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)

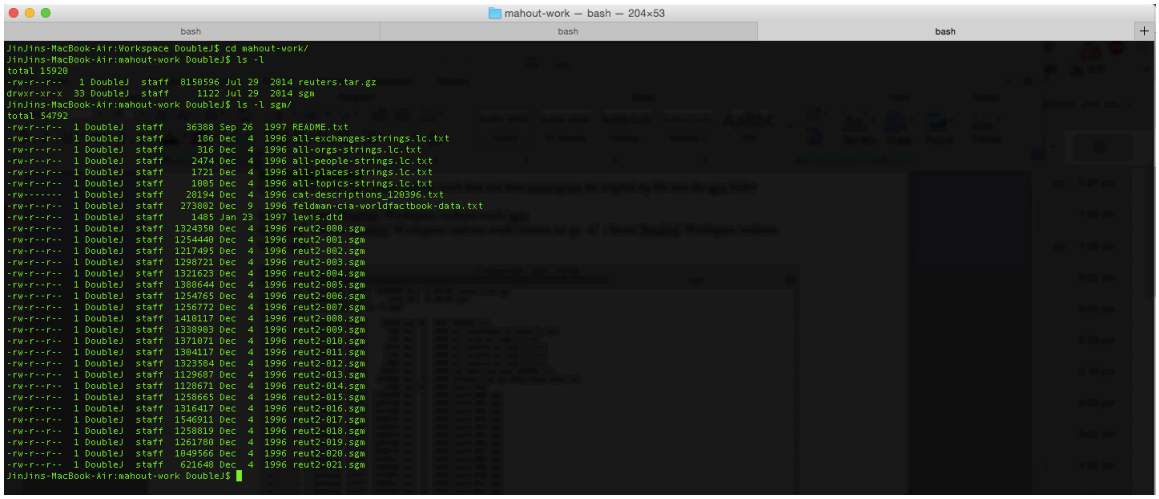
```
$curl http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.tar.gz  -o
/Users/DoubleJ/Workspace/mahout-work/reuters.tar.gz
```

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

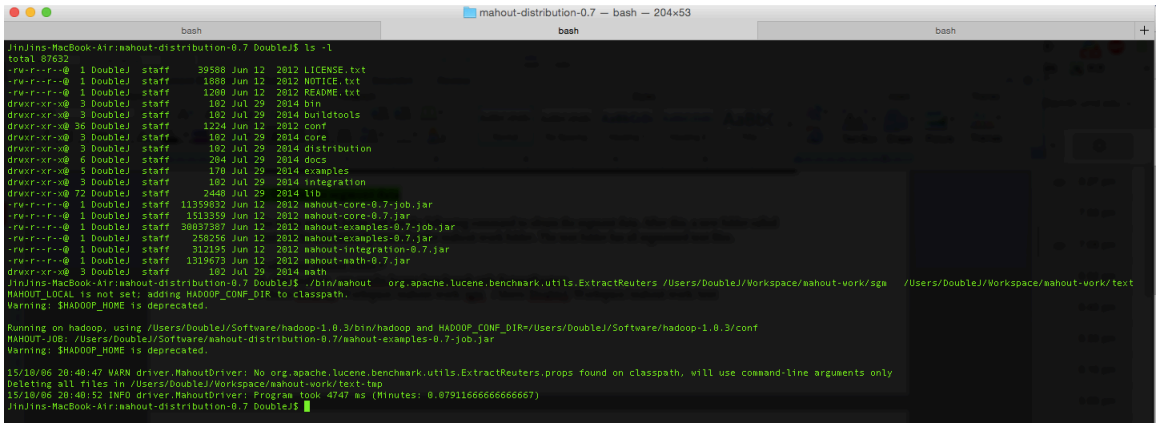
```
$mkdir    /Users/DoubleJ/Workspace/mahout-work/sgm
$tar   zxvf   /Users/DoubleJ/Workspace/mahout-work/reuters.tar.gz  -C  /Users/DoubleJ/Workspace/mahout-work/sgm
```



