Deleting all files in /Users/DoubleJ/Workspace/mahout-work/text/tmp
15/10/06 20:48:52 INFO driver.MahoutDriver: Program took 4747 ms (Minutes: 0.07911666666666667)
jln@jlns-MacBook-Air:~/mahout-distribution-0.7$ ./bin/mahout seqdirectory -i hdfs-mahout/text -o hdfs-mahout/text-seqdir -c UTF-8 -chunk 5
MAHOUT_LOCAL is not set; adding MAHOUT_CONF_DIR to classpath.
Warning: $MAHOUT_HOME is deprecated.

Running on hadoop, using /Users/DoubleJ/Software/hadoop-1.0.3/bin/hadoop and MAHOUT_CONF_DIR=/Users/DoubleJ/Software/hadoop-1.0.3/conf
MAHOUT_JOB=/Users/DoubleJ/Software/mahout-distribution-0.7/mahout-examples-0.7-job.jar
Warning: $MAHOUT_HOME is deprecated.

15/10/06 20:48:11 INFO common.CommandJob: Command line arguments: [--charset=UTF-8, --chunkSize=5, --endPhase=[2147483647], --fileFilterClass=[org.apache.mahout.text.PrefixAdditionFilter], --input=[hdfs-mahout/text], --keyPrefix=[], --output=[hdfs-mahout/text-seqdir], --startPhase=[0], --tempDir=[temp]]
15/10/06 20:49:08 INFO driver.MahoutDriver: Program took 56747 ms (Minutes: 0.9457833333333333)
jln@jlns-MacBook-Air:~/mahout-distribution-0.7$
```

To check new files generated in the hdfs-mahout/text-seqdir directory, using the following command.

```
$ cd <your hadoop folder>
$./bin/hadoop fs -ls hdfs-mahout
$./bin/hadoop fs -ls hdfs-mahout/text-seqdir
```



```
bash
hadoop-1.0.3 -- bash -- 204x53

jln@jlns-MacBook-Air:~/hadoop-1.0.3$ ./bin/hadoop fs -mkdir hdfs-mahout
Warning: $MAHOUT_HOME is deprecated.

jln@jlns-MacBook-Air:~/hadoop-1.0.3$ ./bin/hadoop fs -put /Users/DoubleJ/Workspace/mahout-work/text hdfs-mahout
Warning: $MAHOUT_HOME is deprecated.

jln@jlns-MacBook-Air:~/hadoop-1.0.3$ ./bin/hadoop fs -ls
Warning: $MAHOUT_HOME is deprecated.

Found 1 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout
jln@jlns-MacBook-Air:~/hadoop-1.0.3$ ./bin/hadoop fs -ls hdfs-mahout
Warning: $MAHOUT_HOME is deprecated.

Found 2 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:46 /user/DoubleJ/hdfs-mahout/text
jln@jlns-MacBook-Air:~/hadoop-1.0.3$ ./bin/hadoop fs -ls hdfs-mahout/text-seqdir
Warning: $MAHOUT_HOME is deprecated.

Found 4 items
-rw-r--r-- 3 DoubleJ supergroup 4869392 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout/text-seqdir/chunk-0
-rw-r--r-- 3 DoubleJ supergroup 4866775 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout/text-seqdir/chunk-1
-rw-r--r-- 3 DoubleJ supergroup 4875188 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout/text-seqdir/chunk-2
-rw-r--r-- 3 DoubleJ supergroup 3793653 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout/text-seqdir/chunk-3
jln@jlns-MacBook-Air:~/hadoop-1.0.3$
```

6. Create vector files to represent these documents

```
$cd <your mahout folder>
$./bin/mahout seq2sparse -i hdfs-mahout/text-seqdir/ -o hdfs-mahout/vectors --maxDFPercent 85 --namedVector
```

