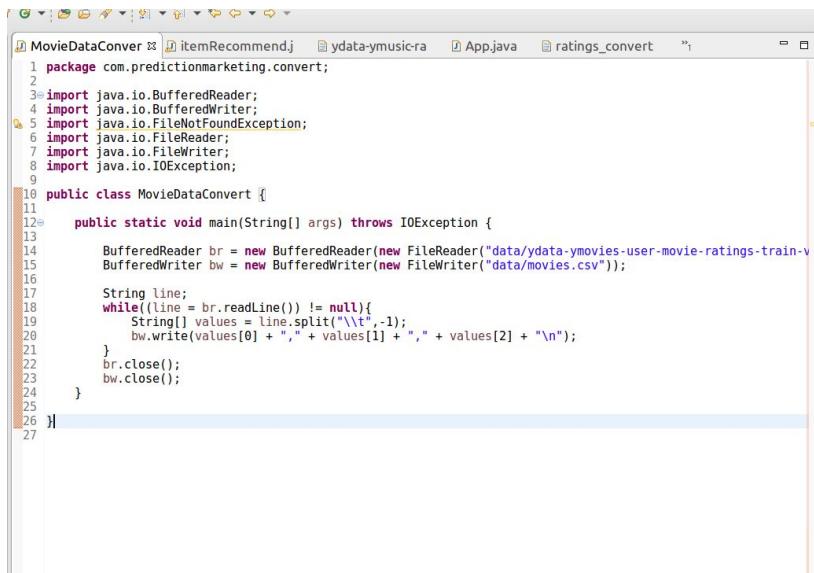


1. Recommender With Eclipse

1-1 User Based Recommender

1-1-1 Data Used ydata-ymovies-user-movie-ratings-train-v1_0.txt

1. Convert data to csv, code

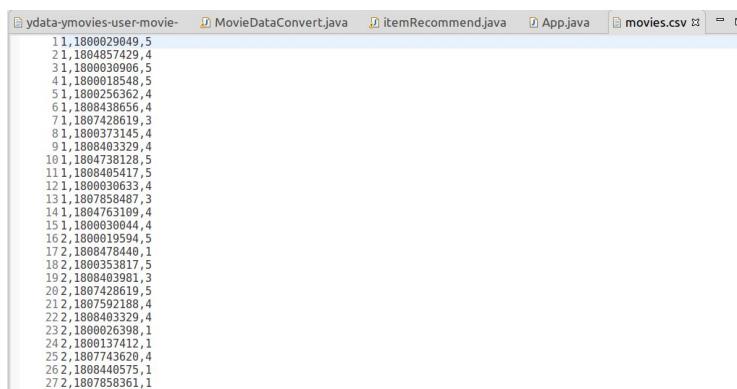


```
MovieDataConver.java itemRecommend.java ydata-ymusic-ra App.java ratings_convert

1 package com.predictionmarketing.convert;
2
3 import java.io.BufferedReader;
4 import java.io.BufferedWriter;
5 import java.io.FileNotFoundException;
6 import java.io.FileReader;
7 import java.io.FileWriter;
8 import java.io.IOException;
9
10 public class MovieDataConvert {
11     public static void main(String[] args) throws IOException {
12         BufferedReader br = new BufferedReader(new FileReader("data/ydata-ymovies-user-movie-ratings-train-v1_0.txt"));
13         BufferedWriter bw = new BufferedWriter(new FileWriter("data/movies.csv"));
14
15         String line;
16         while((line = br.readLine()) != null){
17             String[] values = line.split("\t",-1);
18             bw.write(values[0] + "," + values[1] + "," + values[2] + "\n");
19         }
20         br.close();
21         bw.close();
22     }
23 }
24
25
26 }

27
```

2. Converted data



```
movies.csv

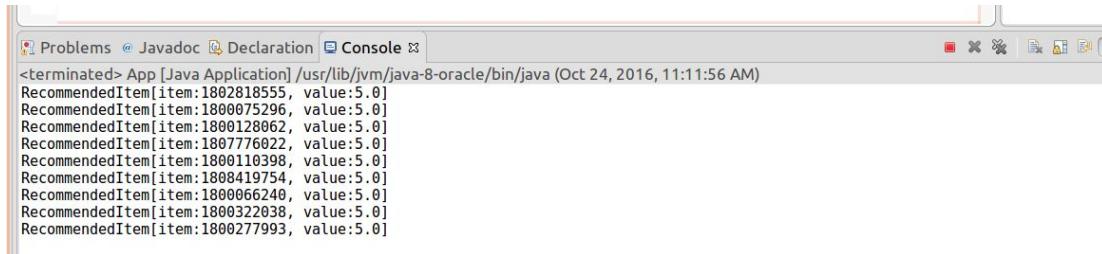
11,1880029849,5
21,1884857429,4
31,1880030986,5
41,1880018549,5
51,1880256362,4
61,1888438656,4
71,1887428619,3
81,1880373145,4
91,1888403329,4
101,1884738128,5
111,1888405417,5
121,1880030633,4
131,1887858487,3
141,1884763109,4
151,1888403329,4
161,1880019594,5
172,1888478440,1
182,1880353817,5
192,1888403081,3
202,1887428619,5
212,18875792188,4
222,1888403329,4
232,1880026398,1
242,1880137412,1
252,1887743628,4
262,1888440575,1
272,1887858361,1
```

3. Build Recommender



```
1 package com.predictionmarketing.RecommenderApp;
2
3 import java.io.File;
4
5 /**
6  * Hello world!
7  *
8  */
9 public class App
10 {
11     public static void main( String[] args ) throws Exception
12     {
13         DataModel model = new FileDataModel(new File("data/movies.csv"));
14         UserSimilarity similarity = new PearsonCorrelationSimilarity(model);
15         UserNeighborhood neighborhood = new ThresholdUserNeighborhood(0.1, similarity, model);
16         UserBasedRecommender recommender =
17             new GenericUserBasedRecommender(model, neighborhood, similarity);
18         List<RecommendedItem> recommendations = recommender.recommend(3, 10);
19         for (RecommendedItem recommendation : recommendations) {
20             System.out.println(recommendation);
21         }
22     }
23 }
```

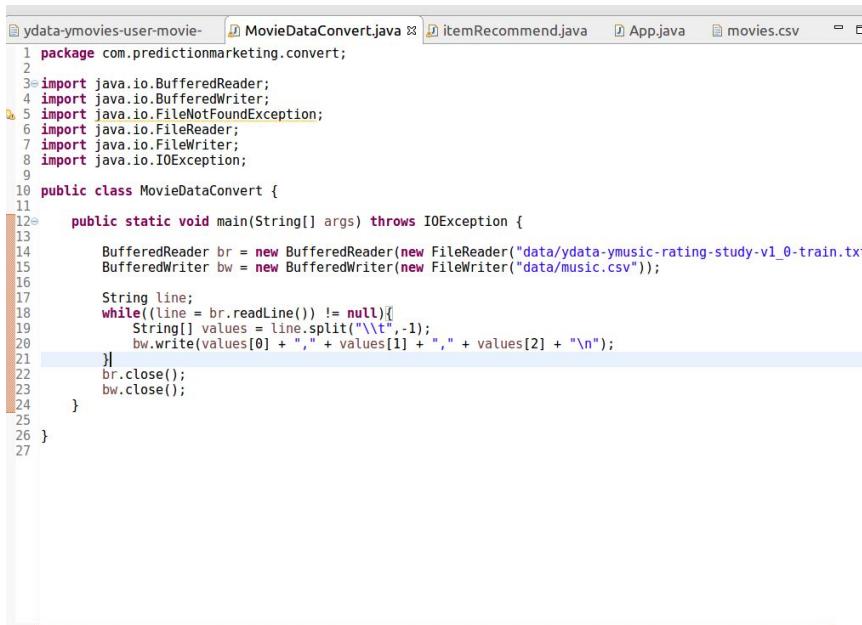
4. Recommender Result



```
<terminated> App [Java Application] /usr/lib/jvm/java-8-oracle/bin/java (Oct 24, 2016, 11:11:56 AM)
RecommendedItem[item:1802818555, value:5.0]
RecommendedItem[item:1800075296, value:5.0]
RecommendedItem[item:1800128062, value:5.0]
RecommendedItem[item:1807776022, value:5.0]
RecommendedItem[item:1800110398, value:5.0]
RecommendedItem[item:1808419754, value:5.0]
RecommendedItem[item:1800066240, value:5.0]
RecommendedItem[item:1800322038, value:5.0]
RecommendedItem[item:1800277993, value:5.0]
```

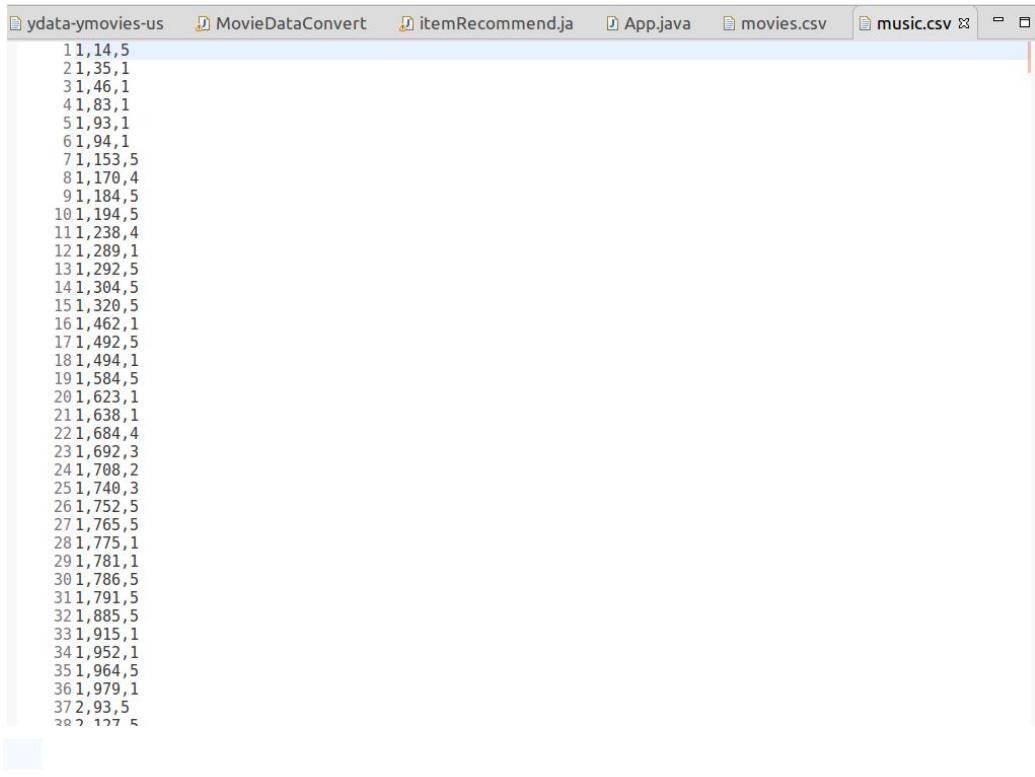
1-1-2 Data Used ydata-ymusic-rating-study-v1_0-train.txt

1.Code to convert to csv



```
1 package com.predictionmarketing.convert;
2
3 import java.io.BufferedReader;
4 import java.io.BufferedWriter;
5 import java.io.FileNotFoundException;
6 import java.io.FileReader;
7 import java.io.FileWriter;
8 import java.io.IOException;
9
10 public class MovieDataConvert {
11
12     public static void main(String[] args) throws IOException {
13
14         BufferedReader br = new BufferedReader(new FileReader("data/ydata-ymusic-rating-study-v1_0-train.txt"));
15         BufferedWriter bw = new BufferedWriter(new FileWriter("data/music.csv"));
16
17         String line;
18         while((line = br.readLine()) != null){
19             String[] values = line.split("\t",-1);
20             bw.write(values[0] + "," + values[1] + "," + values[2] + "\n");
21         }
22         br.close();
23         bw.close();
24     }
25
26 }
```

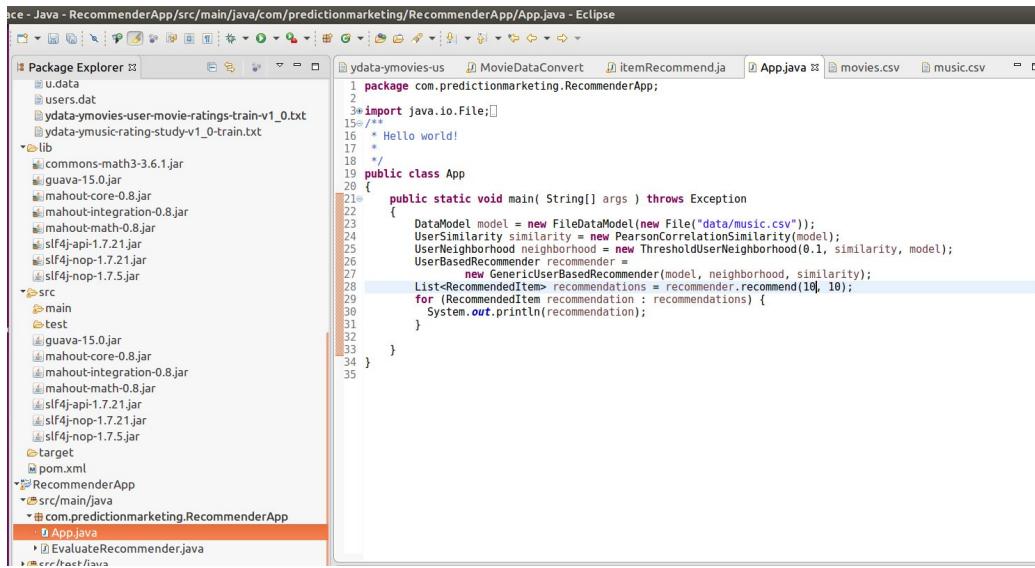
2.CSV result



The screenshot shows a code editor window with multiple tabs. The active tab is 'music.csv', which contains the following data:

```
11,14,5  
21,35,1  
31,46,1  
41,83,1  
51,93,1  
61,94,1  
71,153,5  
81,170,4  
91,184,5  
101,194,5  
111,238,4  
121,289,1  
131,292,5  
141,304,5  
151,320,5  
161,462,1  
171,492,5  
181,494,1  
191,584,5  
201,623,1  
211,638,1  
221,684,4  
231,692,3  
241,708,2  
251,740,3  
261,752,5  
271,765,5  
281,775,1  
291,781,1  
301,786,5  
311,791,5  
321,885,5  
331,915,1  
341,952,1  
351,964,5  
361,979,1  
372,93,5
```