4. Obtain the segment data

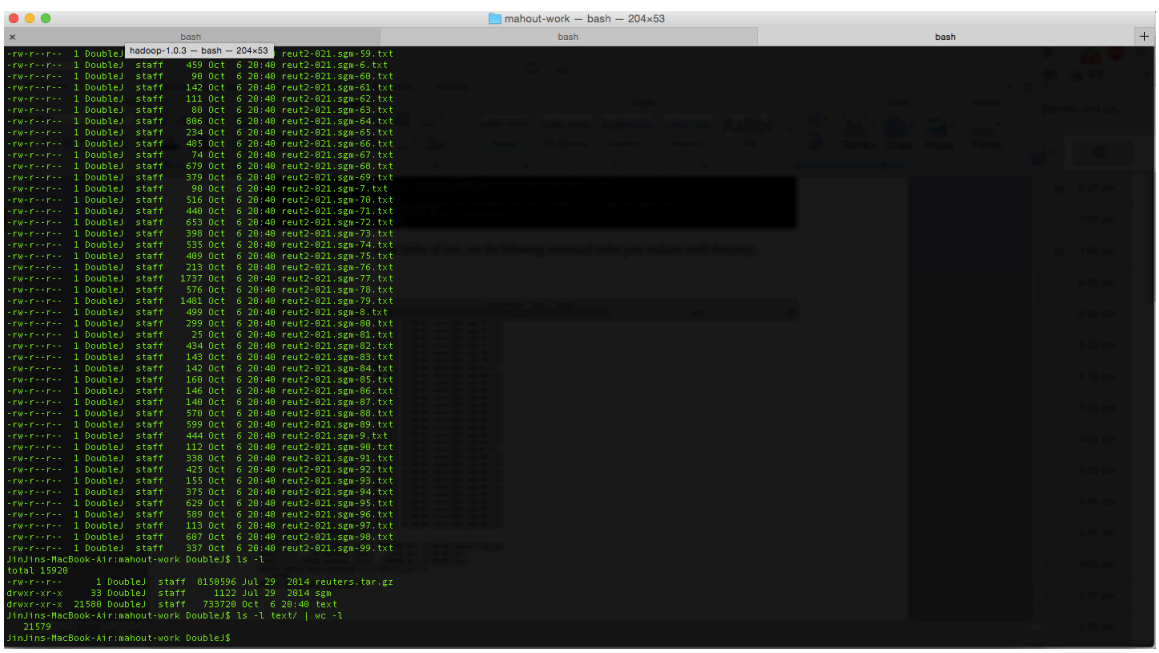
Go to the mahout folder and run the following command to obtain the segment data. After this, a new folder called text will be automatically created under mahout-work folder. The text folder has all segmented text files.

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



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

$ls -l text/ | wc -l



5. Generate the sequence directory, which contains all texts

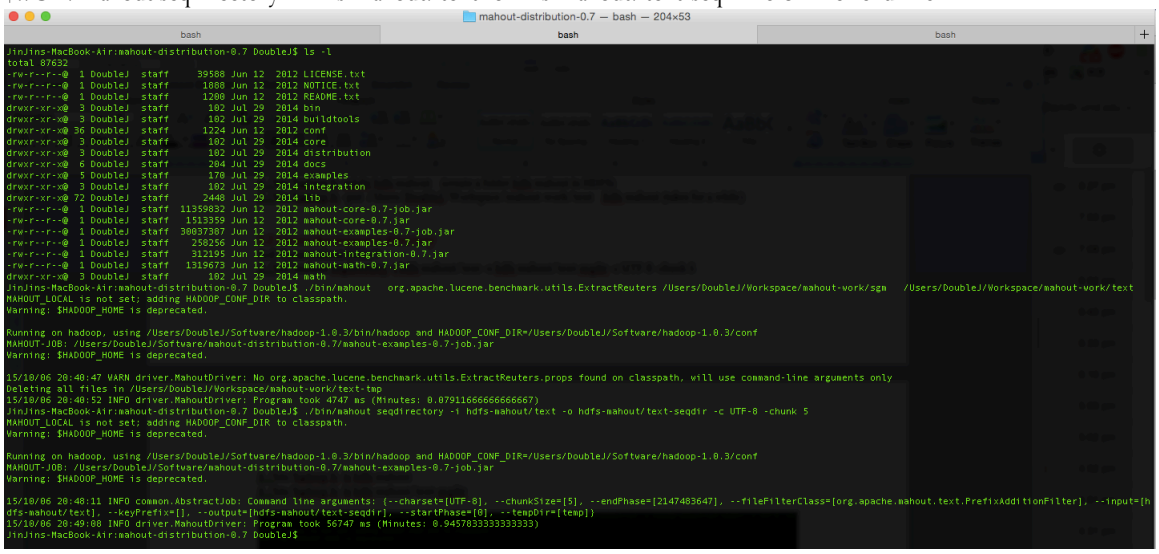
5.1 Send the segment data to the Hadoop file system

