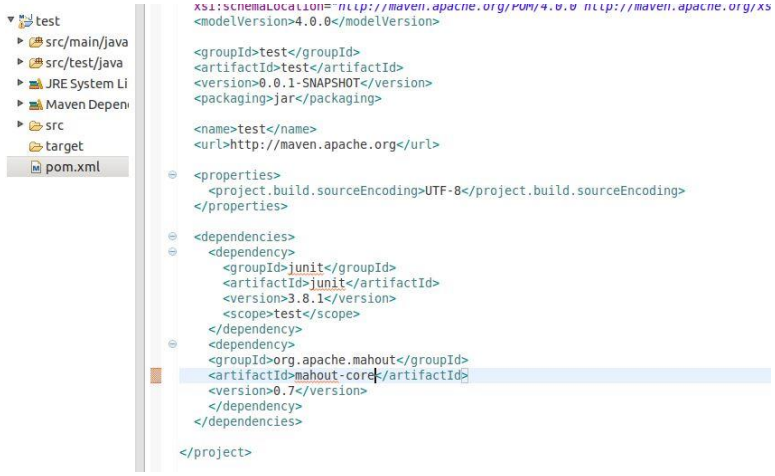
Anubha Bhargava
Homework 2

Recommendation:

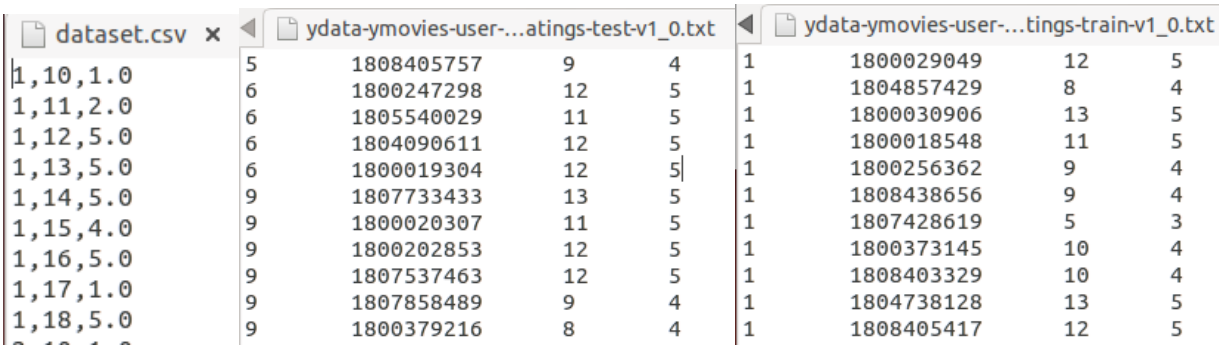
In order to run the recommendation algorithm, I installed Eclipse and configured the pom.xml file as shown below. I also built Maven in eclipse, shown by the second figure.




```
xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xs
  <modelVersion>4.0.0</modelVersion>

  <groupId>test</groupId>
  <artifactId>test</artifactId>
  <version>0.0.1-SNAPSHOT</version>
  <packaging>jar</packaging>

  <name>test</name>
  <url>http://maven.apache.org</url>

  <properties>
    <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
  </properties>

  <dependencies>
    <dependency>
      <groupId>junit</groupId>
      <artifactId>junit</artifactId>
      <version>3.8.1</version>
      <scope>test</scope>
    </dependency>
    <dependency>
    <groupId>org.apache.mahout</groupId>
    <artifactId>mahout-core</artifactId>
    <version>0.7</version>
    </dependency>
  </dependencies>

</project>
```

```
<terminated> /usr/lib/jvm/jdk1.6.0_45/bin/java (Oct 21, 2015, 6:23:16 AM)
Downloading: http://repo.maven.apache.org/maven2/org/apache/maven/maven-archiver/2.4.2/maven-a
Downloading: http://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-archiver/2.0.1/ple
Downloaded: http://repo.maven.apache.org/maven2/org/apache/maven/maven-archiver/2.4.2/maven-ar
Downloaded: http://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-archiver/2.0.1/plex
[INFO] Building jar: /home/bigdata/workspace/Clustering/target/Clustering-0.0.1-SNAPSHOT.jar
[INFO]
[INFO] --- maven-install-plugin:2.3.1:install (default-install) @ Clustering ---
[INFO] Installing /home/bigdata/workspace/Clustering/target/Clustering-0.0.1-SNAPSHOT.jar to /
[INFO] Installing /home/bigdata/workspace/Clustering/pom.xml to /home/bigdata/.m2/repository/R
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 6.734s
[INFO] Finished at: Wed Oct 21 06:23:25 PDT 2015
[INFO] Final Memory: 6M/30M
[INFO] ------------------------------------------------------------------------
```