3. Code for recommender

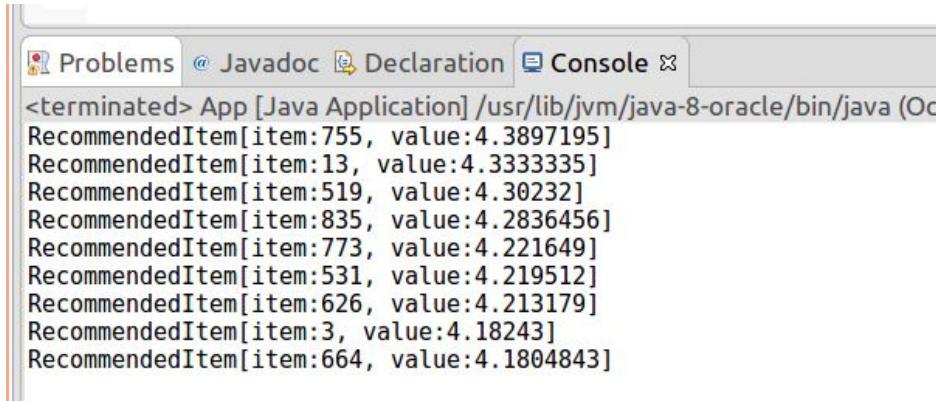


The screenshot shows the Eclipse IDE interface with the following details:

- Project Explorer:** Shows the project structure with packages like 'u.data', 'users.dat', 'ydata-ymovies-user-movie-ratings-train-v1_0.txt', 'ydata-ymusic-rating-studyv1_0-train.txt', 'lib' containing various JAR files, and 'src' containing 'main' and 'test' packages.
- Code Editor:** The 'App.java' file is open, displaying Java code for a recommender system. The code imports java.io.File, defines a public class App with a main method, and uses Mahout's UserBasedRecommender to generate recommendations based on Pearson Correlation Similarity.
- File Explorer:** Shows files like 'pom.xml' and 'EvaluateRecommender.java'.

```
1 package com.predictionmarketing.RecommenderApp;  
2  
3 import java.io.File;  
4  
5 /**  
6  * Hello world!  
7 */  
8  
9 public class App  
10 {  
11     public static void main( String[] args ) throws Exception  
12     {  
13         DataModel model = new FileDataModel(new File("data/music.csv"));  
14         UserSimilarity similarity = new PearsonCorrelationSimilarity(model);  
15         UserNeighborhood neighborhood = new ThresholdUserNeighborhood(0.1, similarity, model);  
16         UserBasedRecommender recommender =  
17             new GenericUserBasedRecommender(model, neighborhood, similarity);  
18         List<RecommendedItem> recommendations = recommender.recommend(10, 10);  
19         for (RecommendedItem recommendation : recommendations) {  
20             System.out.println(recommendation);  
21         }  
22     }  
23 }  
24  
25 }
```

4. Recommender Result

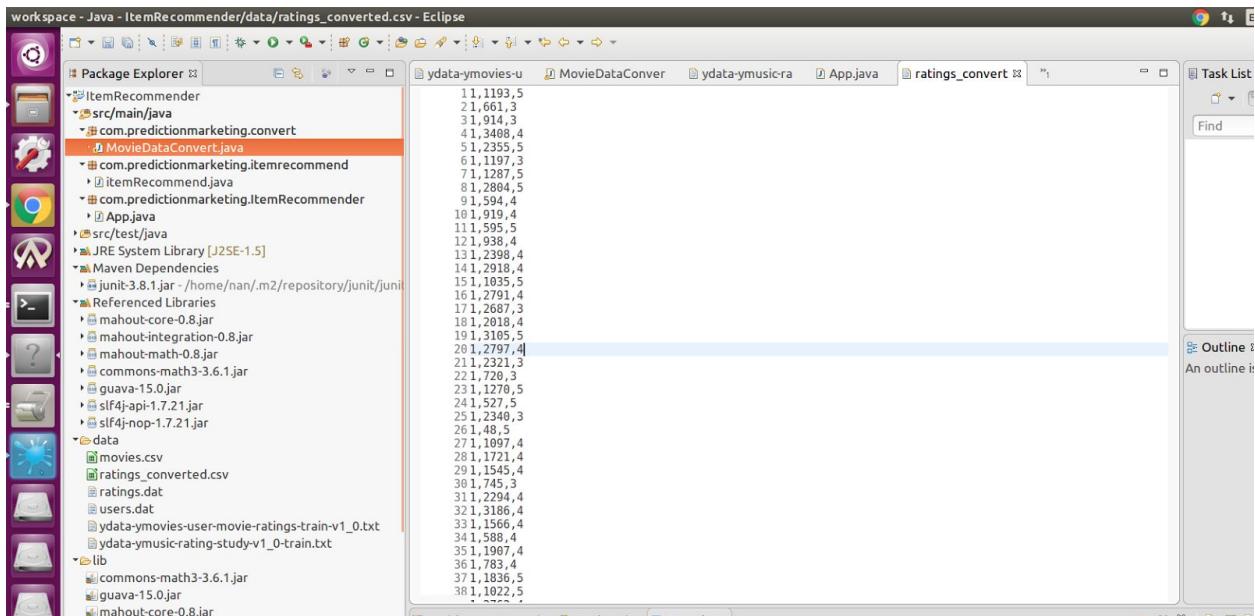


```
<terminated> App [Java Application] /usr/lib/jvm/java-8-oracle/bin/java (OpenJDK 64-Bit Server VM)
RecommendedItem[item:755, value:4.3897195]
RecommendedItem[item:13, value:4.3333335]
RecommendedItem[item:519, value:4.30232]
RecommendedItem[item:835, value:4.2836456]
RecommendedItem[item:773, value:4.221649]
RecommendedItem[item:531, value:4.219512]
RecommendedItem[item:626, value:4.213179]
RecommendedItem[item:3, value:4.18243]
RecommendedItem[item:664, value:4.1804843]
```

1-2 Item Based Recommender

1-2-1 Data Used: <http://grouplens.org/datasets/movielens/> 1M move data set ratings.dat

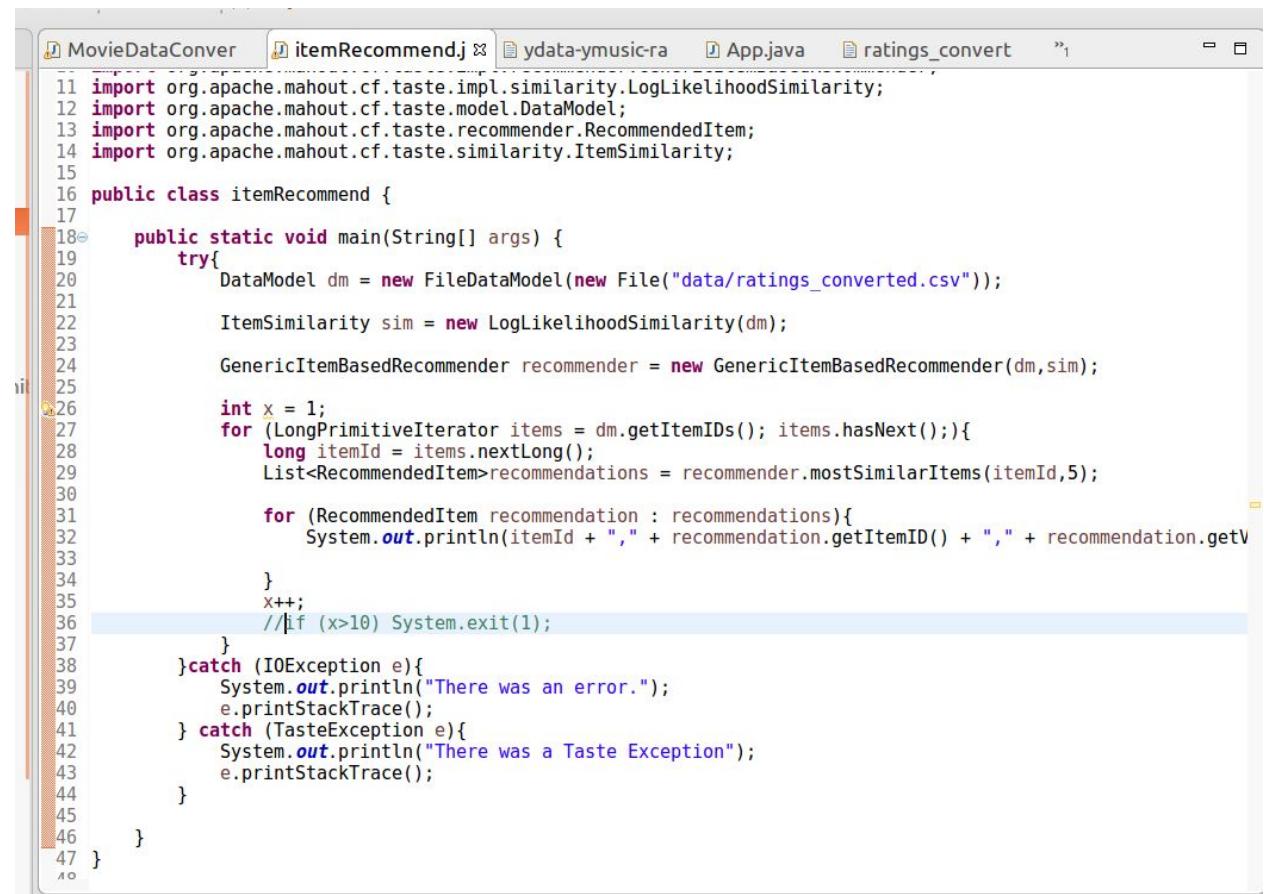
1. Convert to csv file



2. Code for converter

```
1 package com.predictionmarketing.convert;
2
3 import java.io.BufferedReader;
4 import java.io.BufferedWriter;
5 import java.io.FileNotFoundException;
6 import java.io.FileReader;
7 import java.io.FileWriter;
8 import java.io.IOException;
9
10 public class MovieDataConvert {
11
12     public static void main(String[] args) throws IOException {
13
14         BufferedReader br = new BufferedReader(new FileReader("data/ratings.dat"));
15         BufferedWriter bw = new BufferedWriter(new FileWriter("data/ratings_converted.csv"));
16
17         String line;
18         while((line = br.readLine()) != null){
19             String[] values = line.split("::",-1);
20             bw.write(values[0] + "," + values[1] + "," + values[2] + "\n");
21         }
22         br.close();
23         bw.close();
24     }
25
26 }
27
```

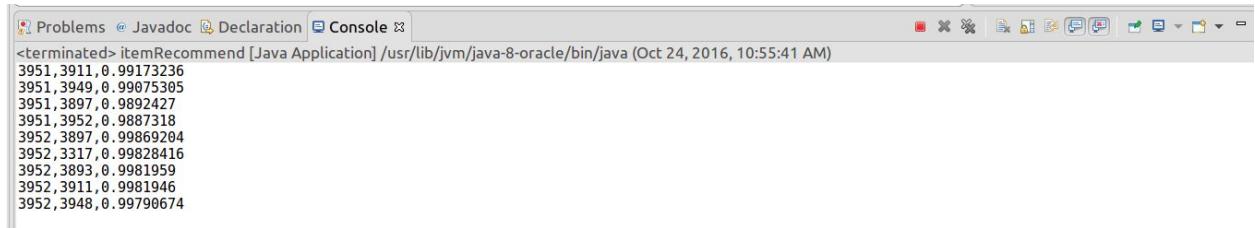
3. Build Recomemder



The screenshot shows an IDE interface with multiple tabs at the top: MovieDataConver, itemRecommend, ydata-ymusic-ra, App.java, ratings_convert, and a file icon. The itemRecommend tab is active, displaying the following Java code:

```
11 import org.apache.mahout.cf.taste.impl.similarity.LogLikelihoodSimilarity;
12 import org.apache.mahout.cf.taste.model.DataModel;
13 import org.apache.mahout.cf.taste.recommender.RecommendedItem;
14 import org.apache.mahout.cf.taste.similarity.ItemSimilarity;
15
16 public class itemRecommend {
17
18     public static void main(String[] args) {
19         try{
20             DataModel dm = new FileDataModel(new File("data/ratings_converted.csv"));
21
22             ItemSimilarity sim = new LogLikelihoodSimilarity(dm);
23
24             GenericItemBasedRecommender recommender = new GenericItemBasedRecommender(dm,sim);
25
26             int x = 1;
27             for (LongPrimitiveIterator items = dm.getItemIDs(); items.hasNext()){
28                 long itemId = items.nextLong();
29                 List<RecommendedItem>recommendations = recommender.mostSimilarItems(itemId,5);
30
31                 for (RecommendedItem recommendation : recommendations){
32                     System.out.println(itemId + "," + recommendation.getItemId() + "," + recommendation.getValue());
33
34                 }
35                 x++;
36             //if (x>10) System.exit(1);
37         }
38     }catch (IOException e){
39         System.out.println("There was an error.");
40         e.printStackTrace();
41     } catch (TasteException e){
42         System.out.println("There was a Taste Exception");
43         e.printStackTrace();
44     }
45
46 }
47 }
```

