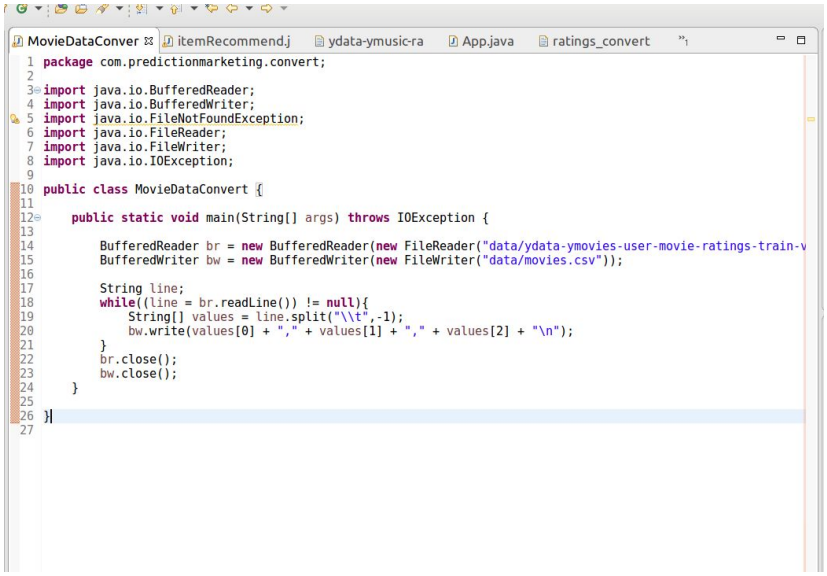
1.Recommender With Eclipse

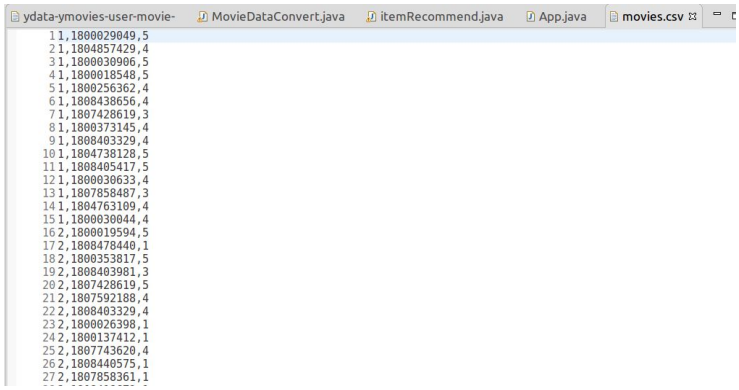
1-1 User Based Recommender

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

1. Convert data to csv, code



2. Converted data



3. Build Recommender

ydata-ymovies-user-movie- | MovieDataConvert.java | itemRecommend.java | App.java | movies.csv

```
1  package com.predictionmarketing.RecommenderApp;
2
3  import java.io.File;
15 /**
16  * Hello world!
17  *
18  */
19 public class App
20 {
21     public static void main( String[] args ) throws Exception
22     {
23         DataModel model = new FileDataModel(new File("data/movies.csv"));
24         UserSimilarity similarity = new PearsonCorrelationSimilarity(model);
25         UserNeighborhood neighborhood = new ThresholdUserNeighborhood(0.1, similarity, model);
26         UserBasedRecommender recommender =
27                 new GenericUserBasedRecommender(model, neighborhood, similarity);
28         List<RecommendedItem> recommendations = recommender.recommend(3, 10);
29         for (RecommendedItem recommendation : recommendations) {
30           System.out.println(recommendation);
31         }
32
33     }
34 }
35
```

4. Recommender Result

Problems | Javadoc | Declaration | Console

```
<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]
```