```
15/10/06 20:58:24 INFO mapred.JobClient: FILE_BYTES_READ=17420430
15/10/06 20:58:24 INFO mapred.JobClient: HDFS_BYTES_WRITTEN=17023387
15/10/06 20:58:24 INFO mapred.JobClient: Job Counters
15/10/06 20:58:24 INFO mapred.JobClient: Launched map tasks=1
15/10/06 20:58:24 INFO mapred.JobClient: Launched reduce tasks=1
15/10/06 20:58:24 INFO mapred.JobClient: SLOTS_MILLIS_REDUCES=10103
15/10/06 20:58:24 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
15/10/06 20:58:24 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=12970
15/10/06 20:58:24 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
15/10/06 20:58:24 INFO mapred.JobClient: Data-Local map tasks=1
15/10/06 20:58:24 INFO mapred.JobClient: File Output Format Counters
15/10/06 20:58:24 INFO mapred.JobClient: Bytes Written=17023387
15/10/06 20:58:24 INFO mapred.JobClient: input.FileInputFormat: Total input paths to process : 1
15/10/06 20:58:25 INFO mapred.JobClient: Running job: job_201510062030_0007
15/10/06 20:58:26 INFO mapred.JobClient: map 0% reduce 0%
15/10/06 20:58:43 INFO mapred.JobClient: map 100% reduce 0%
15/10/06 20:58:50 INFO mapred.JobClient: map 100% reduce 100%
15/10/06 20:59:03 INFO mapred.JobClient: Job complete: job_201510062030_0007
15/10/06 20:59:03 INFO mapred.JobClient: Counters: 26
15/10/06 20:59:03 INFO mapred.JobClient: Map-Reduce Framework
15/10/06 20:59:03 INFO mapred.JobClient: Spilled Records=43156
15/10/06 20:59:03 INFO mapred.JobClient: Map output materialized bytes=16804864
15/10/06 20:59:03 INFO mapred.JobClient: Reduce input records=21578
15/10/06 20:59:03 INFO mapred.JobClient: Map input records=21578
15/10/06 20:59:03 INFO mapred.JobClient: SPLIT_RAW_BYTES=158
15/10/06 20:59:03 INFO mapred.JobClient: Map output bytes=16722273
15/10/06 20:59:03 INFO mapred.JobClient: Reduce shuffle bytes=0
15/10/06 20:59:03 INFO mapred.JobClient: Reduce input groups=21578
15/10/06 20:59:03 INFO mapred.JobClient: Combine output records=0
15/10/06 20:59:03 INFO mapred.JobClient: Reduce output records=21578
15/10/06 20:59:03 INFO mapred.JobClient: Map output records=21578
15/10/06 20:59:03 INFO mapred.JobClient: Combine input records=0
15/10/06 20:59:03 INFO mapred.JobClient: Total committed heap usage (bytes)=204164096
15/10/06 20:59:03 INFO mapred.JobClient: File Input Format Counters
15/10/06 20:59:03 INFO mapred.JobClient: Bytes Read=17023387
15/10/06 20:59:03 INFO mapred.JobClient: FilesystemCounters
15/10/06 20:59:03 INFO mapred.JobClient: HDFS_BYTES_READ=17023387
15/10/06 20:59:03 INFO mapred.JobClient: FILE_BYTES_WRITTEN=33654549
15/10/06 20:59:03 INFO mapred.JobClient: FILE_BYTES_READ=16804864
15/10/06 20:59:03 INFO mapred.JobClient: HDFS_BYTES_WRITTEN=17023387
15/10/06 20:59:03 INFO mapred.JobClient: Job Counters
15/10/06 20:59:03 INFO mapred.JobClient: Launched map tasks=1
15/10/06 20:59:03 INFO mapred.JobClient: Launched reduce tasks=1
15/10/06 20:59:03 INFO mapred.JobClient: SLOTS_MILLIS_REDUCES=13542
15/10/06 20:59:03 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
15/10/06 20:59:03 INFO mapred.JobClient: SLOTS_MILLIS_MAPS=13792
15/10/06 20:59:03 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
15/10/06 20:59:03 INFO mapred.JobClient: Data-Local map tasks=1
15/10/06 20:59:03 INFO mapred.JobClient: File Output Format Counters
15/10/06 20:59:03 INFO mapred.JobClient: Bytes Written=17023387
15/10/06 20:59:03 INFO mapred.JobClient: Deleting hdfs-mahout/vectors/partial-vectors-0
15/10/06 20:59:03 INFO driver.MahoutDriver: Program took 277069 ms (Minutes: 4.6178166666666666)
jin@jin-MacBook-Air:~/mahout-distribution-0.7.0$ double$
```

To check new files generated in the HDFS, use the following commands.