Then, I created a user-based recommendation algorithm and ran it on a set of a few values called dataset.csv. Then, I ran the same algorithm on two sets of data from Yahoo!. All 3 files are shown below.



dataset.csv

```
1,10,1.0
1,11,2.0
1,12,5.0
1,13,5.0
1,14,5.0
1,15,4.0
1,16,5.0
1,17,1.0
1,18,5.0
```

ydata-ymovies-user-...atings-test-v1_0.txt

5	1808405757	9	4
6	1800247298	12	5
6	1805540029	11	5
6	1804090611	12	5
6	1800019304	12	5
9	1807733433	13	5
9	1800020307	11	5
9	1800202853	12	5
9	1807537463	12	5
9	1807858489	9	4
9	1800379216	8	4

ydata-ymovies-user-...tings-train-v1_0.txt

1	1800029049	12	5
1	1804857429	8	4
1	1800030906	13	5
1	1800018548	11	5
1	1800256362	9	4
1	1808438656	9	4
1	1807428619	5	3
1	1800373145	10	4
1	1808403329	10	4
1	1804738128	13	5
1	1808405417	12	5

The user-based algorithm I used is shown below, used on both Yahoo! datasets. The first was used on the Yahoo! Movies User Ratings Test v.1.0 txt file (the second text file shown above).

Anubha Bhargava
Homework 2

```
package RecommenderTest1.Test1;

import java.io.File;

public class App
{
    public static void main( String[] args ) throws IOException, TasteException
    {
        DataModel model = new FileDataModel(new File("/home/bigdata/Documents/Homework2/Ya
        UserSimilarity similarity = new PearsonCorrelationSimilarity(model);
        UserNeighborhood neighborhood = new ThresholdUserNeighborhood(0.1, similarity, mod
        UserBasedRecommender recommender = new GenericUserBasedRecommender(model, neighbor
        List<RecommendedItem> recommendations = recommender.recommend(22,2);
        for (RecommendedItem recommendation: recommendations){
            System.out.println(recommendation);
        }
    }
}
```

```
Problems  @ Javadoc  Declaration  Console  Properties
<terminated> App (1) [Java Application] /usr/lib/jvm/jdk1.6.0_45/bin/java (Oct 20, 2015, 3:04:14 PM)
3 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Creating FileDataModel
72 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Reading file info...
621 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Read lines: 10136
826 [main] INFO org.apache.mahout.cf.taste.impl.model.GenericDataModel - Processed 2309 users
RecommendedItem[item:1802771079, value:13.0]
RecommendedItem[item:1808411996, value:13.0]
```

It produced two recommended items, as shown in the console. Below I ran user-based recommendation on the Yahoo! Movies User Ratings Train v1.0 txt file (the third text file shown above) and received one single output.

```
App.java  Test1/pom.xml  Test1/pom.xml
package RecommenderTest1.Test1;

import java.io.File;

public class App
{
    public static void main( String[] args ) throws IOException, TasteException
    {
        DataModel model = new FileDataModel(new File("/home/bigdata/Documents/Homework2/Ya
        UserSimilarity similarity = new PearsonCorrelationSimilarity(model);
        UserNeighborhood neighborhood = new ThresholdUserNeighborhood(0.1, similarity, mod
        UserBasedRecommender recommender = new GenericUserBasedRecommender(model, neighbor
        List<RecommendedItem> recommendations = recommender.recommend(9,1);
        for (RecommendedItem recommendation: recommendations){
            System.out.println(recommendation);
        }
    }
}
```

```
Problems  @ Javadoc  Declaration  Console  Properties
<terminated> App (1) [Java Application] /usr/lib/jvm/jdk1.6.0_45/bin/java (Oct 20, 2015, 3:02:58 PM)
10 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Creating FileDataMode
160 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Reading file info...
677 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Read lines: 10136
796 [main] INFO org.apache.mahout.cf.taste.impl.model.GenericDataModel - Processed 2309 users
RecommendedItem[item:1800183448, value:13.0]
```