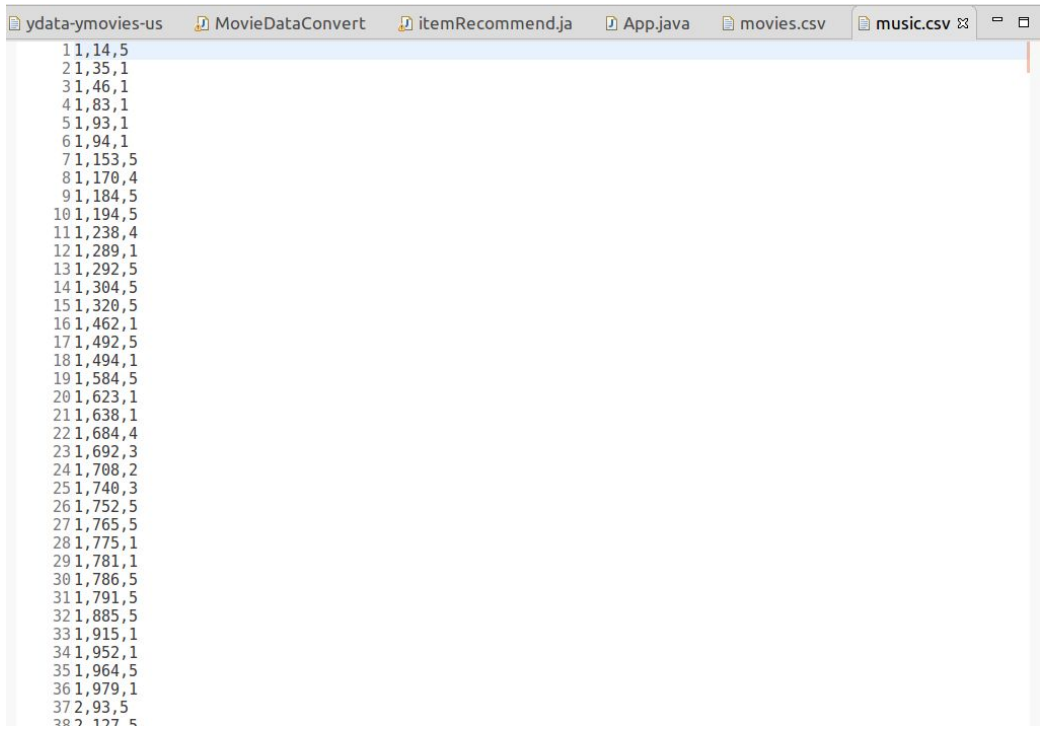
1-1-2 Data Used ydata-ymusic-rating-study-v1_0-train.txt

1.Code to convert to csv

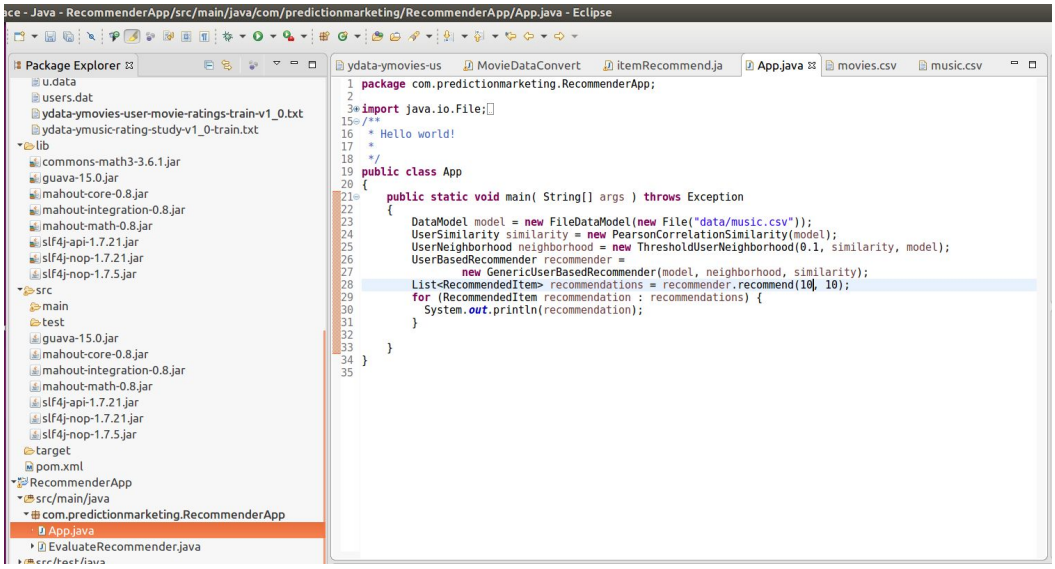
ydata-ymovies-user-movie- | MovieDataConvert.java | itemRecommend.java | App.java | movies.csv

```
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 }
27
```