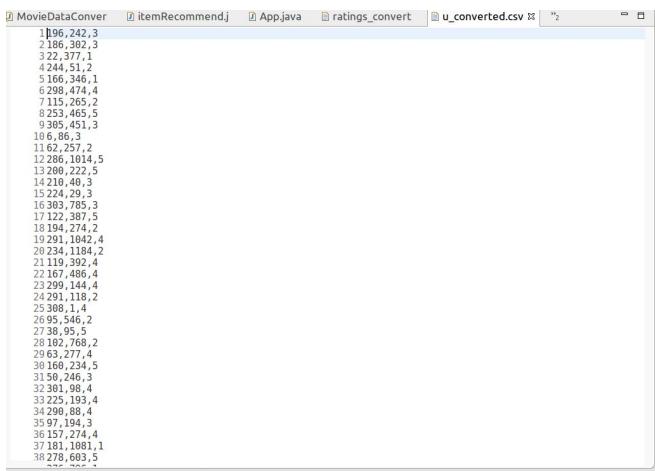
4. Recommender Result



```
<terminated> itemRecommend [Java Application] /usr/lib/jvm/java-8-oracle/bin/java (Oct 24, 2016, 10:55:41 AM)
3951,3911,0.99173236
3951,3949,0.99075305
3951,3897,0.9892427
3951,3952,0.9887318
3952,3897,0.99869294
3952,3317,0.99828416
3952,3893,0.9981959
3952,3911,0.9981946
3952,3948,0.99790674
```

1-2-2 Data Used: <http://grouplens.org/datasets/movielens/> 100K movie data set u.data

1. Convert data



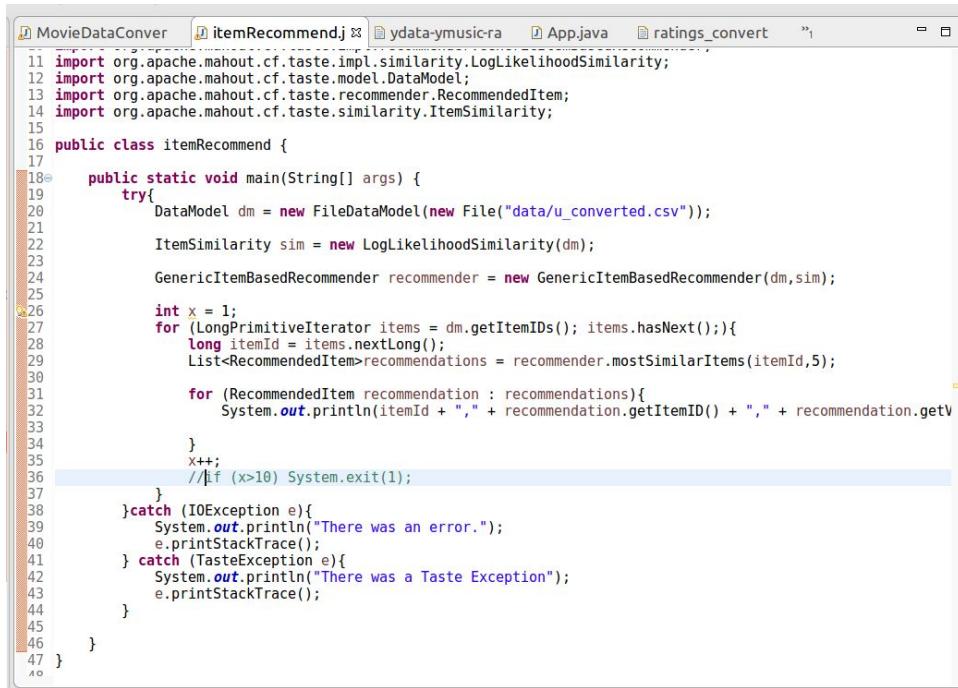
```
1 96, 242, 3
2 186, 362, 3
3 119, 1, 1
4 244, 51, 2
5 166, 346, 1
6 298, 474, 4
7 115, 265, 2
8 235, 265, 5
9 395, 451, 3
10 86, 3
11 62, 257, 2
12 286, 1014, 5
13 30, 222, 5
14 210, 3
15 224, 29, 3
16 303, 785, 3
17 122, 387, 5
18 194, 274, 2
19 234, 1184, 4
20 234, 1184, 2
21 119, 392, 4
22 167, 486, 4
23 299, 144, 4
24 234, 118, 2
25 380, 4
26 95, 546, 2
27 38, 95, 5
28 162, 768, 2
29 60, 74, 4
30 160, 234, 5
31 58, 249, 3
32 381, 99, 4
33 225, 193, 4
34 299, 89, 4
35 234, 118, 4
36 157, 274, 4
37 181, 1081, 1
38 278, 603, 5
```

2. Code to convert data



```
1 package com.predictionmarketing.convert;
2
3 import java.io.BufferedReader;
4 import java.io.BufferedWriter;
5 import java.io.FileNotFoundException;
6 import java.io.FileReader;
7 import java.io.FileWriter;
8 import java.io.IOException;
9
10 public class MovieDataConvert {
11
12     public static void main(String[] args) throws IOException {
13
14         BufferedReader br = new BufferedReader(new FileReader("data/u.data"));
15         BufferedWriter bw = new BufferedWriter(new FileWriter("data/u_converted.csv"));
16
17         String line;
18         while((line = br.readLine()) != null){
19             String[] values = line.split("\t",1);
20             bw.write(values[0] + "," + values[1] + "," + values[2] + "\n");
21         }
22         br.close();
23         bw.close();
24     }
25
26 }
```

3. Build Recommender



```
MovieDataConver itemRecommendj ydata-ymusic-ra App.java ratings_convert
11 import org.apache.mahout.cf.taste.impl.similarity.LogLikelihoodSimilarity;
12 import org.apache.mahout.cf.taste.model.DataModel;
13 import org.apache.mahout.cf.taste.recommender.RecommendedItem;
14 import org.apache.mahout.cf.taste.similarity.ItemSimilarity;
15
16 public class itemRecommend {
17
18     public static void main(String[] args) {
19         try{
20             DataModel dm = new FileDataModel(new File("data/u_converted.csv"));
21
22             ItemSimilarity sim = new LogLikelihoodSimilarity(dm);
23
24             GenericItemBasedRecommender recommender = new GenericItemBasedRecommender(dm,sim);
25
26             int x = 1;
27             for (LongPrimitiveIterator items = dm.getItemIDs(); items.hasNext();){
28                 long itemId = items.nextLong();
29                 List<RecommendedItem>recommendations = recommender.mostSimilarItems(itemId,5);
30
31                 for (RecommendedItem recommendation : recommendations){
32                     System.out.println(itemId + "," + recommendation.getItemId() + "," + recommendation.get
33                 }
34                 x++;
35             }
36             //if (x>10) System.exit(1);
37         }
38     }catch (IOException e){
39         System.out.println("There was an error.");
40         e.printStackTrace();
41     } catch (TasteException e){
42         System.out.println("There was a Taste Exception");
43         e.printStackTrace();
44     }
45
46 }
47 }
```

4. Recommender Results



```
Problems @ Javadoc Declaration Console
terminated: itemRecommend [Java Application] /usr/lib/jvm/java-8-oracle/bin/java (Oct 24, 2016, 11:04:21 AM)
1681,1351,0.9223481
1681,1622,0.9223481
1681,1423,0.918022
1681,1406,0.9144791
1682,1597,0.9144791
1682,1268,0.9019167
1682,1335,0.89995074
1682,767,0.89995074
1682,1428,0.8980945
```

2. Clustering with Mahout

2.1 Reuters clustering

Data used : <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.tar.gz>
In this case, I used cluster-reuter.sh in example and ran fuzzykmeans on data.

2.1.1 Prepare files

```
curl http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.tar.gz -o
${WORK_DIR}/reuters21578.tar.gz
mkdir -p ${WORK_DIR}/reuters-sgm
tar xzf ${WORK_DIR}/reuters21578.tar.gz -C ${WORK_DIR}/reuters-sgm
mkdir -p ${WORK_DIR}/reuters-out
```

```
./mahout org.apache.lucene.benchmark.utils.ExtractReuters ${WORK_DIR}/reuters-sgm
${WORK_DIR}/reuters-out
```

```
nan@nan-ThundeRobot:~$ /usr/local/lib/mahout/bin$ ./mahout org.apache.lucene.benchmark.utils.ExtractReuters ${WORK_DIR}/reuters-sgm ${WORK_DIR}/reuters-out
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT_LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/25 20:36:28 WARN MahoutDriver: No org.apache.lucene.benchmark.utils.ExtractReuters.props found on classpath, will use command-line arguments only
Deleting all files in /tmp/mahout-work-nan/reuters-out-tmp
16/10/25 20:36:30 INFO MahoutDriver: Program took 1616 ms (Minutes: 0.02693333333333333)
```

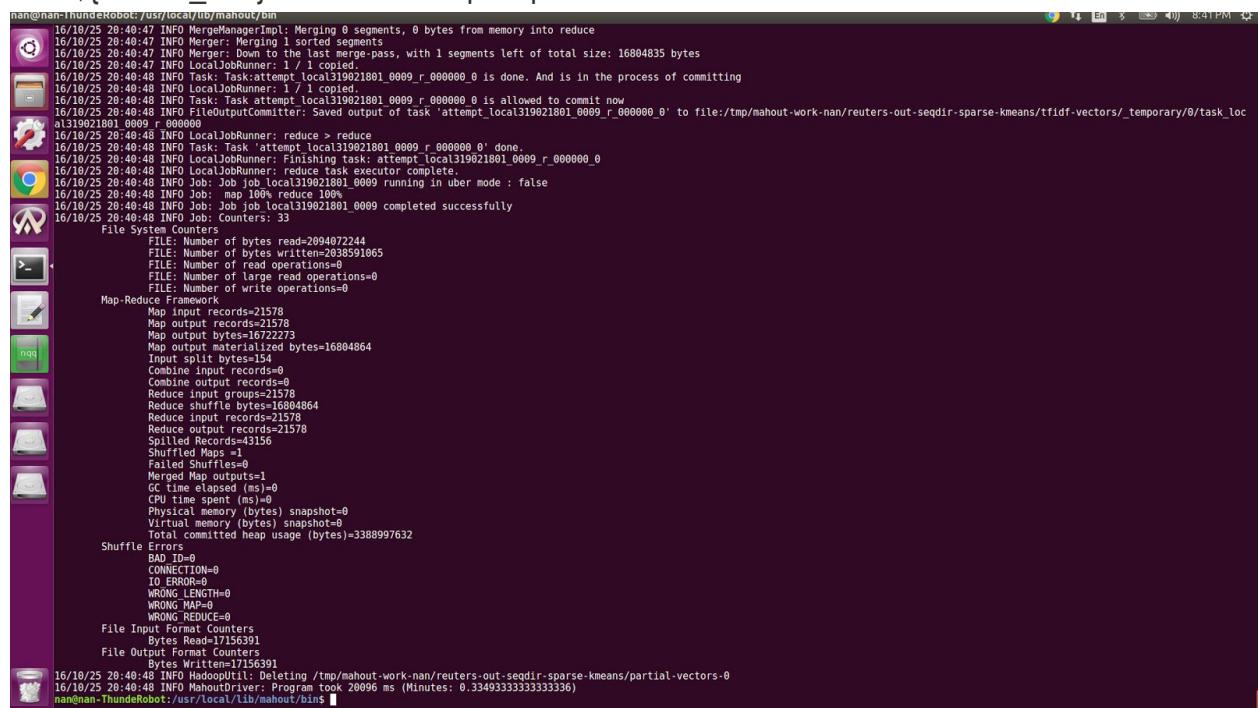
2.1.2 seqdirectory

```
./mahout seqdirectory -i ${WORK_DIR}/reuters-out -o ${WORK_DIR}/reuters-out-seqdir -c
UTF-8 -chunk 64 -xm sequential
```

```
nan@nan-ThundeRobot:~$ /usr/local/lib/mahout/bin$ ./mahout seqdirectory -i ${WORK_DIR}/reuters-out -o ${WORK_DIR}/reuters-out-seqdir -c UTF-8 -chunk 64 -xm sequential
SHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
SHOUT LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/25 20:39:19 INFO AbstractJob: Command line arguments: --charsetName=UTF-8, --endPhase=[2147483647], --fileFilterClass=[org.apache.mahout.text.PrefixAdditionFilter], --input=[/tmp/m
out-work-nan/reuters-out], --keyPrefix[], --method=[sequential], --output=[/tmp/mahout-work-nan/reuters-out-seqdir], --startPhase=[0], --tempDir=[/tmp]
16/10/25 20:39:19 WARN NativeCodeLoader: Unable to load native hadoop library for your platform.. using builtin-Java classes where applicable
16/10/25 20:39:19 INFO NativeCodeLoader: Program loaded in 2268 ms (Minutes: 0.0378)
nan@nan-ThundeRobot:~$ /usr/local/lib/mahout/bin$ █
```