I tried another algorithm on the Yahoo! datasets. I changed the algorithms from user-based recommendation to item-based recommendation and received different recommended outputs:

Anubha Bhargava
Homework 2

```java
import java.util.List;

import org.apache.mahout.cf.taste.common.TasteException;
import org.apache.mahout.cf.taste.impl.model.file.FileDataModel;
import org.apache.mahout.cf.taste.impl.recommender.GenericItemBasedRecommender;
import org.apache.mahout.cf.taste.impl.similarity.PearsonCorrelationSimilarity;
import org.apache.mahout.cf.taste.model.DataModel;
import org.apache.mahout.cf.taste.recommender.RecommendedItem;
import org.apache.mahout.cf.taste.similarity.ItemSimilarity;

public class App
{
    public static void main( String[] args ) throws IOException, TasteException
    {
        DataModel model = new FileDataModel(new File("/home/bigdata/Documents/Homework2/Ya
        ItemSimilarity similarity = new PearsonCorrelationSimilarity(model);
        GenericItemBasedRecommender recommender = new GenericItemBasedRecommender(model, s
        List<RecommendedItem> recommendations = recommender.recommend(1,2);
        for (RecommendedItem recommendation: recommendations){
            System.out.println(recommendation);
        }
    }
}
```

```
Problems  Javadoc  Declaration  Console  Properties

<terminated> App (3) [Java Application] /usr/lib/jvm/jdk1.7.0_79/bin/java (Oct 29, 2015, 11:38:36 AM)
8 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Creating FileDataModel
92 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Reading file info...
1445 [main] INFO org.apache.mahout.cf.taste.impl.model.file.FileDataModel - Read lines: 211231
1806 [main] INFO org.apache.mahout.cf.taste.impl.model.GenericDataModel - Processed 7642 users
RecommendedItem[item:1800127818, value:13.0]
RecommendedItem[item:1800147796, value:13.0]
```

Clustering:

I ran clustering on the synthetic control data, Reuters data and Wikipedia data. After putting the synthetic_control.data file onto HDFS, I ran the clustering algorithm and received the following output results:

```
bigdata@ubuntu:/usr/local/eclipse/mahout-distribution-0.9$ hadoop jar mahout-exa
mples-0.9-job.jar org.apache.mahout.clustering.syntheticcontrol.kmeans.Job
```

```
20.841,21.189,27.999,33.274,37.098,46.94,37.891,31.12,29.039,18.015,21.467,26.29
,31.476,36.344,42.412,39.284,34.666,27.746,19.489,17.466,14.161,18.294,26.514,34
.013,40.329,38.44,38.222,26.448,20.689,17.714,14.618,14.382]
        1.0 : [distance=76.52337014895532]: [26.24,35.063,48.123,49.951,35.547,4
0.524,24.295,20.726,19.711,21.393,25.407,31.618,37.473,35.773,38.988,45.535,37.1
95,26.407,18.761,24.149,11.399,20.869,25.102,33.357,36.633,37.64,41.725,42.374,3
4.15,32.215,22.921,17.116,13.989,15.962,21.162,38.315,35.831,46.373,47.746,33.05
5,34.929,22.782,13.369,21.997,16.906,20.107,32.456,43.196,42.651,47.739,41.398,4
0.825,25.242,17.258,14.702,17.765,17.287,28.627,31.652,38.961]
        1.0 : [distance=56.71859744952735]: [26.891,43.616,46.491,41.635,42.21,3
4.433,20.803,13.055,14.923,15.482,30.787,38.131,39.014,41.858,40.022,35.192,23.0
42,22.366,11.799,18.675,24.888,34.396,38.611,45.156,43.438,38.218,28.174,24.715,
20.483,17.374,22.727,36.105,44.437,42.92,43.577,42.392,35.561,27.745,15.829,21.8
64,18.412,29.645,39.517,44.572,38.629,40.567,35.29,22.683,16.195,19.235,24.257,2
2.592,34.713,44.764,39.443,40.623,36.059,24.643,23.876,16.2]
        1.0 : [distance=45.84111577700582]: [30.889,32.489,37.018,39.645,45.914,
31.508,21.637,21.944,16.339,11.605,23.196,26.46,42.842,40.764,45.469,45.3,37.122
,23.297,21.292,12.301,12.031,23.394,28.881,37.382,37.317,45.073,35.787,39.555,28
.424,14.686,18.112,15.185,26.378,27.795,35.08,37.005,37.855,37.93,36.44,25.022,1
6.131,18.664,18.489,25.011,23.513,42.732,42.599,47.651,36.176,39.035,25.058,21.8
18,10.813,22.535,21.182,31.802,31.936,37.568,46.937,35.429]
15/10/24 14:38:30 INFO ClusterDumper: Wrote 6 clusters
15/10/24 14:38:30 INFO MahoutDriver: Program took 413661 ms (Minutes: 6.89435)
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$
```