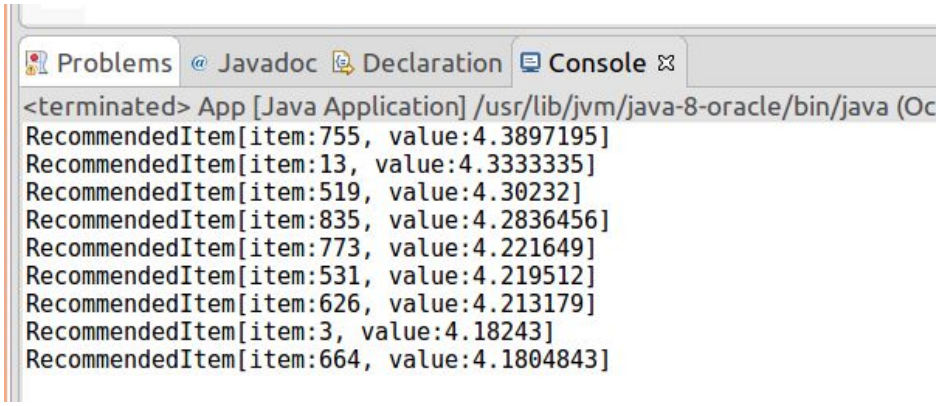
2.CSV result



3. Code for recommender



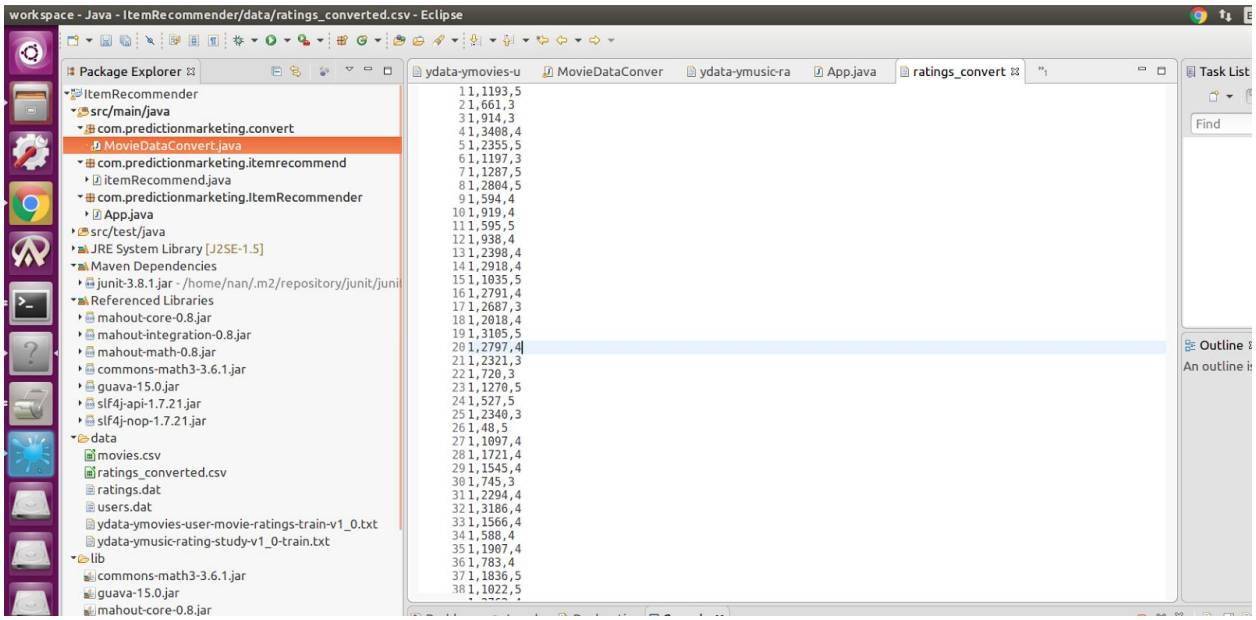
4. Recommender Result



1-2 Item Based Recommender

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

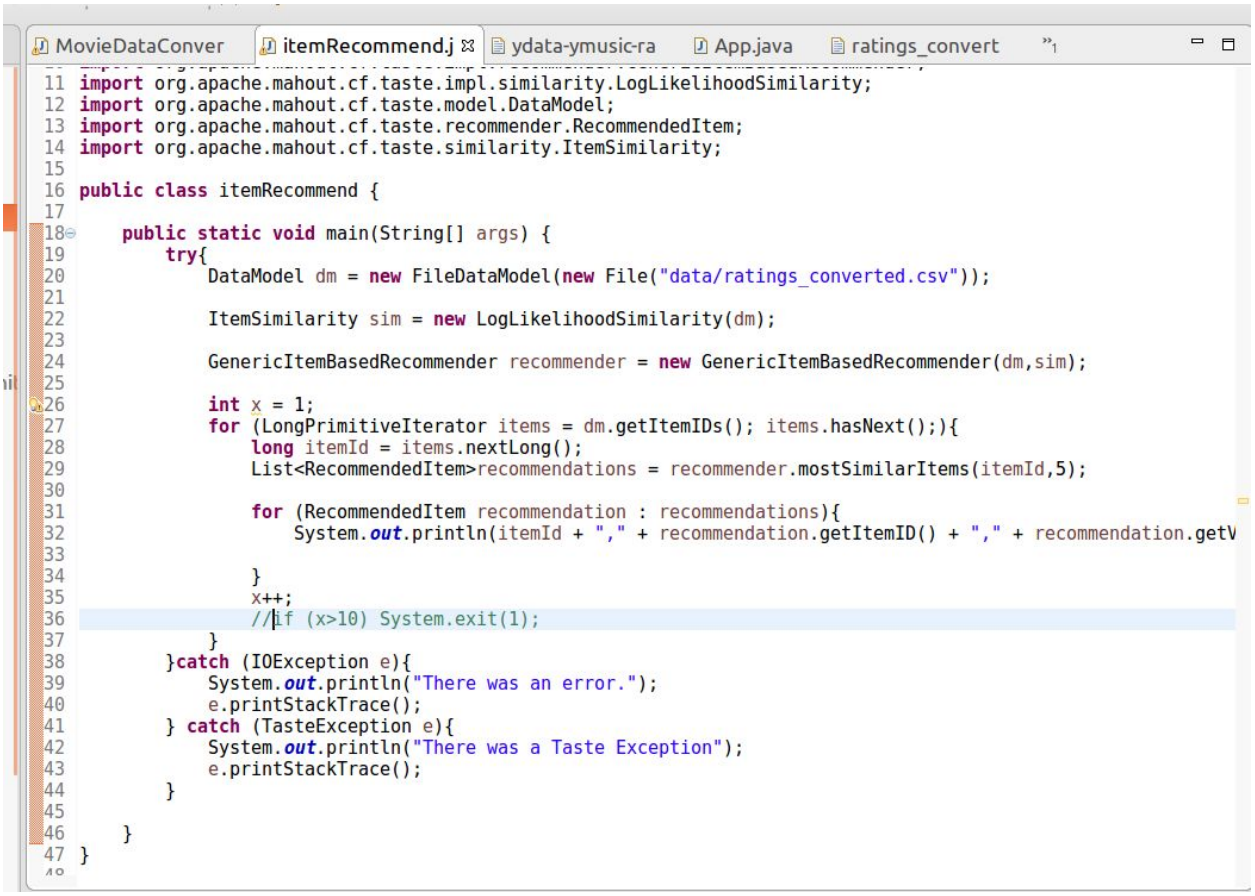
1.Convert to csv file



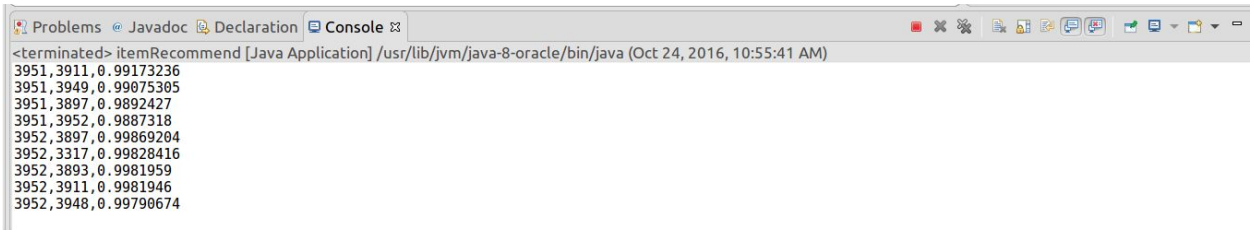
2. Code for coverter

```java
package com.predictionmarketing.convert;

import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;

public class MovieDataConvert {

    public static void main(String[] args) throws IOException {

        BufferedReader br = new BufferedReader(new FileReader("data/ratings.dat"));