2.1.3 seq2sparse

```
./mahout seq2sparse \
-i ${WORK_DIR}/reuters-out-seqdir \
-o ${WORK_DIR}/reuters-out-seqdir-sparse-kmeans --maxDFPercent 85 --namedVector
```

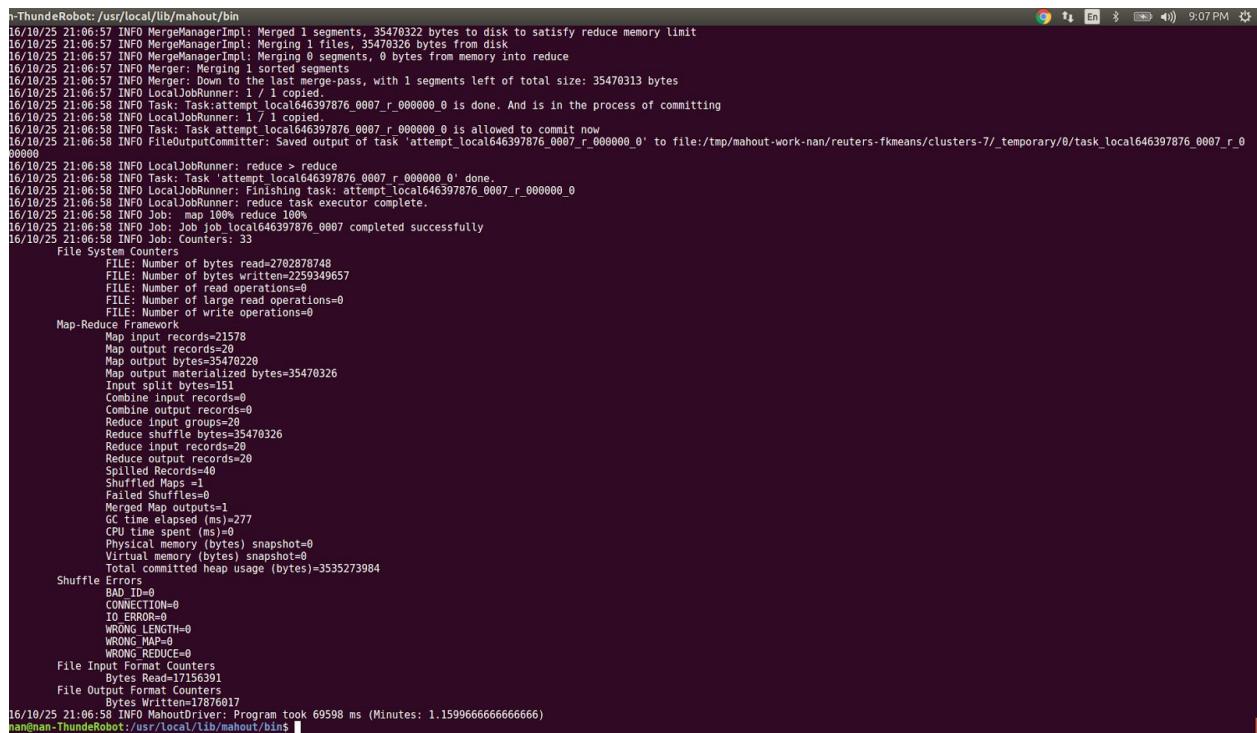


The screenshot shows a terminal window with the following log output:

```
16/10/25 20:40:47 INFO MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/10/25 20:40:47 INFO Merger: Merging 1 sorted segments
16/10/25 20:40:47 INFO Merger: Down to the last merge-pass, with 1 segments left of total size: 16804835 bytes
16/10/25 20:40:47 INFO LocalJobRunner: 1 / 1 copied.
16/10/25 20:40:48 INFO Task: Task-attempt_local319021801_0009_r_000000_0 is done. And is in the process of committing
16/10/25 20:40:48 INFO LocalJobRunner: 1 / 1 copied.
16/10/25 20:40:48 INFO TaskAttempt: Task attempt_local319021801_0009_r_000000_0 is allowed to commit now
16/10/25 20:40:48 INFO FileOutputCommitter: Saved output of task 'attempt_local319021801_0009_r_000000_0' to file:/tmp/mahout-work-nan/reuters-out-seqdir-sparse-kmeans/tfidf-vectors/_temporary/0/task_loc
al319021801_0009_r_000000
16/10/25 20:40:48 INFO LocalJobRunner: reduce > reduce
16/10/25 20:40:48 INFO Task: Task attempt_local319021801_0009_r_000000_0' done.
16/10/25 20:40:48 INFO LocalJobRunner: Flushing task: attempt_local319021801_0009_r_000000_0
16/10/25 20:40:48 INFO LocalJobRunner: Job job local319021801_0009 running in uber mode : false
16/10/25 20:40:48 INFO Job: map 100% reduce 100%
16/10/25 20:40:48 INFO Job: Job job local319021801_0009 completed successfully
16/10/25 20:40:48 INFO Job: Counters: 33
  File System Counters
    File output records=21578
    FILE: Number of bytes written=238591065
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=21578
    Map output records=21578
    Map output bytes=16722273
    Map output materialized bytes=16804864
    Input split bytes=154
    Combine input records=0
    Combine output records=0
    Reduce input groups=21578
    Reduce shuffle bytes=16804864
    Reduce input records=21578
    Reduce output records=21578
    Spilled Records=43196
    Shuffled Maps=21578
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=0
    CPU time spent (ms)=0
    Physical memory (bytes) snapshot=0
    Virtual memory (bytes) snapshot=0
    Total committed heap usage (bytes)=3388997632
  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=17156391
  File Output Format Counters
    Bytes Written=17156391
16/10/25 20:40:48 INFO HadoopUtil: Deleting /tmp/mahout-work-nan/reuters-out-seqdir-sparse-kmeans/partial-vectors-0
16/10/25 20:40:48 INFO MahoutDriver: Program took 20096 ms (Minutes: 0.3349333333333336)
nan@nan-ThundeRobot:~$ /usr/local/lib/mahout/bin$ █
```

2.1.4 fkmeans

```
./mahout fkmeans \
-i ${WORK_DIR}/reuters-out-seqdir-sparse-fkmeans/tfidf-vectors/ \
-c ${WORK_DIR}/reuters-fkmeans-clusters \
-o ${WORK_DIR}/reuters-fkmeans \
-dm org.apache.mahout.common.distance.EuclideanDistanceMeasure \
-x 10 -k 20 -ow -m 1.1
```



The screenshot shows a terminal window titled "ThunderRobot: /usr/local/lib/mahout/bin". The log output is as follows:

```
16/10/25 21:06:57 INFO MergeManagerImpl: Merged 1 segments, 35470322 bytes to disk to satisfy reduce memory limit
16/10/25 21:06:57 INFO MergeManagerImpl: Merging 1 files, 35470326 bytes from disk
16/10/25 21:06:57 INFO MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/10/25 21:06:57 INFO Merger: Merging 1 sorted segments
16/10/25 21:06:57 INFO Merger: Down to the last merge-pass, with 1 segments left of total size: 35470313 bytes
16/10/25 21:06:57 INFO LocalJobRunner: 1 / 1 copied
16/10/25 21:06:57 INFO FileOutputCommitter: local646397876_0007_r_000000_0 is done. And is in the process of committing
16/10/25 21:06:58 INFO LocalJobRunner: 1 / 1 copied
16/10/25 21:06:58 INFO Task: Task attempt local646397876_0007_r_000000_0 is allowed to commit now
16/10/25 21:06:58 INFO FileOutputCommitter: Saved output of task 'attempt_local646397876_0007_r_000000_0' to file:/tmp/mahout-work-nan/reuters-fkmeans/clusters-7/_temporary/0/task_local646397876_0007_r_0_00000
16/10/25 21:06:58 INFO LocalJobRunner: reduce > reduce
16/10/25 21:06:58 INFO Task: Task attempt local646397876_0007_r_000000_0 done.
16/10/25 21:06:58 INFO LocalJobRunner: Finishing task: attempt local646397876_0007_r_000000_0
16/10/25 21:06:58 INFO LocalJobRunner: The task executor complete.
16/10/25 21:06:58 INFO Job: map 100% reduce 100%
16/10/25 21:06:58 INFO Job: Job local646397876_0007 completed successfully
16/10/25 21:06:58 INFO Job: Counters: 33
  File System Counters
    FILE: Number of bytes read=2702878748
    FILE: Number of bytes written=259349657
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=21578
    Map output records=20
    Map output bytes=35470220
    Map output spilt bytes=35470326
    Input split bytes=151
    Combine input records=0
    Combine output records=0
    Reduce input groups=20
    Reduce shuffle bytes=35470326
    Reduce input records=20
    Reduce output records=20
    Spilled Records=0
    Shuffled Maps=1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=277
    CPU time spent (ms)=0
    Physical memory (bytes) snapshot=0
    Virtual memory (bytes) snapshot=0
    Total committed heap usage (bytes)=3535273984
  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=17156391
  File Output Format Counters
    Bytes Written=17876017
16/10/25 21:06:58 INFO MahoutDriver: Program took 69598 ms (Minutes: 1.1599666666666666)
nan@nan-ThunderRobot:/usr/local/lib/mahout/bin$
```

2.1.5 clusterdump and show data

```
./mahout clusterdump \
-i ${WORK_DIR}/reuters-fkmeans/clusters-*-final \
-o ${WORK_DIR}/reuters-fkmeans/clusterdump \
-d ${WORK_DIR}/reuters-out-seqdir-sparse-fkmeans/dictionary.file-0 \
-dt sequencefile -b 100 -n 20 -sp 0 \
&& \
```

```
cat ${WORK_DIR}/reuters-fkmeans/clusterdump
```

```
nan@nan-ThundeRobot: /usr/local/lib/mahout/bin
mar >> 0.8722887544517766
year >> 0.8720021865203545
has >> 0.7944532516661233
he >> 0.7934456481661056
is >> 0.7934456481661056
billion >> 0.7341027472479633
cts >> 0.722689174970288
would >> 0.717880331727681
inc >> 0.711367182936889
company >> 0.70754739995950376
which >> 0.6806368592056907
net >> 0.655223228972878
apr >> 0.6514752639465702
:{"identifier":"SV-13854","r": [{"":0":0.642}, {"":0.003":0.097}, {"":0.006913":0.097}, {"":0.007050":0.098}, {"":0.0
Top Terms:
said >> 1.705329682127908
mln >> 1.2636185368044586
dlrs >> 1.141607966674459
pct >> 1.0481647161294545
from >> 0.995669182934359
its >> 0.995669182934359
year >> 0.8940993082407854
mar >> 0.8792651075587474
vs >> 0.8497679013210386
has >> 0.8202809276529438
he >> 0.7824557240565266
u.s >> 0.7624981912768823
billion >> 0.756410677231527
world >> 0.75206470002503
company >> 0.726090189341719
inc >> 0.717323694751672
which >> 0.705737105902384
cts >> 0.6999531044656514
corp >> 0.6595230293727834
new >> 0.6574043387008863
:{"identifier":"SV-3557","r": [{"":0":0.616}, {"":0.003":0.093}, {"":0.006913":0.092}, {"":0.007050":0.092}, {"":0.0
Top Terms:
said >> 1.5880793700612341
mln >> 1.297578976621919
dlrs >> 1.078863534816195
pct >> 0.9720932162355935
from >> 0.9203449506121562
its >> 0.9202879945854752
vs >> 0.897861512981982
mar >> 0.897861512981982
year >> 0.8355749244721761
has >> 0.7567840334860426
cts >> 0.741372589150705
he >> 0.7084659669390404
u.s >> 0.7039426326258195
billion >> 0.7008303918880217
inc >> 0.6972258143534369
company >> 0.6972258143534369
world >> 0.6774699970369502
net >> 0.6547624615861033
which >> 0.6446871799194829
apr >> 0.6359738384145118
```