For the Reuters dataset, I copied the Reuters dataset to the directory and extracted the articles from .sgm files to .txt files. The process is shown in the below 3 figures.

Anubha Bhargava
Homework 2

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/mahout org.
apache.lucene.benchmark.utils.ExtractReuters /home/bigdata/Documents/Homework2/Reuters/Reut
ers21578/ /home/bigdata/Documents/Homework2/Reuters/Reuters-Outk2/Reuters/Reuter
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/target/mahou
t-examples-0.11.1-SNAPSHOT-job.jar
15/10/27 20:07:53 WARN MahoutDriver: No org.apache.lucene.benchmark.utils.ExtractReuters.pr
ops found on classpath, will use command-line arguments only
Deleting all files in /home/bigdata/Documents/Homework2/Reuters/Reuters-Out-tmp
15/10/27 20:08:00 INFO MahoutDriver: Program took 6808 ms (Minutes: 0.11346666666666666)
```

```
bigdata@ubuntu:~/Documents/Homework2/Reuters/Reuters21578$ ls
all-exchanges-strings.lc.txt        README.txt     reut2-007.sgm  reut2-015.sgm
all-orgs-strings.lc.txt             reut2-000.sgm  reut2-008.sgm  reut2-016.sgm
all-people-strings.lc.txt           reut2-001.sgm  reut2-009.sgm  reut2-017.sgm
all-places-strings.lc.txt           reut2-002.sgm  reut2-010.sgm  reut2-018.sgm
all-topics-strings.lc.txt           reut2-003.sgm  reut2-011.sgm  reut2-019.sgm
cat-descriptions_120396.txt         reut2-004.sgm  reut2-012.sgm  reut2-020.sgm
feldman-cia-worldfactbook-data.txt  reut2-005.sgm  reut2-013.sgm  reut2-021.sgm
lewis.dtd                           reut2-006.sgm  reut2-014.sgm
```

Reuters-Out directory:

```
reut2-005.sgm-427.txt  reut2-010.sgm-783.txt  reut2-016.sgm-238.txt  reut2-021.sgm-73.txt
reut2-005.sgm-428.txt  reut2-010.sgm-784.txt  reut2-016.sgm-239.txt  reut2-021.sgm-74.txt
reut2-005.sgm-429.txt  reut2-010.sgm-785.txt  reut2-016.sgm-23.txt   reut2-021.sgm-75.txt
reut2-005.sgm-42.txt   reut2-010.sgm-786.txt  reut2-016.sgm-240.txt  reut2-021.sgm-76.txt
reut2-005.sgm-430.txt  reut2-010.sgm-787.txt  reut2-016.sgm-241.txt  reut2-021.sgm-77.txt
reut2-005.sgm-431.txt  reut2-010.sgm-788.txt  reut2-016.sgm-242.txt  reut2-021.sgm-78.txt
reut2-005.sgm-432.txt  reut2-010.sgm-789.txt  reut2-016.sgm-243.txt  reut2-021.sgm-79.txt
```

Afterwards, I converted the raw data into a Hadoop sequence file:

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/mahout seqdirectory -i /
uters/Reuters-Out/ -o /Reuters/Reuters-Out-SeqDir/ -c UTF-8 -chunk 5
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/target/mahout-examples-0.
.1-SNAPSHOT-job.jar
```

```
5/10/27 21:04:29 INFO Job: Counters: 23
        File System Counters
                FILE: Number of bytes read=237483063
                FILE: Number of bytes written=242307932
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=49018603
                HDFS: Number of bytes written=30146692
                HDFS: Number of read operations=239338
                HDFS: Number of large read operations=88
                HDFS: Number of write operations=24
        Map-Reduce Framework
                Map input records=21578
                Map output records=21578
                Input split bytes=1724120
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=310
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=369098752
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=10827429
5/10/27 21:04:29 INFO MahoutDriver: Program took 76672 ms (Minutes: 1.277866666
666667)
```