```
$cd <your hadoop folder>
$ ./bin/hadoop fs -ls hdfs-mahout
$ ./bin/hadoop fs -ls hdfs-mahout/vectors
```

```
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$
Warning: $HADOOP_HOME is deprecated.
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$ ./bin/hadoop fs -mkdir hdfs-mahout
Warning: $HADOOP_HOME is deprecated.
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$ ./bin/hadoop fs -put /Users/Double/Workspace/mahout-work/text hdfs-mahout
Warning: $HADOOP_HOME is deprecated.
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$ ./bin/hadoop fs -ls
Warning: $HADOOP_HOME is deprecated.
Found 1 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:40 /user/Double/hdfs-mahout
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$ ./bin/hadoop fs -ls hdfs-mahout
Warning: $HADOOP_HOME is deprecated.
Found 2 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:46 /user/Double/hdfs-mahout/text
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:48 /user/Double/hdfs-mahout/text-seqdir
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$ ./bin/hadoop fs -ls hdfs-mahout/text-seqdir
Warning: $HADOOP_HOME is deprecated.
Found 4 items
-rw-r--r-- 3 DoubleJ supergroup 4069392 2015-10-06 20:40 /user/Double/hdfs-mahout/text-seqdir/chunk-0
-rw-r--r-- 3 DoubleJ supergroup 4066775 2015-10-06 20:40 /user/Double/hdfs-mahout/text-seqdir/chunk-1
-rw-r--r-- 3 DoubleJ supergroup 4075189 2015-10-06 20:40 /user/Double/hdfs-mahout/text-seqdir/chunk-2
-rw-r--r-- 3 DoubleJ supergroup 3793653 2015-10-06 20:40 /user/Double/hdfs-mahout/text-seqdir/chunk-3
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$ ./bin/hadoop fs -ls hdfs-mahout
Warning: $HADOOP_HOME is deprecated.
Found 3 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:46 /user/Double/hdfs-mahout/text
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:48 /user/Double/hdfs-mahout/text-seqdir
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:59 /user/Double/hdfs-mahout/vectors
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$ ./bin/hadoop fs -ls hdfs-mahout/vectors
Warning: $HADOOP_HOME is deprecated.
Found 7 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:57 /user/Double/hdfs-mahout/vectors/df-count
-rw-r--r-- 3 DoubleJ supergroup 824086 2015-10-06 20:55 /user/Double/hdfs-mahout/vectors/dictionary.file-0
-rw-r--r-- 3 DoubleJ supergroup 844593 2015-10-06 20:57 /user/Double/hdfs-mahout/vectors/frequency.file-0
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:57 /user/Double/hdfs-mahout/vectors/tf-vectors
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:58 /user/Double/hdfs-mahout/vectors/tfidf-vectors
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:54 /user/Double/hdfs-mahout/vectors/wordcount-documents
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:55 /user/Double/hdfs-mahout/vectors/wordcount
jin@jin-MacBook-Air:~/hadoop-1.0.3$ double$
```

In addition, when you check the completed jobs using <http://localhost:50030>, you will see the following screen.

Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)

Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

none

Completed Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201510062030_0004	NORMAL	DoubleJ	PartialVectorMerger::MergePartialVectors	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0005	NORMAL	DoubleJ	VectorTfidf Document Frequency Count running over input: hdfs-mahout/vectors/tf-vectors	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0006	NORMAL	DoubleJ	: MakePartialVectors: input-folder: hdfs-mahout/vectors/tf-vectors, dictionary-file: hdfs-mahout/vectors/frequency.file-0	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0007	NORMAL	DoubleJ	PartialVectorMerger::MergePartialVectors	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0001	NORMAL	DoubleJ	DocumentProcessor::DocumentTokenizer: input-folder: hdfs-mahout/text-seqdir	100.00%	4	4	100.00%	0	0	NA	NA
job_201510062030_0002	NORMAL	DoubleJ	DictionaryVectorizer::WordCount: input-folder: hdfs-mahout/vectors/tokenized-documents	100.00%	4	4	100.00%	1	1	NA	NA
job_201510062030_0003	NORMAL	DoubleJ	DictionaryVectorizer::MakePartialVectors: input-folder: hdfs-mahout/vectors/tokenized-documents, dictionary-file: hdfs-mahout/vectors/dictionary.file-0	100.00%	4	4	100.00%	1	1	NA	NA

7. Running K-means

\$cd <your mahout folder>

\$./bin/mahout kmeans -i hdfs-mahout/vectors/tfidf-vectors -c hdfs-mahout/cluster-centroids -o hdfs-mahout/kmeans -dm org.apache.mahout.common.distance.CosineDistanceMeasure -x 10 -k 20 -ow --clustering -cl