2.2 Wikipedia cluster

2.2.1 Build working directories

```
export WORK_DIR=/tmp/mahout-work-nan
hdfs dfs -mkdir -p $WORK_DIR
mkdir -p $WORK_DIR
mkdir -p ${WORK_DIR}/wiki-sgm
mkdir -p ${WORK_DIR}/wiki-out
cp ~/Downloads/enwiki-articles.xml ${WORK_DIR}/wiki-sgm/
cp ~/Downloads/enwiki-articles.xml ${WORK_DIR}/wiki-out/
cp ~/Downloads/categories.txt ${WORK_DIR}
hdfs dfs -mkdir -p ${WORK_DIR}/
hdfs dfs -mkdir ${WORK_DIR}/wiki-sgm
hdfs dfs -mkdir ${WORK_DIR}/wiki-out
hdfs dfs -put ${WORK_DIR}/wiki-sgm ${WORK_DIR}/wiki-sgm
hdfs dfs -put ${WORK_DIR}/wiki-out ${WORK_DIR}/wiki-out
```

```

corp ~ 0.0321729207223004
nan@nan-ThundeRobot:/usr/local/lib/mahout/examples/bin$ cd ${WORK_DIR}
nan@nan-ThundeRobot:$ export WORK_DIR=/tmp/mahout-work-nan
nan@nan-ThundeRobot:$ cd ${WORK_DIR}
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ mkdir -p ${WORK_DIR}/wiki-sgm
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -mkdir -p $WORK_DIR
16/10/24 22:13:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ cp ~/Downloads/enwiki-articles.xml wiki-sgm
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -mkdir ${WORK_DIR}/wiki-sgm
16/10/24 22:14:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -mkdir ${WORK_DIR}/wiki-out
16/10/24 22:14:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -put ${WORK_DIR}/wiki-sgm ${WORK_DIR}/wiki-sgm
16/10/24 22:16:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -put ${WORK_DIR}/wiki-sgm ${WORK_DIR}/wiki-out
16/10/24 22:16:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
put: '/tmp/mahout-work-nan/wiki-out': No such file or directory
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -mkdir ${WORK_DIR}/wiki-out
Error: Could not find or load main class dfs-mkdir
16/10/24 22:16:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
mkdir: '/tmp/mahout-work-nan/wiki-out': File exists
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -put ${WORK_DIR}/wiki-out ${WORK_DIR}/wiki-out
16/10/24 22:16:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
put: '/tmp/mahout-work-nan/wiki-out': No such file or directory
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ mkdir -p ${WORK_DIR}/wiki-out
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -put ${WORK_DIR}/wiki-out ${WORK_DIR}/wiki-out
16/10/24 22:17:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ 

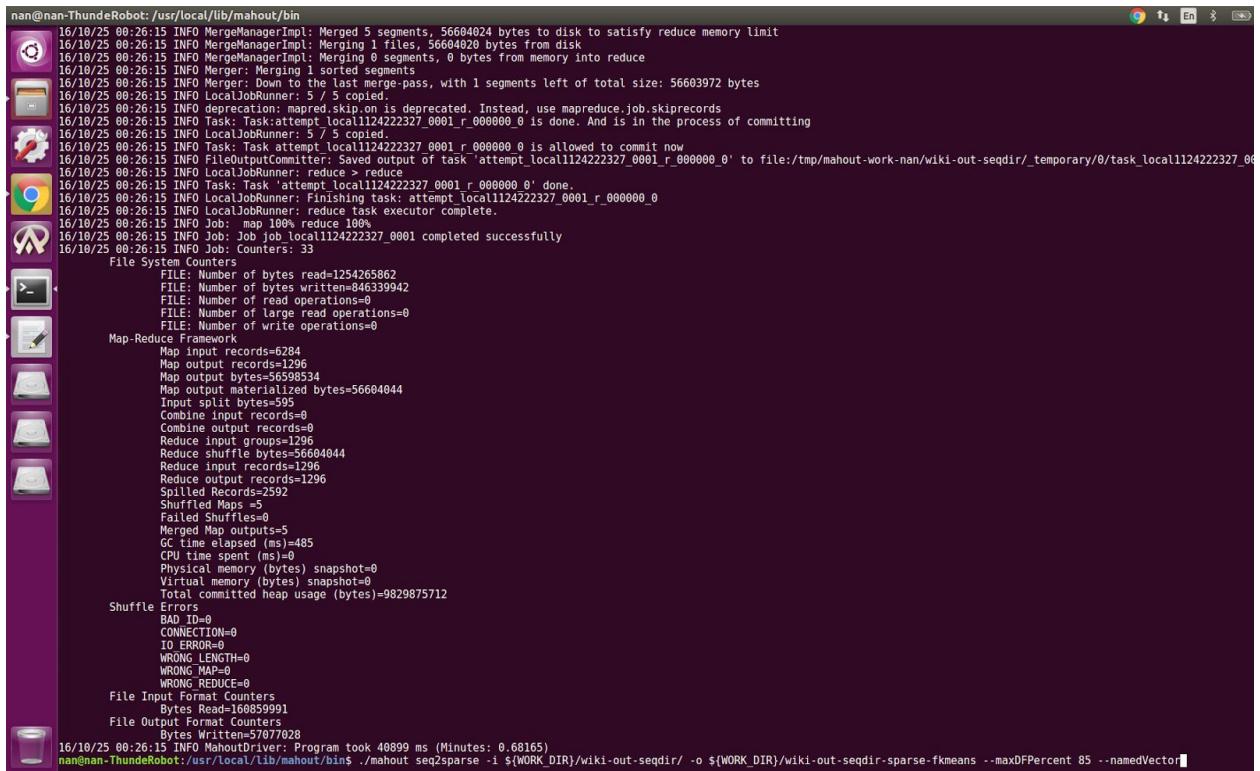
```

2.2.2 seqwiki

```

./mahout seqwiki -c ${WORK_DIR}/categories.txt -i ${WORK_DIR}/wiki-out/enwiki-articles.xml
-o ${WORK_DIR}/wiki-out-seqdir

```



```

nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$ ./mahout seqwiki -c ${WORK_DIR}/categories.txt -i ${WORK_DIR}/wiki-out/enwiki-articles.xml
-o ${WORK_DIR}/wiki-out-seqdir
16/10/25 00:26:15 INFO MergeManagerImpl: Merged 5 segments, 56604024 bytes to disk to satisfy reduce memory limit
16/10/25 00:26:15 INFO MergeManagerImpl: Merging 1 files, 56604020 bytes from disk
16/10/25 00:26:15 INFO MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/10/25 00:26:15 INFO Merger: Merging 1 sorted segments
16/10/25 00:26:15 INFO Merger: Down to the last merge-pass, with 1 segments left of total size: 56603972 bytes
16/10/25 00:26:15 INFO LocalJobRunner: 5 / 5 copied.
16/10/25 00:26:15 INFO Task: Task-attempt_local1124222327_0001_r_000000_0 is done. And is in the process of committing
16/10/25 00:26:15 INFO LocalJobRunner: 5 / 5 copied.
16/10/25 00:26:15 INFO Task: Task-attempt_local1124222327_0001_r_000000_0 is allowed to commit now
16/10/25 00:26:15 INFO LocalJobRunner: Reducing task input for task 'attempt_local1124222327_0001_r_000000_0' to file:/tmp/mahout-work-nan/wiki-out-seqdir/_temporary/0/task_local1124222327_0001_r_000000_0
16/10/25 00:26:15 INFO LocalJobRunner: reduce > reduce
16/10/25 00:26:15 INFO LocalJobRunner: Task 'attempt_local1124222327_0001_r_000000_0' done.
16/10/25 00:26:15 INFO LocalJobRunner: Finishing task: attempt_local1124222327_0001_r_000000_0
16/10/25 00:26:15 INFO LocalJobRunner: reduce task executor complete.
16/10/25 00:26:15 INFO Job: map 100% reduce 100%
16/10/25 00:26:15 INFO Job: Job job_local1124222327_0001 completed successfully
16/10/25 00:26:15 INFO Job: Counters: 33
File System Counters
FILE: Number of bytes read=1254265862
FILE: Number of bytes written=846339942
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map input records=6284
Map output records=1296
Map output bytes=6598534
Map output materialized bytes=56604044
Input split bytes=595
Combine input records=8
Combine output records=8
Reduce input bytes=106
Reduce shuffle bytes=56604044
Reduce input records=1296
Reduce output records=1296
Spilled Records=2592
Shuffled Maps=5
Failed Shuffles=0
Merged Map outputs=5
GC time elapsed (ms)=485
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=9829875712
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=6284000
File Output Format Counters
Bytes Written=57077028
16/10/25 00:26:15 INFO MahoutDriver: Program took 40899 ms (Minutes: 0.68165)
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$ ./mahout seq2sparse -i ${WORK_DIR}/wiki-out-seqdir/ -o ${WORK_DIR}/wiki-out-seqdir-sparse-fkmeans --maxDFPercent 85 --namedVector

```

2.2.3 seq2sparse

```

./mahout seq2sparse -i ${WORK_DIR}/wiki-out-seqdir/ -o
${WORK_DIR}/wiki-out-seqdir-sparse-fkmeans --maxDFPercent 85 --namedVector

```

```

nan@nan-ThundeRobot:/usr/local/lib/mahout/bin
16/10/25 00:30:18 INFO MergeManagerImpl: Merging 1 files, 128174886 bytes from disk
16/10/25 00:30:18 INFO MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/10/25 00:30:18 INFO Merger: Merging 1 sorted segments
16/10/25 00:30:18 INFO LocalJobRunner: Down to the last merge-pass, with 1 segments left of total size: 128174793 bytes
16/10/25 00:30:18 INFO Job: map 10% reduce 0%
16/10/25 00:30:20 INFO Task: Task attempt local1890104047_0010_r_000000_0 is done. And is in the process of committing
16/10/25 00:30:20 INFO LocalJobRunner: 1 / 1 copied.
16/10/25 00:30:20 INFO Task: Task attempt local1890104047_0010_r_000000_0 is allowed to commit now
16/10/25 00:30:20 INFO fileOutputCommitter: Saved output of task 'attempt_local1890104047_0010_r_000000_0' to file:/tmp/mahout-work-nan/wiki-fkmeans/clusters-10/_temporary/0/task_local1890104047_0010_r_0_000000
16/10/25 00:30:20 INFO LocalJobRunner: Job: Counters: 33
    File System Counters
        FILE: Number of bytes read=11339520330
        FILE: Number of bytes written=908931583
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
    Map-Reduce Framework
        Map input records=1296
        Map output records=28
        Map output bytes=128174700
        Map output materialized bytes=128174886
        Input split bytes=148
        Combine input records=0
        Combine output records=0
        Reduce input groups=20
        Reduce input bytes=128174886
        Reduce input records=29
        Reduce output records=29
        Spilled Records=68
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=472
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
        Total committed heap usage (bytes)=3834118144
    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=21966018
    File Output Format Counters
        Bytes Written=64590385
16/10/25 00:30:21 INFO MahoutDriver: Program took 160620 ms (Minutes: 2.677)
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$
```

2.2.4 fkmeans

```

./mahout fkmeans -i ${WORK_DIR}/wiki-out-seqdir-sparse-fkmeans/tfidf-vectors/ -c
${WORK_DIR}/wiki-fkmeans-clusters -o ${WORK_DIR}/wiki-fkmeans -dm
org.apache.mahout.common.distance.EuclideanDistanceMeasure -x 10 -k 20 -ow -m 1.1

```

```

nan@nan-ThundeRobot:/usr/local/lib/mahout/bin
16/10/25 00:30:16 INFO MergeManagerImpl: Merging 1 files, 128174886 bytes from disk
16/10/25 00:30:18 INFO MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/10/25 00:30:18 INFO Merger: Merging 1 sorted segments
16/10/25 00:30:18 INFO LocalJobRunner: Down to the last merge-pass, with 1 segments left of total size: 128174793 bytes
16/10/25 00:30:18 INFO Job: map 10% reduce 0%
16/10/25 00:30:20 INFO Task: Task attempt local1890104047_0010_r_000000_0 is done. And is in the process of committing
16/10/25 00:30:20 INFO LocalJobRunner: 1 / 1 copied.
16/10/25 00:30:20 INFO Task: Task attempt local1890104047_0010_r_000000_0 is allowed to commit now
16/10/25 00:30:20 INFO fileOutputCommitter: Saved output of task 'attempt_local1890104047_0010_r_000000_0' to file:/tmp/mahout-work-nan/wiki-fkmeans/clusters-10/_temporary/0/task_local1890104047_0010_r_0_000000
16/10/25 00:30:20 INFO LocalJobRunner: Job: Counters: 33
    File System Counters
        FILE: Number of bytes read=11339520330
        FILE: Number of bytes written=908931583
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
    Map-Reduce Framework
        Map input records=1296
        Map output records=28
        Map output bytes=128174700
        Map output materialized bytes=128174886
        Input split bytes=148
        Combine input records=0
        Combine output records=0
        Reduce input groups=20
        Reduce input records=29
        Reduce shuffle bytes=128174886
        Reduce input records=29
        Reduce output records=29
        Spilled Records=68
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=472
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
        Total committed heap usage (bytes)=3834118144
    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=21966018
    File Output Format Counters
        Bytes Written=64590385
16/10/25 00:30:21 INFO MahoutDriver: Program took 160620 ms (Minutes: 2.677)
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$
```

2.2.5 show data

```
./mahout clusterdump \
-i ${WORK_DIR}/wiki-fkmeans/clusters-*-final \
-o ${WORK_DIR}/wiki-fkmeans/clusterdump \
-d ${WORK_DIR}/wiki-out-seqdir-sparse-fkmeans/dictionary.file-0 \
-dt sequencefile -b 100 -n 20 -sp 0 \
&& \
cat ${WORK_DIR}/wiki-fkmeans/clusterdump
```

3. Classification with Mahout

```
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=2
    CPU time spent (ms)=0
    Physical memory (bytes) snapshot=0
    Virtual memory (bytes) snapshot=0
    Total committed heap usage (bytes)=4767875072
    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=122
    File Output Format Counters
        Bytes Written=102
16/10/24 22:34:03 INFO HadoopUtil: Deleting /tmp/mahout-work-nan/wiki-out-seqdir
-sparse-fkmeans/partial-vectors-0
16/10/24 22:34:03 INFO MahoutDriver: Program took 97289 ms (Minutes: 1.621483333
3333334)
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$
```

3-1 Classification on 20 newsgroups data

I used the pre-implemented algorithms from \$MAHOUT_HOME/examples/bin/classify-20newsgroups.sh for this part. The only difference between 3-1 and 3-2 is the data difference. Here I just paste 3-1 result for simplification.