Then, I generated vectors from the sequence file:

Anubha Bhargava
Homework 2

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/mahout seq2sparse -i /Reuters/R
euters-Out-SeqDir/ -o /Reuters/Reuters-Out-SeqDir-sparse-kmeans
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/target/mahout-examples-0.11.1-SN
APSHOT-job.jar
```

```
15/10/27 21:09:19 INFO Job: Job job_local1847790156_0009 completed successfully
15/10/27 21:09:19 INFO Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=1371814836
                FILE: Number of bytes written=1404282137
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=298184980
                HDFS: Number of bytes written=224009025
                HDFS: Number of read operations=449
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=128
        Map-Reduce Framework
                Map input records=21578
                Map output records=21578
                Map output bytes=16606217
                Map output materialized bytes=16688735
                Input split bytes=158
                Combine input records=0
                Combine output records=0
                Reduce input groups=21578
                Reduce shuffle bytes=16688735
                Reduce input records=21578
                Reduce output records=21578
                Spilled Records=43156
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=176
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=394280960
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=16906791
        File Output Format Counters
                Bytes Written=16906791
15/10/27 21:09:19 INFO HadoopUtil: Deleting /Reuters/Reuters-Out-SeqDir-sparse-kmeans/partial-vectors-0
15/10/27 21:09:19 INFO MahoutDriver: Program took 58373 ms (Minutes: 0.9728833333333333)
```

I ran the commands to cluster with k-means as shown below:

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/mahout kmeans -i /Reuters/Reute
rs-Out-SeqDir-sparse-kmeans/tfidf-vectors/ -c /Reuters/kmeans-clusters -o /Reuters/reuters-kmeans -dm org.apach
e.mahout.common.distance.CosineDistanceMeasure -cd 0.1 -x 10 -k 20 -ow -cl
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/target/mahout-examples-0.11.1-SN
APSHOT-job.jar
```

Anubha Bhargava
Homework 2

```
                HDFS: Number of bytes written=29273258
                HDFS: Number of read operations=188
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=49
        Map-Reduce Framework
                Map input records=21579
                Map output records=21579
                Input split bytes=154
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=54
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=183730176
        File Input Format Counters
                Bytes Read=16909529
        File Output Format Counters
                Bytes Written=17927868
15/10/27 22:11:16 INFO MahoutDriver: Program took 28898 ms (Minutes: 0.481633333
33333336)
```

Lastly, I dumped the results to files.

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/
mahout clusterdump -i /Reuters/Reuters-kmeans/clusters-*-final -d /Reuters/Reute
rs-Out-SeqDir-sparse-kmeans/dictionary.file-0 -dt sequencefile -o /home/bigdata/
Documents/Homework2/Reuters/Reuters-kmeans-dump -n 5 -b 100
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/t
arget/mahout-examples-0.11.1-SNAPSHOT-job.jar
15/10/27 22:28:35 INFO AbstractJob: Command line arguments: {--dictionary=[/Reut
ers/Reuters-Out-SeqDir-sparse-kmeans/dictionary.file-0], --dictionaryType=[seque
ncefile], --distanceMeasure=[org.apache.mahout.common.distance.SquaredEuclideanD
istanceMeasure], --endPhase=[2147483647], --input=[/Reuters/Reuters-kmeans/clust
ers-*-final], --numWords=[5], --output=[/home/bigdata/Documents/Homework2/Reuter
s/Reuters-kmeans-dump], --outputFormat=[TEXT], --startPhase=[0], --substring=[10
0], --tempDir=[temp]}
15/10/27 22:28:41 INFO ClusterDumper: Wrote 20 clusters
15/10/27 22:28:41 INFO MahoutDriver: Program took 5276 ms (Minutes: 0.0879333333
3333334)
```

When viewing the output, I received these results:

Anubha Bhargava
Homework 2