```

mahout-distribution-0.7 -- bash -- 204x53
bash
15/10/06 21:11:34 INFO wapped.JobClient: Combine input records=0
15/10/06 21:11:34 INFO wapped.JobClient: Total committed heap usage (bytes)=306184192
15/10/06 21:11:34 INFO wapped.JobClient: File Input Format Counters
15/10/06 21:11:34 INFO wapped.JobClient: Bytes Read=17023387
15/10/06 21:11:34 INFO wapped.JobClient: FileSystemCounters
15/10/06 21:11:34 INFO wapped.JobClient: HDFS_BYTES_READ=24992083
15/10/06 21:11:34 INFO wapped.JobClient: FILE_BYTES_WRITTEN=1649121
15/10/06 21:11:34 INFO wapped.JobClient: FILE_BYTES_READ=7801658
15/10/06 21:11:34 INFO wapped.JobClient: HDFS_BYTES_WRITTEN=3021459
15/10/06 21:11:34 INFO wapped.JobClient: Job Counters
15/10/06 21:11:34 INFO wapped.JobClient: Launched map tasks=1
15/10/06 21:11:34 INFO wapped.JobClient: Launched reduce tasks=1
15/10/06 21:11:34 INFO wapped.JobClient: SLOTS_MILLIS_REDUCES=16207
15/10/06 21:11:34 INFO wapped.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
15/10/06 21:11:34 INFO wapped.JobClient: SLOTS_MILLIS_MAPS=16887
15/10/06 21:11:34 INFO wapped.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
15/10/06 21:11:34 INFO wapped.JobClient: Data-local map tasks=1
15/10/06 21:11:34 INFO wapped.JobClient: File Output Format Counters
15/10/06 21:11:34 INFO wapped.JobClient: Bytes Written=3821459
15/10/06 21:11:34 INFO kmeans.KMeansDriver: Clustering data
15/10/06 21:11:34 INFO kmeans.KMeansDriver: Running Clustering
15/10/06 21:11:34 INFO kmeans.KMeansDriver: Input: hdfs-mahout/vectors/tfidf-vectors Clusters In: hdfs-mahout/kmeans Out: hdfs-mahout/kmeans Distance: org.apache.mahout.common.distance.CosineDistanceMeasure
15/10/06 21:11:35 INFO wapped.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
15/10/06 21:11:35 INFO input.FileInputFormat: Total input paths to process : 1
15/10/06 21:11:35 INFO wapped.JobClient: Running job: job_201510062030_0010
15/10/06 21:11:36 INFO wapped.JobClient: map 0% reduce 0%
15/10/06 21:11:56 INFO wapped.JobClient: map 100% reduce 0%
15/10/06 21:12:01 INFO wapped.JobClient: Job complete: job_201510062030_0010
15/10/06 21:12:01 INFO wapped.JobClient: Counters: 16
15/10/06 21:12:01 INFO wapped.JobClient: Map-Reduce Framework
15/10/06 21:12:01 INFO wapped.JobClient: Spilled Records=0
15/10/06 21:12:01 INFO wapped.JobClient: Map input records=21578
15/10/06 21:12:01 INFO wapped.JobClient: SPIT_RAW_BYTES=146
15/10/06 21:12:01 INFO wapped.JobClient: Map output records=21578
15/10/06 21:12:01 INFO wapped.JobClient: Total committed heap usage (bytes)=111673344
15/10/06 21:12:01 INFO wapped.JobClient: File Input Format Counters
15/10/06 21:12:01 INFO wapped.JobClient: Bytes Read=17023387
15/10/06 21:12:01 INFO wapped.JobClient: FileSystemCounters
15/10/06 21:12:01 INFO wapped.JobClient: HDFS_BYTES_READ=20045106
15/10/06 21:12:01 INFO wapped.JobClient: FILE_BYTES_WRITTEN=21999
15/10/06 21:12:01 INFO wapped.JobClient: HDFS_BYTES_READ=16766819
15/10/06 21:12:01 INFO wapped.JobClient: Job Counters
15/10/06 21:12:01 INFO wapped.JobClient: Launched map tasks=1
15/10/06 21:12:01 INFO wapped.JobClient: SLOTS_MILLIS_REDUCES=0
15/10/06 21:12:01 INFO wapped.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
15/10/06 21:12:01 INFO wapped.JobClient: SLOTS_MILLIS_MAPS=16822
15/10/06 21:12:01 INFO wapped.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
15/10/06 21:12:01 INFO wapped.JobClient: Data-local map tasks=1
15/10/06 21:12:01 INFO wapped.JobClient: File Output Format Counters
15/10/06 21:12:01 INFO wapped.JobClient: Bytes Written=16766819
15/10/06 21:12:01 INFO driver.MahoutDriver: Program took 113687 ms (Minutes: 1.8947833333333333)
$./bin/hadoop fs -ls hdfs-mahout/
$./bin/hadoop fs -ls hdfs-mahout/cluster-centroids

```

To check new files generated in HDFS, use the following commands.

\$cd <your hadoop folder>

\$./bin/hadoop fs -ls hdfs-mahout/

\$./bin/hadoop fs -ls hdfs-mahout/cluster-centroids