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

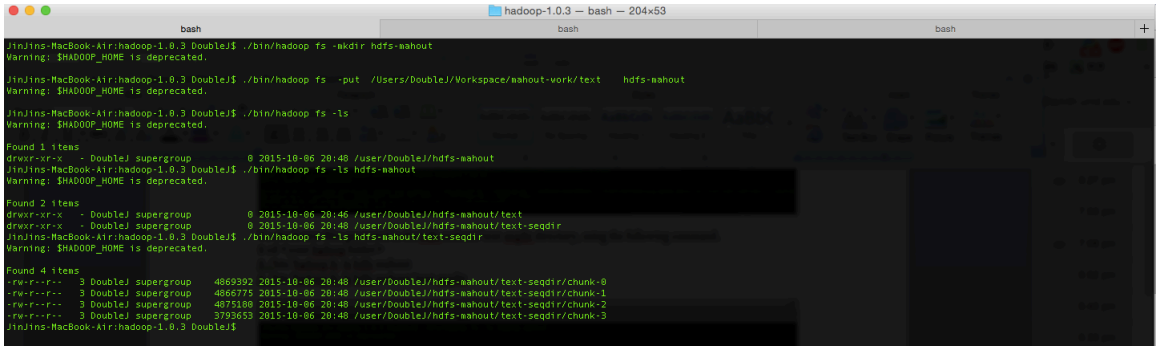
$cd <your hadoop folder>
$./bin/start-all.sh
$./bin/hadoop fs -mkdir hdfs-mahout (create a folder hdfs-mahout in HDFS)
$./bin/hadoop fs -put /Users/DoubleJ/Workspace/mahout-work/text hdfs-mahout (takes for a while)

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



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



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

```
mahout-distribution-0.7 — bash — 204×53
bash    bash    bash
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=10193
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 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 20:58:58 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=150
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)=284164096
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=17023537
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 common.HadoopUtil: Deleting hdfs-mahout/vectors/partial-vectors-0
15/10/06 20:59:03 INFO driver.MahoutDriver: Program took 277069 ms (Minutes: 4.617816666666666)
JinJins-MacBook-Air:mahout-distribution-0.7 DoubleJ$
```

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

```
hadoop-1.0.3 — bash — 204×53
bash    bash    bash
JinJins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs -mkdir hdfs-mahout
Warning: $HADOOP_HOME is deprecated.

JinJins-MacBook-Air:hadoop-1.0.3 DoubleJ$ ./bin/hadoop fs  -put  /Users/DoubleJ/Workspace/mahout-work/text    hdfs-mahout
Warning: $HADOOP_HOME is deprecated.

JinJins-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
JinJins-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
JinJins-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 20:48 /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    4875180 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
JinJins-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 20:59 /user/DoubleJ/hdfs-mahout/vectors
JinJins-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     824086 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
JinJins-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

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

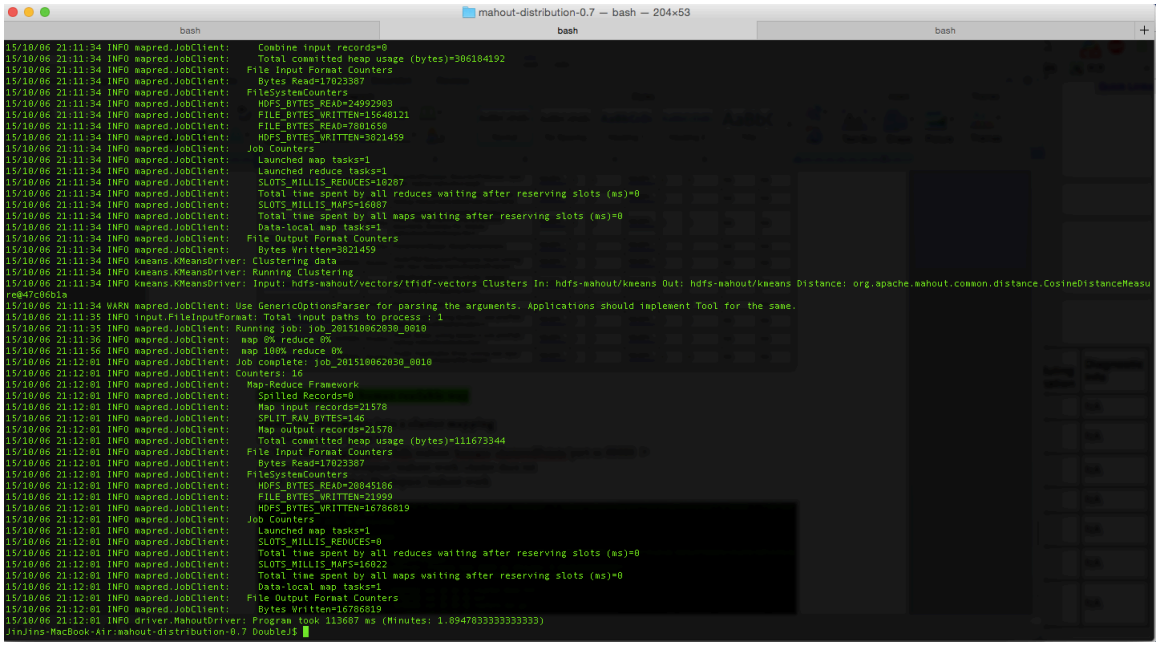
none

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



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 — 204×53
bash    bash    bash

JinJins-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
JinJins-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
JinJins-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 20:48 /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    4875180 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
JinJins-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 20:59 /user/DoubleJ/hdfs-mahout/vectors
JinJins-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     824086 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
JinJins-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 20:59 /user/DoubleJ/hdfs-mahout/vectors
JinJins-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
JinJins-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