```
                he                                      =>  1.6951136977467316
:{"r":[{"0.003":0.338},{"0.006913":0.338},{"0.007050":0.239},{"0.05":0.206},{"0
06":0.239},{"0.07":0.
        Top Terms:
                bank                                    =>  2.9254567329312713
                said                                    =>  2.9071813958211683
                banks                                   =>  2.7144628161555726
                he                                      =>   2.446499566656883
                market                                  =>  2.2318716627051574
:{"r":[{"0":0.26},{"0.01":0.314},{"0.59":0.246},{"00":0.606},{"00.03":0.235},{"
0.11":0.231},{"00.13"
        Top Terms:
                said                                    =>  1.7359546045931618
                its                                     =>  1.2637971808591226
                inc                                     =>   1.234903767665853
                corp                                    =>  1.1474406615955624
                mar                                     =>    1.09485300835662
:{"r":[{"00":0.622},{"00.18":0.444},{"00.20":0.421},{"00.56":0.444},{"00.84":0.
11},{"00.89":0.435},{
        Top Terms:
                officer                                 =>   3.826718148220791
                president                               =>  3.8211704173784575
                chief                                   =>   3.813937988709868
                executive                               =>  3.8014617866344667
                chairman                                =>  2.9587232139673127
:{"r":[{"0.01":0.291},{"0.02":0.222},{"0.05":0.349},{"0.07":0.349},{"0.1":0.736
,{"0.10":0.576},{"0.1
```

```
Key: 18816: Value: wt: 1.0 distance: 0.6519258161127899  vec: [{"2689":3.135},{"
3731":3.051},{"5085":3.633},{"5245":3.4
Key: 7052: Value: wt: 1.0 distance: 0.8337635422085907  vec: [{"2689":3.135},{"3
731":3.051},{"5085":3.633},{"6001":1.1
Key: 18816: Value: wt: 1.0 distance: 0.5395841312700682  vec: [{"2460":2.742},{"
2689":3.135},{"3731":3.051},{"4922":3.6
Key: 7427: Value: wt: 1.0 distance: 0.6029572342398424  vec: [{"1512":7.514},{"1
528":9.602},{"1890":9.37},{"2215":9.18
Key: 3051: Value: wt: 1.0 distance: 0.8065843938926417  vec: [{"2689":3.135},{"3
731":3.051},{"3949":3.953},{"4751":4.0
Key: 5067: Value: wt: 1.0 distance: 0.8371259470140552  vec: [{"2689":3.135},{"3
731":3.051},{"4485":7.647},{"4751":4.0
Key: 21148: Value: wt: 1.0 distance: 0.8128485258514935  vec: [{"2490":7.129},{"
2689":3.135},{"3731":3.051},{"3862":8.4
Key: 5067: Value: wt: 1.0 distance: 0.8393221999654109  vec: [{"2563":8.677},{"2
689":3.135},{"3556":3.035},{"3731":3.0
Key: 7427: Value: wt: 1.0 distance: 0.544270057257653  vec: [{"834":8.677},{"155
5":7.453},{"2689":3.135},{"3731":3.051
Count: 21579
15/10/27 22:30:44 INFO MahoutDriver: Program took 15795 ms (Minutes: 0.26325)
```

I conducted the same steps for the Wikipedia file – enwiki-latest-pages-articles1.xml-p000000010p000010000. The first step for the Wikipedia files was slightly different. It is shown below.

Anubha Bhargava
Homework 2

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/
mahout seqwiki -all -i /Wikipedia/enwiki-latest-pages-articles1.xml-p000000010p0
00010000 -o /Wikipedia/Wikipedia-out-seqdir
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/t
arget/mahout-examples-0.11.1-SNAPSHOT-job.jar
15/10/29 10:57:43 INFO WikipediaToSequenceFile: Input: /Wikipedia/enwiki-latest-
pages-articles1.xml-p000000010p000010000 Out: /Wikipedia/Wikipedia-out-seqdir Ca
tegories:  All Files: true
```

```
                Combine input records=0
                Combine output records=0
                Reduce input groups=6269
                Reduce shuffle bytes=159715397
                Reduce input records=6269
                Reduce output records=6269
                Spilled Records=17589
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=2554
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=732966912
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=168902037
        File Output Format Counters
                Bytes Written=159832019
15/10/29 10:54:19 INFO MahoutDriver: Program took 145336 ms (Minutes: 2.42226666
6666667)
```

I generated vectors from the sequence file.