3.1.1 Prepare data

```
company          => 0.6784786878366332
would          => 0.6774699920369592
net              => 0.6547624615861833
which          => 0.6446871769194829
apr             => 0.6359738384145118
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$ export WORK_DIR=/tmp/mahout-work-nan
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$ mkdir -p ${WORK_DIR}
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$ mkdir -p ${WORK_DIR}/20news-all
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$ cd ${WORK_DIR}/20news-all
nan@nan-ThundeRobot:/tmp/mahout-work-nan/20news-all$ curl http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz -o ${WORK_DIR}/20news-bydate.tar.gz
% Total % Received % Xferd Average Speed Time Time Current
          Upload Total Spent Left Speed
100 13.7M 100 13.7M 0 0 2449k 0 0:00:05 0:00:05 ...:---- 2806K
nan@nan-ThundeRobot:/tmp/mahout-work-nan/20news-all$ mkdir -p ${WORK_DIR}/20news-bydate
nan@nan-ThundeRobot:/tmp/mahout-work-nan/20news-all$ cd ${WORK_DIR}/20news-bydate && tar xzf ../20news-bydate.tar.gz && cd .. && cd ..
nan@nan-ThundeRobot:/tmp$ cp -R ${WORK_DIR}/20news-bydate/* ${WORK_DIR}/20news-all
```

3.1.2 seqdirectory

```
./bin/mahout seqdirectory \
-i ${WORK_DIR}/20news-all \
```

```
-o ${WORK_DIR}/20news-seq -ow
```

```
nan@nan-ThundeRobot: /usr/local/lib/mahout
[s-all/sci.electronics/53729:0+331,/tmp/mahout-work-nan/20news-all/sci.electronics/53525:0+933,/tmp/mahout-work-nan/20news-all/sci.electronics/53924:0+495,/tmp/mahout-work-nan/20news-all/sci.electronics/52796:0+258,/tmp/mahout-work-nan/20news-all/sci.electronics/54093:0+675,/tmp/mahout-work-nan/20news-all/sci.electronics/53968:0+1920,/tmp/mahout-work-nan/20news-all/sci.electronics/53548:0+875,/tmp/mahout-work-nan/20news-all/sci.electronics/53720:0+839,/tmp/mahout-work-nan/20news-all/sci.electronics/53703:0+1673,/tmp/mahout-work-nan/20news-all/sci.electronics/53925:0+1495,/tmp/mahout-work-nan/20news-all/sci.electronics/53721:0+684,/tmp/mahout-work-nan/20news-all/sci.electronics/53896:0+806,/tmp/mahout-work-nan/20news-all/sci.electronics/54278:0+433,/tmp/mahout-work-nan/20news-all/sci.electronics/54093:0+675,/tmp/mahout-work-nan/20news-all/sci.electronics/53615:0+1256,/tmp/mahout-work-nan/20news-all/sci.electronics/54032:0+322,/tmp/mahout-work-nan/20news-all/sci.electronics/53935:0+1294,/tmp/mahout-work-nan/20news-all/sci.electronics/54347:0+1600,/tmp/mahout-work-nan/20news-all/sci.electronics/53510:0+1274,/tmp/mahout-work-nan/20news-all/sci.electronics/53788:0+972,/tmp/mahout-work-nan/20news-all/sci.electronics/54332:0+392,/tmp/mahout-work-nan/20news-all/sci.electronics/53779:0+1145,/tmp/mahout-work-nan/20news-all/sci.electronics/53776:0+1908,/tmp/mahout-work-nan/20news-all/sci.electronics/54021:0+485,/tmp/mahout-work-nan/20news-all/sci.electronics/54249:0+656,/tmp/mahout-work-nan/20news-all/sci.electronics/53746:0+314
16/10/25 21:17:59 INFO CodecPool: Got brand-new compressor [.deflate]
16/10/25 21:18:00 INFO Job: Job job_local698623683_0001 running in uber mode : false
16/10/25 21:18:00 INFO Job: map 0% reduce 0%
16/10/25 21:18:03 INFO LocalJobRunner:
16/10/25 21:18:03 INFO Task: Task attempt_local698623683_0001_m_000000_0 is done. And is in the process of committing
16/10/25 21:18:03 INFO LocalJobRunner:
16/10/25 21:18:03 INFO Task: Task attempt_local698623683_0001_m_000000_0 is allowed to commit now
16/10/25 21:18:03 INFO FileOutputCommitter: Saved output of task 'attempt_local698623683_0001_m_000000_0' to file:/tmp/mahout/work/nan/20news-all/sci.electronics/53746:0+314
16/10/25 21:18:03 INFO LocalJobRunner: map
16/10/25 21:18:03 INFO Task: Task 'attempt_local698623683_0001_m_000000_0' done.
16/10/25 21:18:03 INFO LocalJobRunner: Finishing task: attempt_local698623683_0001_m_000000_0
16/10/25 21:18:03 INFO LocalJobRunner: map task executor complete.
16/10/25 21:18:04 INFO Job: map 100% reduce 0%
16/10/25 21:18:04 INFO Job: Job job_local698623683_0001 completed successfully
16/10/25 21:18:04 INFO Job: Counters: 18
File System Counters
FILE: Number of bytes read=117876417
FILE: Number of bytes written=102224986
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
Map-Reduce Framework
Map input records=18846
Map output records=18846
Input split bytes=1427950
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=0
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=1604321280
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=19352119
16/10/25 21:18:04 INFO MahoutDriver: Program took 30870 ms (Minutes: 0.5145)
nan@nan-ThundeRobot: /usr/local/lib/mahout$
```

3.1.3 seq2sparse

```
./bin/mahout seq2sparse \
-i ${WORK_DIR}/20news-seq \
-o ${WORK_DIR}/20news-vectors -lnorm -nv -wt tfidf
```

```

Terminal File Edit View Search Terminal Help
16/10/25 21:19:43 INFO Merger: Merging 1 sorted segments
16/10/25 21:19:43 INFO Merger: Down to the last merge-pass, with 1 segments left of total size: 28437723 bytes
16/10/25 21:19:43 INFO LocalJobRunner: 1 / 1 copied.
16/10/25 21:19:43 INFO Job: Job job local1716044990_0009 running in uber mode : false
16/10/25 21:19:43 INFO Job: map 100% reduce 0%
16/10/25 21:19:43 INFO Task: Task:attempt local1716044990_0009_r_000000_0 is done. And is in the process of committing
16/10/25 21:19:43 INFO LocalJobRunner: 1 / 1 copied.
16/10/25 21:19:43 INFO Task: Task attempt local1716044990_0009_r_000000_0 is allowed to commit now
16/10/25 21:19:43 INFO FileOutputCommitter: Saved output of task 'attempt local1716044990_0009_r_000000_0' to file:/tmp/mahout-work-nan/20news-vectors/tfidf-vectors/_temporal/_r_000000
16/10/25 21:19:43 INFO LocalJobRunner: reduce > reduce
16/10/25 21:19:43 INFO Task: Task attempt local1716044990_0009_r_000000_0' done.
16/10/25 21:19:43 INFO LocalJobRunner: Finishing task: attempt_local1716044990_0009_r_000000_0
16/10/25 21:19:43 INFO LocalJobRunner: reduce task executor complete.
16/10/25 21:19:44 INFO Job: map 100% reduce 100%
16/10/25 21:19:44 INFO Job: Job job local1716044990_0009 completed successfully
16/10/25 21:19:44 INFO Job: Counters: 33
    File System Counters
        FILE: Number of bytes read=254175420
        FILE: Number of bytes written=2457394371
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
    Map-Reduce Framework
        Map input records=18846
        Map output records=18846
        Map output bytes=28362505
        Map output materialized bytes=28437750
        Input split bytes=136
        Combine input records=0
        Combine output records=0
        Reduce input groups=18846
        Reduce input records=18846
        Reduce output records=18846
        Spilled Records=37692
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map Outputs=1
        GC time elapsed (ms)=3
        CPU time spent (ms)=0
        Physical memory (bytes) snapshot=0
        Virtual memory (bytes) snapshot=0
        Total committed heap usage (bytes)=3551002624
    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=28913427
    File Output Format Counters
        Bytes Written=28913427
16/10/25 21:19:44 INFO HadoopUtil: Deleting /tmp/mahout-work-nan/20news-vectors/partial-vectors-0
16/10/25 21:19:44 INFO MahoutDriver: Program took 32329 ms (Minutes: 0.5388166666666667)
nan@nan-ThundeRobot:~/usr/local/lib/mahout$ 

```

3.1.4 split

```

./bin/mahout split \
-i ${WORK_DIR}/20news-vectors/tfidf-vectors \
--trainingOutput ${WORK_DIR}/20news-train-vectors \
--testOutput ${WORK_DIR}/20news-test-vectors \
--randomSelectionPct 40 --overwrite --sequenceFiles -xm sequential

```

```

16/10/25 21:20:47 INFO HadoopUtil: Deleting /tmp/mahout-work-nan/20news-vectors/partial-vectors-0 (Minutes: 0.5388166666666667)
nan@nan-ThundeRobot:~/usr/local/lib/mahout$ ./bin/mahout split \
> -i ${WORK_DIR}/20news-vectors/tfidf-vectors \
> -o ${WORK_DIR}/20news-vectors \
> -testOutput ${WORK_DIR}/20news-test-vectors \
> -r randomSelectionPct 40 --overwrite --sequenceFiles -xm sequential
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in /usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/25 21:20:47 WARN MahoutDriver: No split.props found on classpath, will use command-line arguments only
16/10/25 21:20:47 INFO AbstractJob: Command line arguments: {-endPhase=[2147483647], -input=[/tmp/mahout-work-nan/20news-vectors/tfidf-vectors], -method=[sequential], -overwrite=null, -randomSelectonPct=[40], --sequenceFiles=null, -startPhase=[0], -tempDir=[/tmp], -testOutput=[/tmp/mahout-work-nan/20news-test-vectors], -trainingOutput=[/tmp/mahout-work-nan/20news-train-vectors]}
16/10/25 21:20:48 INFO NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
16/10/25 21:20:48 INFO NativeCodeLoader: NativeCodeLoader has loaded native-hadoop
16/10/25 21:20:48 INFO SplitInput: part-r-00000 test split size is 64968 based on random selection percentage 40
16/10/25 21:20:48 INFO CodecPool: Got brand-new compressor [.deflate]
16/10/25 21:20:48 INFO CodecPool: Got brand-new compressor [.deflate]
16/10/25 21:20:51 INFO SplitInput: file: part-r-00000, input: 162419 train: 11277, test: 7569 starting at 0
16/10/25 21:20:51 INFO MahoutDriver: Program took 4292 ms (Minutes: 0.07153333333333334)
nan@nan-ThundeRobot:~/usr/local/lib/mahout$ 

```

3.1.5 train different model and test model and show result

C-NB result

```
./bin/mahout trainnb \
-i ${WORK_DIR}/20news-train-vectors \
-o ${WORK_DIR}/model \
-li ${WORK_DIR}/labelindex \
-ow $c

./bin/mahout testnb \
-i ${WORK_DIR}/20news-train-vectors\
-m ${WORK_DIR}/model \
-I ${WORK_DIR}/labelindex \
-ow -o ${WORK_DIR}/20news-testing $c

./bin/mahout testnb \
-i ${WORK_DIR}/20news-test-vectors\
-m ${WORK_DIR}/model \
-I ${WORK_DIR}/labelindex \
-ow -o ${WORK_DIR}/20news-testing $c
```

```
root@nan-ThundeRobot:/usr/local/lib/mahout# ./mahout
[...]
| 396 b =
| 388 c =
| 375 d =
| 385 e =
| 398 f =
| 371 g =
| 409 h =
| 379 i =
| 406 j =
| 391 k =
| 387 l =
| 363 m =
| 388 n =
| 399 o =
| 418 p =
| 367 q =
| 378 r =
| 332 s =
| 271 t =
=====
Statistics
-----
Kappa                                0.8548
Accuracy                             89.0136%
Reliability                           84.4586%
Reliability (standard deviation)    0.2197
Weighted precision                   0.9314
Weighted recall                      0.8981
Weighted F1 score                    0.8978

16/10/23 21:49:26 INFO MahoutDriver: Program took 5144 ms (Minutes: 0.0857333333333333)
root@nan-ThundeRobot:/usr/local/lib/mahout#
```

NB result

```
./bin/mahout trainnb \
-i ${WORK_DIR}/20news-train-vectors \
-o ${WORK_DIR}/model \
```

```

-li ${WORK_DIR}/labelindex \
-ow $c

./bin/mahout testnb \
-i ${WORK_DIR}/20news-train-vectors\
-m ${WORK_DIR}/model \
-I ${WORK_DIR}/labelindex \
-ow -o ${WORK_DIR}/20news-testing $c

./bin/mahout testnb \
-i ${WORK_DIR}/20news-test-vectors\
-m ${WORK_DIR}/model \
-I ${WORK_DIR}/labelindex \
-ow -o ${WORK_DIR}/20news-testing $c