```
hadoop-1.0.3 — bash — 204x53
bash
jindins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs -ls
Warning: $HADOOP_HOME is deprecated.

Found 1 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout
jindins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs -ls hdfs-mahout
Warning: $HADOOP_HOME is deprecated.

Found 2 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:46 /user/DoubleJ/hdfs-mahout/text
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout/text-seqdir
jindins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs -ls hdfs-mahout/text-seqdir
Warning: $HADOOP_HOME is deprecated.

Found 4 items
-rw-r--r-- 3 DoubleJ supergroup 4869392 2015-10-06 20:40 /user/DoubleJ/hdfs-mahout/text-seqdir/chunk-0
-rw-r--r-- 3 DoubleJ supergroup 4866775 2015-10-06 20:40 /user/DoubleJ/hdfs-mahout/text-seqdir/chunk-1
-rw-r--r-- 3 DoubleJ supergroup 4875180 2015-10-06 20:40 /user/DoubleJ/hdfs-mahout/text-seqdir/chunk-2
-rw-r--r-- 3 DoubleJ supergroup 2723653 2015-10-06 20:40 /user/DoubleJ/hdfs-mahout/text-seqdir/chunk-3
jindins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs -ls hdfs-mahout
Warning: $HADOOP_HOME is deprecated.

Found 3 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:46 /user/DoubleJ/hdfs-mahout/text
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout/text-seqdir
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:49 /user/DoubleJ/hdfs-mahout/vectors
jindins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs -ls hdfs-mahout/vectors
Warning: $HADOOP_HOME is deprecated.

Found 7 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:57 /user/DoubleJ/hdfs-mahout/vectors/df-count
-rw-r--r-- 3 DoubleJ supergroup 824886 2015-10-06 20:55 /user/DoubleJ/hdfs-mahout/vectors/dictionary.file-0
-rw-r--r-- 3 DoubleJ supergroup 844593 2015-10-06 20:57 /user/DoubleJ/hdfs-mahout/vectors/frequency.file-0
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:57 /user/DoubleJ/hdfs-mahout/vectors/tf-vectors
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:58 /user/DoubleJ/hdfs-mahout/vectors/tfidf-vectors
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:54 /user/DoubleJ/hdfs-mahout/vectors/tokenized-documents
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:55 /user/DoubleJ/hdfs-mahout/vectors/wordcount
jindins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs -ls hdfs-mahout
Warning: $HADOOP_HOME is deprecated.

Found 5 items
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 21:10 /user/DoubleJ/hdfs-mahout/cluster-centroids
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 21:11 /user/DoubleJ/hdfs-mahout/kmeans
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:46 /user/DoubleJ/hdfs-mahout/text
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:48 /user/DoubleJ/hdfs-mahout/text-seqdir
drwxr-xr-x - DoubleJ supergroup 0 2015-10-06 20:49 /user/DoubleJ/hdfs-mahout/vectors
jindins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs -ls hdfs-mahout/cluster-centroids
Warning: $HADOOP_HOME is deprecated.