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/
mahout seq2sparse -i /Wikipedia/Wikipedia-out-seqdir -o /Wikipedia/Wikipedia-out
-seqdir-sparse-kmeans
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/t
arget/mahout-examples-0.11.1-SNAPSHOT-job.jar
```

Then, I clustered the produced vectors with k-means.

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/
mahout kmeans -i /Wikipedia/Wikipedia-out-seqdir-sparse-kmeans/tfidf-vectors/ -c
 /Wikipedia/kmeans-clusters -o /Wikipedia/Wikipedia-kmeans -dm org.apache.mahout
.common.distance.CosineDistanceMeasure -cd 0.1 -x 10 -k 20 -ow -cl
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/t
arget/mahout-examples-0.11.1-SNAPSHOT-job.jar
```

Anubha Bhargava
Homework 2

Finally, I dumped the clusters into files.

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./bin/
mahout clusterdump -i /Wikipedia/Wikipedia-kmeans/clusters-*-final -d /Wikipedia
/Wikipedia-out-seqdir-sparse-kmeans/dictionary.file-0 -dt sequencefile -o /home/
bigdata/Documents/Homework2/Wikipedia/Wikipedia-kmeans-dump -n 5 -b 100
Running on hadoop, using /usr/local/hadoop-2.5.0/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/apache-mahout-distribution-0.11.0/mahout-trunk/examples/t
arget/mahout-examples-0.11.1-SNAPSHOT-job.jar
15/10/29 11:26:21 INFO AbstractJob: Command line arguments: {--dictionary=[/Wiki
pedia/Wikipedia-out-seqdir-sparse-kmeans/dictionary.file-0], --dictionaryType=[s
equencefile], --distanceMeasure=[org.apache.mahout.common.distance.SquaredEuclid
eanDistanceMeasure], --endPhase=[2147483647], --input=[/Wikipedia/Wikipedia-kmea
ns/clusters-*-final], --numWords=[5], --output=[/home/bigdata/Documents/Homework
2/Wikipedia/Wikipedia-kmeans-dump], --outputFormat=[TEXT], --startPhase=[0], --s
ubstring=[100], --tempDir=[temp]}
15/10/29 11:26:36 INFO ClusterDumper: Wrote 20 clusters
15/10/29 11:26:36 INFO MahoutDriver: Program took 14846 ms (Minutes: 0.247433333
33333334)
```

Here are the generated results from running the Wikipedia commands:

```
:{"r":[{"11th":0.724},{"1914":0.688},{"1956":0.641},{"1970s":0.592},{"19th":0.53
},{"2008":0.355},{"4t
        Top Terms:
                nahmc                                   =>   4.480807799559373
                subpage                                 =>  1.8690807635967548
                redirect                                =>   1.816250360929049
                aba                                     =>  1.5084163959209735
                relations                               =>  1.4852837232443004
:{"r":[{"0":1.305},{"0,3":0.922},{"0.0":0.703},{"0.00":0.852},{"0.05":2.146},{"0
.1":0.78},{"0.13":0.7
        Top Terms:
                telecommunications                      =>   7.862010623469497
```

```
Key: 5023: Value: wt: 1.0 distance: 0.8115312160816286  vec: [{"2500":2.457},{"6
186":3.148},{"8517":1.999},{"10932":2.
Key: 5035: Value: wt: 1.0 distance: 0.7793875613407406  vec: [{"102587":2.858},{
"127111":3.684},{"190246":6.91},{"2032
Key: 5023: Value: wt: 1.0 distance: 0.7280789820938538  vec: [{"0":3.75},{"1862"
:2.479},{"3334":3.001},{"4537":2.329},
Key: 5035: Value: wt: 1.0 distance: 0.6059475083876485  vec: [{"84623":2.871},{"
130923":4.023},{"226150":2.146}]
Key: 5035: Value: wt: 1.0 distance: 0.7590970363806326  vec: [{"84737":8.645},{"
226150":2.146}]
Key: 6048: Value: wt: 1.0 distance: 0.81
```

Classification:

I ran classification on the 20 newsgroups data and the Wikipedia data. For the 20 newsgroups data, I selected the 1st choice – cnaivebayes. Then, it proceeded to download data. The results I received are in the second figure below.

Anubha Bhargava
Homework 2