```

```

Terminal | File | Edit | View | Search | Terminal | Help
0   318  5   12   9   17   9   0   0   1   0   3   6   0   2   0   0   0   0   0
1   18    254  77   22   17   6   1   0   1   0   7   2   0   2   0   1   0   0   0
0   6    4   334  28   1   8   0   0   0   1   0   0   11  0   0   0   0   0   0   0
1   7    2   15   325  1   4   0   0   0   0   0   0   11  0   0   0   0   0   0   0
0   15   3   5    3   356  4   0   0   0   0   0   1   0   1   2   0   0   0   0   0
1   4    1   19   8   0    320  10  2   1   4   0   8   0   4   0   2   0   0   0   2
0   1    0   0    1   0    11   367  5   0   0   0   2   0   2   1   1   0   0   1
rec.autos
0   0    0    1   0    0   3   4   369  0   0   0   1   2   0   0   0   0   0   0
rec.motorcycles
0   0    0    0   0    0   2   2   384  4   0   0   0   0   0   0   0   0   0   0
rec.sport.baseball
0   0    0    0   0    1   2   0   0   5   390  0   0   0   0   0   0   1   0   2
rec.sport.hockey
1   3    0    0   0    3   0   0   0   1   0   367  0   1   0   0   0   2   0   3
sci.crypt
0   4    0    17   9   1   8   5   0   1   0   2   360  1   0   0   0   1   0   1
sci.electronics
0   2    1    1   4   1   0   0   2   0   0   0   5   386  7   0   1   1   1   2
sci.med
1   0    0    0   1   0   1   1   0   0   0   1   0   2   362  0   0   1   0   1
sci.space
3   0    0    1   1   0   2   0   0   1   0   0   1   1   0   405  1   1   0   0
soc.religion.christian
0   0    1   0    0   0   0   0   1   0   0   0   4   0   0   0   1   361  0   11
talk.politics.guns
1   0    0    0   0   0   0   0   0   0   1   0   0   0   0   1   1   3   398  4
talk.politics.mideast
1   0    0    0   2   0   0   0   0   1   0   1   2   1   4   0   0   14  3   304
talk.politics.misc
26  1    0    0   0   1   0   1   1   0   0   0   0   0   0   1   15  9   1   5
talk.religion.misc

=====
Statistics
-----
Kappa          0.8734
Accuracy       98.127%
Reliability    85.6636%
Reliability (standard deviation) 0.2143
Weighted precision 0.9049
Weighted recall 0.9013
Weighted F1 score 0.9006

```

16/10/23 21:12:54 INFO MahoutDriver: Program took 4033 ms (Minutes: 0.06721666666666666)
root@nan-Thundekobot:/usr/local/lib/mahout#

SGD result

```

./bin/mahout org.apache.mahout.classifier.sgd.TrainNewsGroups
${WORK_DIR}/20news-bydate/20news-bydate-train/
./bin/mahout org.apache.mahout.classifier.sgd.TestNewsGroups --input
${WORK_DIR}/20news-bydate/20news-bydate-test/ --model /tmp/news-group.model

```

```

0   0   4   291   5   15   6   7   2   3   0   6   10   1   1   3   6
comp.graphics      0   3   4   11   193   0   9   4   5   3   3   4   4   1   0   27   11
soc.religion.christian  0   1   2   43   3   240   7   5   12   1   3   1   7   1   1   0   1
comp.windows.x     0   3   2   10   2   3   333   4   2   1   1   3   1   5   0   2   0
sci.space          0   0   6   14   2   2   8   292   12   1   1   3   2   0   0   1   0
rec.autos          1   0   10   4   0   0   4   13   313   0   0   1   19   0   0   0   0
misc.forsale       0   14   0   5   0   0   5   7   0   142   4   85   0   7   4   0   25
talk.politics.misc 0   2   0   3   1   1   0   2   2   1   336   1   1   0   0   2   3
rec.sport.hockey    0   6   1   3   6   1   2   11   1   10   0   282   0   2   3   3   17
talk.politics.guns 0   0   25   23   0   3   3   3   11   0   0   1   242   0   0   0   0
comp.sys.ibm.pc.hardware 0   4   6   23   12   3   6   17   11   9   4   4   13   192   1   7   9
sci.med            0   3   0   11   2   2   5   9   1   5   0   12   2   2   305   2   0
sci.crypt          0   11   0   4   21   0   5   7   2   2   1   1   1   3   1   195   51
alt.atheism         1   7   2   5   14   0   10   5   1   4   1   12   1   3   0   33   14
talk.religion.misc 0   3   0   7   1   0   0   1   6   3   22   1   0   1   0   1   3
rec.sport.baseball 0   1   17   44   0   8   6   4   3   2   1   0   51   0   2   4   7
comp.os.ms-windows.misc 0   1   17   39   2   2   14   15   8   0   1   5   17   0   4   2   1
sci.electronics

=====
Statistics
-----
Kappa                  0.6052
Accuracy               64.9363%
Reliability            61.5632%
Reliability (standard deviation) 0.2264
Weighted precision     0.6929
Weighted recall        0.6494
Weighted F1 score      0.6422
Log-likelihood
  mean      : -1.4236
  25%-ile   : -2.1524
  75%-ile   : -0.3053

16/10/23 21:17:39 INFO MahoutDriver: Program took 3304 ms (Minutes: 0.05506666666666667)
root@nan-ThundeRobot:/usr/local/lib/mahout#
```

Analysis

C-Naive Bayes and Naive Bayes have similar outcomes. SGD(Stochastic Gradient Decent) has much worse performance in terms of accuracy/reliability but better performance on run-time comparing to C-Naive Bayes and Naive Bayes.

3-2 Classification on wiki data

3.2.1 Prepare data

```

export WORK_DIR=/tmp/mahout-work-wiki
mkdir -p ${WORK_DIR}
mkdir -p ${WORK_DIR}/wikixml
cd ${WORK_DIR}/wikixml
root@nan-ThundeRobot:/tmp# cp /home/nan/Downloads/enwiki-articles.xml
${WORK_DIR}/wikixml
rm -rf ${WORK_DIR}/wiki
```

```
mkdir ${WORK_DIR}/wiki
root@nan-ThundeRobot:/tmp/mahout-work-wiki# cp
$MAHOUT_HOME/examples/bin/resources/country10.txt ${WORK_DIR}/country.txt
```

```
hadoop dfs -mkdir -p ${WORK_DIR}
hadoop dfs -put ${WORK_DIR}/wikixml ${WORK_DIR}/wikixml
```

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin
[sudo] password for nan:
[bash]: /usr/local/Cellar: No such file or directory
root@nan-ThundeRobot:~ export WORK_DIR=/tmp/mahout-work-wiki
root@nan-ThundeRobot:~ cd $MAHOUT_HOME/examples/bin/resources/country10.txt ${WORK_DIR}/country.txt
cp: cannot create regular file '/tmp/mahout-work-wiki/country.txt': No such file or directory
root@nan-ThundeRobot:~ mkdir -p ${WORK_DIR}/wikixml
root@nan-ThundeRobot:~ cp /home/nan/Downloads/enwiki-articles.xml ${WORK_DIR}/wikixml
root@nan-ThundeRobot:~ cd $MAHOUT_HOME/examples/bin/resources/country10.txt ${WORK_DIR}/country.txt
root@nan-ThundeRobot:~ cd /usr/local/lib/mahout/b
[bash]: cd /usr/local/lib/mahout/b: No such file or directory
root@nan-ThundeRobot:~ cd /usr/local/lib/mahout/bin/
root@nan-ThundeRobot:~ /usr/local/lib/mahout/bin# ls
compute-classpath.sh mahout-load-spark-env.sh mahout-spark-class.sh
root@nan-ThundeRobot:~ /usr/local/lib/mahout/bin# ./mahout seqwiki -c ${WORK_DIR}/country.txt -i ${WORK_DIR}/wikixml/enwiki-articles.xml -o ${WORK_DIR}/wikipediainput
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [/jarfile:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [/jarfile:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 01:32:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/24 01:32:42 INFO WikipediaInputFormat: Input: /tmp/mahout-work-wiki/wikixml/enwiki-articles.xml Out: /tmp/mahout-work-wiki/wikipediainput Categories: /tmp/mahout-work-wiki/country.txt All Files: false
16/10/24 01:32:42 INFO depreciation: session.id is deprecated. Instead, use dfs.metrics.session.id
16/10/24 01:32:42 INFO JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/10/24 01:32:42 WARN JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
```

3.2.2 seqwiki

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout seqwiki -c
${WORK_DIR}/country.txt -i ${WORK_DIR}/wikixml/enwiki-articles.xml -o
${WORK_DIR}/wikipediainput
```

```
Reduce output records=288
Spilled Records=576
Shuffled Maps =5
Failed Shuffles=0
Merged Map outputs=5
GC time elapsed (ms)=446
CPU time spent (ms)=9
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=10169614336
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=160859991
File Output Format Counters
  Bytes Written=14777035
16/10/24 01:33:00 INFO MahoutDriver: Program took 17936 ms (Minutes: 0.2989333333333333)
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout seq2sparse -i ${WORK_DIR}/wikipediainput -o ${WORK_DIR}/wikipediaVecs -wt tfidf -lnorm -nv -ow -ng 2
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [/jarfile:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [/jarfile:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 01:33:46 INFO SparseVectorsFromSequenceFiles: Maximum n-gram size is: 2
16/10/24 01:33:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/24 01:33:47 INFO SparseVectorsFromSequenceFiles: Minimum LLR value: 1.0
16/10/24 01:33:47 INFO SparseVectorsFromSequenceFiles: Number of reduce tasks: 1
```

3.2.3 seq2sparse

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout seq2sparse -i
${WORK_DIR}/wikipediainput -o ${WORK_DIR}/wikipediaVecs -wt tfidf -lnorm -nv -ow -ng 2
```

```

Reducer Output Records=288
Spilled Records=576
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=3415212032

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  INVALID_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=10266207
File Output Format Counters
  Bytes Written=10266207

16/10/24 01:34:37 INFO HadoopUtil: Deleting /tmp/mahout-work/wiki/wikipediaVecs/partial_vectors-0
16/10/24 01:34:37 INFO MahoutDriver: Program took 58299 ms (Minutes: 0.8383166666666667)
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout split -i ${WORK_DIR}/tmp/wikipediaVecs/tfidf-vectors/ --trainingOutput ${WORK_DIR}/training --testOutput ${WORK_DIR}/testing -rp 20 -ow -seq -xm sequential
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 01:34:45 WARN MahoutDriver: No split.props found on classpath, will use command-line arguments only
16/10/24 01:34:45 INFO AbstractJob: Command line arguments: [-endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/tmp/wikipediaVecs/tfidf-vectors/], --method=[sequential], --overwrite=null, --randomSelectionPct=[20], --sequenceFiles=null, --startPhase=[0], --tempDir=[/tmp], --testOutput=[/tmp/mahout-work-wiki/testing], --trainingOutput=[/tmp/mahout-work-wiki/training]}
16/10/24 01:34:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
Exception in thread "main" java.io.FileNotFoundException: /tmp/mahout-work-wiki/tmp/wikipediaVecs/tfidf-vectors does not exist

```

3.2.4 split

```

root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout split -i
${WORK_DIR}/wikipediaVecs/tfidf-vectors/ --trainingOutput ${WORK_DIR}/training --testOutput
${WORK_DIR}/testing -rp 20 -ow -seq -xm sequential

```

```

root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout split -i ${WORK_DIR}/wikipediaVecs/tfidf-vectors/ --trainingOutput ${WORK_DIR}/training --testOutput ${WORK_DIR}/testing
ntial.
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 01:34:59 WARN MahoutDriver: No split.props found on classpath, will use command-line arguments only
16/10/24 01:34:59 INFO AbstractJob: Command line arguments: [-endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/wikipediaVecs/tfidf-vectors/], --method=[sequential], --overwrite=null, --randomSelectionPct=[20], --sequenceFiles=null, --startPhase=[0], --tempDir=[/tmp], --testOutput=[/tmp/mahout-work-wiki/testing], --trainingOutput=[/tmp/mahout-work-wiki/training]}
16/10/24 01:35:00 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
16/10/24 01:35:00 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/training
16/10/24 01:35:00 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/testing
16/10/24 01:35:00 INFO SplitInput: part-r-00000 has 130607 lines
16/10/24 01:35:00 INFO SplitInput: part-r-00000 test split size is 26121 based on random selection percentage 20
16/10/24 01:35:00 INFO CodecPool: Got brand-new compressor [deflate]
16/10/24 01:35:00 INFO CodecPool: Got brand-new compressor [deflate]
16/10/24 01:35:01 INFO MahoutDriver: Program took 1392 ms (Minutes: 0.0232)
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout trainnb -i ${WORK_DIR}/training -o ${WORK_DIR}/model -li ${WORK_DIR}/labelindex -ow -c
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 01:36:13 WARN MahoutDriver: No trainnb.props found on classpath, will use command-line arguments only
16/10/24 01:36:13 INFO AbstractJob: Command line arguments: [-alpha=[1.0], --endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/training], --labelIndex=[/tmp/mahout-work-wiki/training/model], --overwrite=null, --startPhase=[0], --tempDir=[/tmp], --trainComplementary=null]
16/10/24 01:36:13 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable

```

Naive Bayes

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout trainnb -i ${WORK_DIR}/training -o ${WORK_DIR}/model -li ${WORK_DIR}/labelindex -ow -c
```

```
Comding input records=1
Combining output records=1
Reduce input groups=1
Reduce shuffle bytes=106
Reduce input records=1
Reduce output records=1
Spilled Records=2
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC Time elapsed (ms)=160
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=991952896
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=4263402
File Output Format Counters
  Bytes Written=197
16/10/24 01:36:19 INFO MahoutDriver: Program took 6317 ms (Minutes: 0.1052833333333334)
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout testnb -i ${WORK_DIR}/model -l ${WORK_DIR}/labelindex -ow -o ${WORK_DIR}/output -c
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT LOCAL is set, running in local mode.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 01:38:55 WARN MahoutDriver: No testnb.props found on classpath, will use command-line arguments only
16/10/24 01:38:55 INFO AbstractJob: Command line arguments: {-endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/training], --labelIndex=[/tmp/mahout-work-wiki/labelindex], --model=[/tmp/mahout-wiki/model], --output=/tmp/mahout-work-wiki/output}, --overwrite=null, --startPhase=[0], --tempDir=[temp], --testComplementary=null}
16/10/24 01:38:56 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/24 01:38:56 INFO Deprecation: mapred deprecated instead of mapreduce. Instead, use mapreduce.map.output.compress=true
16/10/24 01:38:56 INFO Deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
16/10/24 01:38:56 INFO Deprecation: fileoutputformat.outputdir is deprecated. Instead, use file.outputformat.outputdir
```