Found 1 items
-rw-r--r-- 3 DoubleJ supergroup 17969 2015-10-06 21:10 /user/DoubleJ/hdfs-mahout/cluster-centroids/part-randomSeed
jindins-MacBook-Air:hadoop-1.0.3 DoubleJ$
```

In addition, when you check the completed jobs using <http://localhost:50030>, you will see the following screen.

Quick Links

Running Jobs

none

Completed Jobs

JobId	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201510062030_0001	NORMAL	DoubleJ	DocumentProcessor::DocumentTokenizer: input-folder: hdfs-mahout/text-seqdir	100.00%	4	4	100.00%	0	0	NA	NA
job_201510062030_0002	NORMAL	DoubleJ	DictionaryVectorizer::WordCount: input-folder: hdfs-mahout/vectors/tokenized-documents	100.00%	4	4	100.00%	1	1	NA	NA
job_201510062030_0003	NORMAL	DoubleJ	DictionaryVectorizer::MakePartialVectors: input-folder: hdfs-mahout/vectors/tokenized-documents, dictionary-file: hdfs-mahout/vectors/dictionary.file-0	100.00%	4	4	100.00%	1	1	NA	NA
job_201510062030_0004	NORMAL	DoubleJ	PartialVectorMerger::MergePartialVectors	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0005	NORMAL	DoubleJ	VectorTfidf Document Frequency Count running over input: hdfs-mahout/vectors/tf-vectors	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0006	NORMAL	DoubleJ	: MakePartialVectors: input-folder: hdfs-mahout/vectors/tf-vectors, dictionary-file: hdfs-mahout/vectors/frequency.file-0	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0007	NORMAL	DoubleJ	PartialVectorMerger::MergePartialVectors	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0008	NORMAL	DoubleJ	Cluster iterator running iteration 1 over priorPath: hdfs-mahout/kmeans/clusters-0	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0009	NORMAL	DoubleJ	Cluster iterator running iteration 2 over priorPath: hdfs-mahout/kmeans/clusters-1	100.00%	1	1	100.00%	1	1	NA	NA
job_201510062030_0010	NORMAL	DoubleJ	Cluster Classification Driver running over input: hdfs-mahout/vectors/tfidf-vectors	100.00%	1	1	100.00%	0	0	NA	NA

Retired Jobs

8. View results in a human readable way

8.1 Dump the documents into a cluster mapping

```
$cd <your mahout folder>
$./bin/mahout seqdumper -i hdfs-mahout/kmeans/clusteredPoints/part-m-00000 >
/Users/DoubleJ/Workspace/mahout-work/cluster-docs.txt
```