```
mahout-distribution-0.7 — bash — 204×53
bash    bash    bash
15/10/06 21:11:34 INFO mapred.JobClient:     FILE_BYTES_READ=7801650
15/10/06 21:11:34 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=3821459
15/10/06 21:11:34 INFO mapred.JobClient:   Job Counters
15/10/06 21:11:34 INFO mapred.JobClient:     Launched map tasks=1
15/10/06 21:11:34 INFO mapred.JobClient:     Launched reduce tasks=1
15/10/06 21:11:34 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=10287
15/10/06 21:11:34 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
15/10/06 21:11:34 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=16007
15/10/06 21:11:34 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
15/10/06 21:11:34 INFO mapred.JobClient:     Data-local map tasks=1
15/10/06 21:11:34 INFO mapred.JobClient:   File Output Format Counters
15/10/06 21:11:34 INFO mapred.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.CosineDistanceMeasu
re@47c06b1a
15/10/06 21:11:34 WARN mapred.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 mapred.JobClient: Running job: job_201510062030_0010
15/10/06 21:11:36 INFO mapred.JobClient:  map 0% reduce 0%
15/10/06 21:11:56 INFO mapred.JobClient:  map 100% reduce 0%
15/10/06 21:12:01 INFO mapred.JobClient: Job complete: job_201510062030_0010
15/10/06 21:12:01 INFO mapred.JobClient: Counters: 16
15/10/06 21:12:01 INFO mapred.JobClient:   Map-Reduce Framework
15/10/06 21:12:01 INFO mapred.JobClient:     Spilled Records=0
15/10/06 21:12:01 INFO mapred.JobClient:     Map input records=21578
15/10/06 21:12:01 INFO mapred.JobClient:     SPLIT_RAW_BYTES=146
15/10/06 21:12:01 INFO mapred.JobClient:     Map output records=21578
15/10/06 21:12:01 INFO mapred.JobClient:     Total committed heap usage (bytes)=111673344
15/10/06 21:12:01 INFO mapred.JobClient:   File Input Format Counters
15/10/06 21:12:01 INFO mapred.JobClient:     Bytes Read=17023387
15/10/06 21:12:01 INFO mapred.JobClient:   FileSystemCounters
15/10/06 21:12:01 INFO mapred.JobClient:     HDFS_BYTES_READ=20845106
15/10/06 21:12:01 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=21999
15/10/06 21:12:01 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=16786819
15/10/06 21:12:01 INFO mapred.JobClient:   Job Counters
15/10/06 21:12:01 INFO mapred.JobClient:     Launched map tasks=1
15/10/06 21:12:01 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=0
15/10/06 21:12:01 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
15/10/06 21:12:01 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=16022
15/10/06 21:12:01 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
15/10/06 21:12:01 INFO mapred.JobClient:     Data-local map tasks=1
15/10/06 21:12:01 INFO mapred.JobClient:   File Output Format Counters
15/10/06 21:12:01 INFO mapred.JobClient:     Bytes Written=16786819
15/10/06 21:12:01 INFO driver.MahoutDriver: Program took 113687 ms (Minutes: 1.8947833333333333)
JinJins-MacBook-Air:mahout-distribution-0.7 DoubleJ$ ./bin/mahout seqdumper -i hdfs-mahout/kmeans/clusteredPoints/part-m-00000 > /Users/DoubleJ/Workspace/mahout-work/cluster-docs.txt
Warning: $HADOOP_HOME is deprecated.

Warning: $HADOOP_HOME is deprecated.

15/10/06 21:16:52 INFO common.AbstractJob: Command line arguments: {--endPhase=[2147483647], --input=[hdfs-mahout/kmeans/clusteredPoints/part-m-00000], --startPhase=[0], --tempDir=[temp]}
15/10/06 21:17:20 INFO driver.MahoutDriver: Program took 27762 ms (Minutes: 0.4627)
JinJins-MacBook-Air:mahout-distribution-0.7 DoubleJ$
```

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.