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout testnb -i ${WORK_DIR}/training -m ${WORK_DIR}/model -l ${WORK_DIR}/labelindex -ow -o ${WORK_DIR}/output -c
```

```
Bytes Written=23254
16/10/24 01:38:58 INFO TestNaiveBayesDriver: Complementary Results:
=====
Summary
-----
Correctly Classified Instances : 227 100%
Incorrectly Classified Instances : 0 0%
Total Classified Instances : 227
=====

Confusion Matrix
-----
a   b   c   d   e   f   g   h   i   j   <-Classified as
44   0   0   0   0   0   0   0   0   0   | 44   a   = australia
0   30   0   0   0   0   0   0   0   0   | 30   b   = austria
0   0   3   0   0   0   0   0   0   0   | 3   c   = bahamas
0   0   0   43   0   0   0   0   0   0   | 43   d   = canada
0   0   0   0   10   0   0   0   0   0   | 10   e   = colombia
0   0   0   0   0   8   0   0   0   0   | 8   f   = cuba
0   0   0   0   0   0   14   0   0   0   | 14   g   = pakistan
0   0   0   0   0   0   0   1   0   0   | 1   h   = panama
0   0   0   0   0   0   0   0   66   0   | 66   i   = united kingdom
0   0   0   0   0   0   0   0   0   8   | 8   j   = vietnam
=====

Statistics
-----
Kappa : 0.9492
Accuracy : 100%
Reliability : 99.9091%
Reliability (standard deviation) : 0.3015
Weighted precision : 1
Weighted recall : 1
Weighted F1 score : 1
16/10/24 01:38:58 INFO MahoutDriver: Program took 2941 ms (Minutes: 0.0490166666666667)
```

C Naive Bayes

```
root@nan-ThundeRobot:/tmp/mahout-work-wiki# cp  
$MAHOUT_HOME/examples/bin/resources/country.txt ${WORK_DIR}/country.txt
```

```
vectordump: : Dump vectors from a sequence file to text  
viterbi: : Viterbi decoding of hidden states from given output states sequence  
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# cp $MAHOUT_HOME/examples/bin/resources/country.txt ${WORK_DIR}/country.txt  
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout seqwiki -c ${WORK_DIR}/country.txt -i ${WORK_DIR}/wikixml/enwiki-articles.xml -o ${WORK_DIR}/wikipediainput  
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.  
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
16/10/24 01:56:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
16/10/24 01:56:49 INFO WikipediaInputToSequenceFile: Input: /tmp/mahout-work-wiki/wikixml/enwiki-articles.xml Out: /tmp/mahout-work-wiki/wikipediainput Categories: /tmp/mahout-work-wiki/country.txt All File  
File  
16/10/24 01:56:49 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/wikipediainput  
16/10/24 01:56:49 INFO deprecation: session_id is deprecated. Instead, use dfs.metrics.session-id  
16/10/24 01:56:49 INFO JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
```

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout seqwiki -c  
${WORK_DIR}/country.txt -i ${WORK_DIR}/wikixml/enwiki-articles.xml -o  
${WORK_DIR}/wikipediainput
```

```
Spilled Records=3058  
Shuffled Maps=5  
Failed Shuffles=0  
Merged Map outputs=5  
GC time elapsed (ms)=501  
CPU time spent (ms)=0  
Physical memory (bytes) snapshot=0  
Virtual memory (bytes) snapshot=0  
Total committed heap usage (bytes)=9836691456  
Shuffle Errors  
BAD ID=8  
CONNECTION=0  
IO_ERROR=0  
WRONG_LENGTH=0  
WRONG_MAP=0  
WRONG_REDUCE=0  
File Input Format Counters  
Bytes Read=160825991  
File Output Format Counters  
Bytes Written=61612465  
16/10/24 01:57:30 INFO MahoutDriver: Program took 41856 ms (Minutes: 0.6976)  
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout seq2sparse -i ${WORK_DIR}/wikipediainput -o ${WORK_DIR}/wikipediaVecs -wt tfidf -lnorm -nv -ow -ng 2  
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.  
MAHOUT LOCAL is set, running locally  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
16/10/24 01:57:41 INFO SparseVectorsFromSequenceFiles: Maximum n-gram size is: 2  
16/10/24 01:57:41 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
16/10/24 01:57:41 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/wikipediainput  
16/10/24 01:57:41 INFO SparseVectorsFromSequenceFiles: Merging all RLE values: 1  
16/10/24 01:57:41 INFO SparseVectorsFromSequenceFiles: Number of reduce tasks: 1  
16/10/24 01:57:41 INFO SparseVectorsFromSequenceFiles: Tokenizing documents in /tmp/mahout-work-wiki/wikipediainput  
16/10/24 01:57:41 INFO deprecation: session_id is deprecated. Instead, use dfs.metrics.session-id  
16/10/24 01:57:41 INFO JvmMetrics: Initializing JVM Metrics with processName=Jobtracker, sessionId=
```

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout seq2sparse -i  
${WORK_DIR}/wikipediainput -o ${WORK_DIR}/wikipediaVecs -wt tfidf -lnorm -nv -ow -ng 2
```

```
Spilled Records=3058  
Shuffled Maps=2  
Failed Shuffles=0  
Merged Map outputs=2  
GC time elapsed (ms)=0  
CPU time spent (ms)=0  
Physical memory (bytes) snapshot=0  
Virtual memory (bytes) snapshot=0  
Total committed heap usage (bytes)=5745672192  
Shuffle Errors  
BAD_ID=8  
CONNECTION=0  
IO_ERROR=0  
WRONG_LENGTH=0  
WRONG_MAP=0  
WRONG_REDUCE=0  
File Input Format Counters  
Bytes Read=52337918  
File Output Format Counters  
Bytes Written=52268866  
16/10/24 02:00:18 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/wikipediaVecs/partial-vectors-0  
16/10/24 02:00:18 INFO MahoutDriver: Program took 150099 ms (Minutes: 2.633483333333334)  
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout split -i ${WORK_DIR}/wikipediaVecs/tfidf-vectors/ --trainingOutput ${WORK_DIR}/training --testOutput ${WORK_DIR}/testing -rp 20 -ow -seq -xm sequential  
MAHOUT LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.  
MAHOUT LOCAL is set, running locally  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
16/10/24 02:01:13 WARN MahoutDriver: No split.props found on Classpath, will use command-line arguments only  
16/10/24 02:01:13 INFO Abstraction: Command line arguments: {-endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/wikipediaVecs/tfidf-vectors/], --method=[sequential], --overwrite=null, --randomSelect  
ionPct=[20], --sequencefiles=null, --startPhase=[0], --tempDir=[/tmp], --testOutput=[/tmp/mahout-work-wiki/testing], --trainingOutput=[/tmp/mahout-work-wiki/training]}  
16/10/24 02:01:13 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
16/10/24 02:01:13 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/training  
16/10/24 02:01:13 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/testing  
16/10/24 02:01:14 INFO SplitInput: part-r-00000 has 304258 lines  
16/10/24 02:01:14 INFO SplitInput: part-r-00000 test split size is 60850 based on random selection percentage 20
```

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout split -i ${WORK_DIR}/wikipediaVecs/tfidf-vectors/ --trainingOutput ${WORK_DIR}/training --testOutput ${WORK_DIR}/testing -rp 20 -ow -seq -xm sequential
```

```
SLF4J: Found binding in [/usr/libexec/java/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 02:01:13 WARN MahoutDriver: No split.props found on classpath, will use command-line arguments only
16/10/24 02:01:13 INFO AbstractJob: Command line arguments: {--endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/wikipediaVecs/tfidf-vectors/], --method=[sequential], --ionPct=[20], --sequenceFiles=null, --startPhase=[0], --tempDir=[/tmp], --testOutput=[/tmp/mahout-work-wiki/testing], --trainingOutput=[/tmp/mahout-work-wiki/training]}
16/10/24 02:01:13 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/24 02:01:13 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/training
16/10/24 02:01:13 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/testing
16/10/24 02:01:14 INFO SplitInput: part-r-00000 has 304250 lines
16/10/24 02:01:14 INFO SplitInput: part-r-00000 test split size is 60850 based on random selection percentage 20
16/10/24 02:01:14 INFO CodecPool: Got brand-new compressor [.deflate]
16/10/24 02:01:14 INFO CodecPool: Got brand-new compressor [.deflate]
16/10/24 02:01:18 INFO SplitInput: file: part-r-00000, input: 304250 train: 1214, test: 315 starting at 0
16/10/24 02:01:18 INFO MahoutDriver: Program took 5626 ms (Minutes: 0.0937666666666666)
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout trainnb -i ${WORK_DIR}/training -o ${WORK_DIR}/model -li ${WORK_DIR}/labelindex -ow -c
MAHOUT_LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT_LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 02:01:41 WARN MahoutDriver: No trainnb.props found on classpath, will use command-line arguments only
16/10/24 02:01:41 INFO AbstractJob: Command line arguments: {--alpha=[1.0], --endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/training], --labelIndex=[/tmp/mahout-work-wiki/model], --overwrite=null, --startPhase=[0], --tempDir=[/tmp], --trainComplementary=null}
16/10/24 02:01:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout trainnb -i ${WORK_DIR}/training -o ${WORK_DIR}/model -li ${WORK_DIR}/labelindex -ow -c
```

```
16/10/24 02:02:11 INFO Job: Counters: 33
  File System Counters
    FILE: Number of bytes read=768008504
    FILE: Number of bytes written=634675287
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=115
    Map output records=1
    Map output bytes=927
    Map output materialized bytes=838
    Input split bytes=132
    Combine input records=1
    Combine output records=1
    Reduce input groups=1
    Reduce shuffle bytes=838
    Reduce input records=1
    Reduce output records=1
    Spilled Records=2
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=60
    CPU time spent (ms)=0
    Physical memory (bytes) snapshot=0
    Virtual memory (bytes) snapshot=0
    Total committed heap usage (bytes)=1110441984
  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=24525810
  File Output Format Counters
    Bytes Written=1045
16/10/24 02:02:12 INFO MahoutDriver: Program took 30569 ms (Minutes: 0.5094833333333333)
root@nan-ThundeRobot:/usr/local/lib/mahout/bin#
```

```
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout testnb -i ${WORK_DIR}/training -m ${WORK_DIR}/model -l ${WORK_DIR}/labelindex -ow -o ${WORK_DIR}/output -c
```

```
0      0      0      0      0      0      0      0      0      0      0  
0      0      0      0      0      0      0      0      0      0      0  
0      0      0      0      0      0      0      0      0      0      0  
  
=====  
statistics  
-----  
f1 score                                0.9075  
accuracy                                 99.9176%  
reliability                               99.111%  
reliability (standard deviation)          0.0929  
weighted precision                        0.9992  
weighted recall                           0.9992  
weighted F1 score                         0.9992  
  
6/10/24 02:03:33 INFO MahoutDriver: Program took 51093 ms (Minutes: 0.85155)  
root@nan-ThundeRobot:/usr/local/lib/mahout/bin#
```

```
=====
Statistics
-----
Kappa                      0.9058
Accuracy                   100%
Reliability                99.1666%
Reliability (standard deviation) 0.0913
Weighted precision          1
Weighted recall              1
Weighted F1 score            1

16/10/24 02:08:12 INFO MahoutDriver: Program took 55113 ms (Minutes: 0.91855)
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# █
```

Analysis

Both Naive Bayes and C Naive Bayes provides similar outcomes. While the amount of categories has increased from 10 to 229, the accuracy did not change very much. Thus we can say, Naive Bayes is a reliable classification method in this circumstance.

4. Spark

4.1 Installation

```
nanhan-ThunderRobot:~/usr/local/spark-2.0.1-bin-hadoop2.7/bin$ ./pyspark
Python 2.7.12 |Anaconda custom (64-bit)| (default, Jul 2 2016, 17:42:40)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-1)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
Using Spark's default Log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
16/10/24 02:33:08 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/24 02:33:08 WARN Utils: Your hostname, nan-ThunderRobot resolves to a loopback address: 127.0.1.1; using 192.168.1.7 instead (on interface p3p1
16/10/24 02:33:08 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Welcome to

    _____
   /     \
  /       \
 /  \  /  \
/____\ /____\
           version 2.0.1

Using Python version 2.7.12 (default, Jul 2 2016 17:42:40)
SparkSession available as 'spark'.
>>> █
```

4.2 Test Cases