```
15/10/06 21:11:34 INFO wmapred.JobClient: FILE_BYTES_READ=7801650
15/10/06 21:11:34 INFO wmapred.JobClient: HDFS_BYTES_WRITTEN=3021459
15/10/06 21:11:34 INFO wmapred.JobClient: Job Counters
15/10/06 21:11:34 INFO wmapred.JobClient: Launched map tasks=1
15/10/06 21:11:34 INFO wmapred.JobClient: Launched reduce tasks=1
15/10/06 21:11:34 INFO wmapred.JobClient: SLOTS_MILLIS_REDUCES=10207
15/10/06 21:11:34 INFO wmapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
15/10/06 21:11:34 INFO wmapred.JobClient: SLOTS_MILLIS_MAPS=16007
15/10/06 21:11:34 INFO wmapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
15/10/06 21:11:34 INFO wmapred.JobClient: Data-local map tasks=1
15/10/06 21:11:34 INFO wmapred.JobClient: File Output Format Counters
15/10/06 21:11:34 INFO wmapred.JobClient: Bytes Written=3021459
15/10/06 21:11:34 INFO kmeans.KMeansDriver: Clustering data
15/10/06 21:11:34 INFO kmeans.KMeansDriver: Running Clustering
15/10/06 21:11:34 INFO kmeans.KMeansDriver: Input: hdfs-mahout/vectors/tfidf-vectors Clusters In: hdfs-mahout/kmeans Out: hdfs-mahout/kmeans Distance: org.apache.mahout.common.distance.CosineDistanceMeasu
red7000016
15/10/06 21:11:34 WARN wmapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
15/10/06 21:11:35 INFO input.FileInputFormat: Total input paths to process : 1
15/10/06 21:11:35 INFO wmapred.JobClient: Running job: job_201510062030_0010
15/10/06 21:11:36 INFO wmapred.JobClient: map 0% reduce 0%
15/10/06 21:11:56 INFO wmapred.JobClient: map 100% reduce 0%
15/10/06 21:12:01 INFO wmapred.JobClient: Job complete: job_201510062030_0010
15/10/06 21:12:01 INFO wmapred.JobClient: Counters: 16
15/10/06 21:12:01 INFO wmapred.JobClient: Map-Reduce Framework
15/10/06 21:12:01 INFO wmapred.JobClient: Spilled Records=0
15/10/06 21:12:01 INFO wmapred.JobClient: Map Input Records=21578
15/10/06 21:12:01 INFO wmapred.JobClient: SPLIT_RAW_BYTES=146
15/10/06 21:12:01 INFO wmapred.JobClient: Map output records=21578
15/10/06 21:12:01 INFO wmapred.JobClient: Total committed heap usage (bytes)=111673344
15/10/06 21:12:01 INFO wmapred.JobClient: File Input Format Counters
15/10/06 21:12:01 INFO wmapred.JobClient: Bytes Read=17023307
15/10/06 21:12:01 INFO wmapred.JobClient: FileSystemCounters
15/10/06 21:12:01 INFO wmapred.JobClient: HDFS_BYTES_READ=20845106
15/10/06 21:12:01 INFO wmapred.JobClient: FILE_BYTES_WRITTEN=21999
15/10/06 21:12:01 INFO wmapred.JobClient: HDFS_BYTES_WRITTEN=16786819
15/10/06 21:12:01 INFO wmapred.JobClient: Job Counters
15/10/06 21:12:01 INFO wmapred.JobClient: Launched map tasks=1
15/10/06 21:12:01 INFO wmapred.JobClient: SLOTS_MILLIS_REDUCES=0
15/10/06 21:12:01 INFO wmapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
15/10/06 21:12:01 INFO wmapred.JobClient: SLOTS_MILLIS_MAPS=16022
15/10/06 21:12:01 INFO wmapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
15/10/06 21:12:01 INFO wmapred.JobClient: Data-local map tasks=1
15/10/06 21:12:01 INFO wmapred.JobClient: File Output Format Counters
15/10/06 21:12:01 INFO wmapred.JobClient: Bytes Written=16786819
15/10/06 21:12:01 INFO driver.MahoutDriver: Program took 113687 ms (Minutes: 1.8947833333333333)
jln@jlns-MacBook-Air:mahout-distribution-0.7 Double$ ./bin/mahout seqdumper -i hdfs-mahout/kmeans/clusteredPoints/part-n-000000 -o /Users/DoubleJ/Workspace/mahout-work/cluster-docs.txt
Warning: $HADOO_HOME is deprecated.
Warning: $HADOO_HOME is deprecated.
15/10/06 21:16:52 INFO common.AbstractJob: Command line arguments: [--endPhase=[2147483647], --input=[hdfs-mahout/kmeans/clusteredPoints/part-n-000000], --startPhase=[0], --tempDir=[temp]]
15/10/06 21:17:28 INFO driver.MahoutDriver: Program took 27762 ms (Minutes: 0.4627)
jln@jlns-MacBook-Air:mahout-distribution-0.7 Double$
```

A new file (cluster-docs.txt) will be generated in your local system: /Users/DoubleJ/Workspace/cluster-docs.txt

```
$ls -l /Users/DoubleJ/Workspace/mahout-work/
```

8.2 Download the perl view.pl and run it using the following command.

```
$perl view.pl /Users/DoubleJ/Workspace/mahout-work/cluster-docs.txt /Users/DoubleJ/Workspace/mahout-work/cluster-results.txt
```

The cluster-results.txt is a text file where each line has two columns with a tab separated: clustered and its corresponding file name.