```
bigdata@ubuntu:/usr/local/eclipse/mahout-distribution-0.9$ ./examples/bin/classi
fy-20newsgroups.sh
Please select a number to choose the corresponding task to run
1. cnaivebayes
2. naivebayes
3. sgd
4. clean -- cleans up the work area in /tmp/mahout-work-bigdata
Enter your choice : 1
```

```
        0       0       0       0       1       0       1       0       1
|  394          i     = rec.motorcycles
0       1       0       0       0       0       0       0       2       386     6
        0       2       0       1       2       0       4       1       0
|  405          j     = rec.sport.baseball
1       0       1       0       0       0       1       0       1       1       3
69      0       0       0       0       1       0       0       0       2
|  377          k     = rec.sport.hockey
0       2       1       0       1       2       0       0       0       0       0
        391     1       0       0       0       0       2       1       2
|  403          l     = sci.crypt
0       2       1       9       5       2       6       2       1       1       2
        3       354     2       5       7       1       1       2       3
|  409          m     = sci.electronics
1       0       1       0       0       1       6       1       2       0       0
        0       2       418     4       3       2       6       2       2
|  451          n     = sci.med
2       0       0       0       1       0       1       2       0       0       0
        2       0       2       388     0       2       0       2       0
|  402          o     = sci.space
0       0       1       3       0       1       0       0       1       1       0
        0       1       3       2       386     2       0       4       1
|  406          p     = soc.religion.christian
3       0       0       1       0       0       0       0       0       1       0
        0       0       0       1       3       386     0       0       0
|  395          q     = talk.politics.mideast
0       0       1       0       0       2       0       0       1       0       1
        2       0       0       1       1       1       357     1       7
|  375          r     = talk.politics.guns
24      0       0       0       1       0       1       0       0       1       1
        0       0       0       0       24      1       8       189     6
|  256          s     = talk.religion.misc
3       0       1       0       0       1       1       0       0       3       2
        1       1       2       2       0       10      10      2       259
|  298          t     = talk.politics.misc

=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.8622
Accuracy                                   89.7072%
Reliability                                85.1807%
Reliability (standard deviation)             0.212
Weighted precision                          0.8967
Weighted recall                             0.8971
Weighted F1 score                           0.8957
```

I ran the classification algorithms on Wikipedia and selected CBayes.

Anubha Bhargava
Homework 2

```
bigdata@ubuntu:/usr/local/apache-mahout-distribution-0.11.0/mahout-trunk$ ./exam
ples/bin/classify-wikipedia.sh
Discovered Hadoop v2.
Setting dfs command to /usr/local/hadoop-2.5.0/bin/hdfs dfs, dfs rm to /usr/loca
l/hadoop-2.5.0/bin/hdfs dfs -rm -r -skipTrash.
Please select a number to choose the corresponding task to run
1. CBayes (may require increased heap space on yarn)
2. BinaryCBayes
3. clean -- cleans up the work area in /tmp/mahout-work-wiki
Enter your choice : 1
ok. You chose 1 and we'll use CBayes
creating work directory at /tmp/mahout-work-wiki
Downloading wikipedia XML dump
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0  282M    0 1073k    0     0  54546      0  1:30:25  0:00:20  1:30:05 55094
```

The results I received are displayed below. It provides a summary of the data categorized by country. Also, the statistics of these results are displayed at the bottom. At the top, the number of classified instances are displayed.

```
=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j
--Classified as
350     9       12      1       4       3       1       14      16      7
|  417          a     = australia
0       98      2       0       4       1       0       3       0       1
|  109          b     = austria
0       0       6       0       1       1       0       0       0       0
|  8            c     = bahamas
0       4       8       362     11      3       2       8       10      6
|  414          d     = canada
0       0       0       0       23      1       0       1       0       0
|  25           e     = colombia
0       0       1       0       1       20      0       0       0       2
|  24           f     = cuba
0       0       0       0       1       1       58      0       0       0
|  60           g     = pakistan
0       0       1       0       4       0       0       3       1       2
|  11           h     = panama
3       15      32      5       20      9       9       40      360     6
|  499          i     = united kingdom
1       0       0       0       1       0       0       0       0       48
|  50           j     = vietnam

=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.7469
Accuracy                                   82.1274%
Reliability                                73.0634%
Reliability (standard deviation)            0.3098
Weighted precision                          0.9111
Weighted recall                             0.8213
Weighted F1 score                           0.8539
```