        BufferedWriter bw = new BufferedWriter(new FileWriter("data/ratings_converted.csv"));

        String line;
        while((line = br.readLine()) != null){
            String[] values = line.split("::",-1);
            bw.write(values[0] + "," + values[1] + "," + values[2] + "\n");
        }
        br.close();
        bw.close();
    }

}
```

3. Build Recomemder

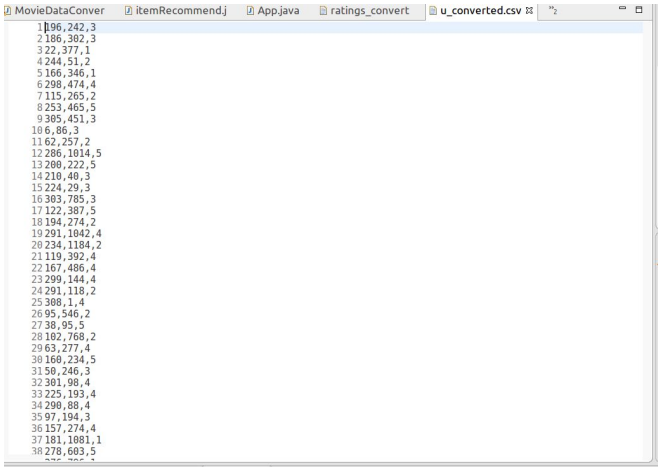


4. Recommender Result

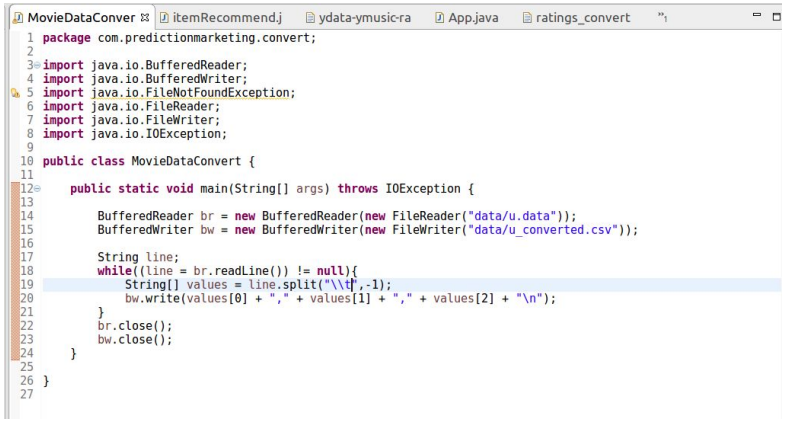


1-2-2 Data Used: http://grouplens.org/datasets/movielens/ 100K move data set u.data

1.Convert data



2. Code to convert data



3. Build Recommender

4. Recommender Results

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$ mkdir -p ${WORK_DIR}/reuters-out
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$ ./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: 0.026933333333333333)
```

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 ${WORK_DIR}/reuters-out-seqdir -c UTF-8 -chunk 64 -xm sequential
MAHOUT_LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT_LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/lib/mahout/examples/target/dependency/slf4j-log4j12-1.7.21.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/25 20:39:19 INFO AbstractJob: Command line arguments: {--charset=[UTF-8], --chunkSize=[64], --endPhase=[2147483647], --fileFilterClass=[org.apache.mahout.text.PrefixAdditionFilter], --input=[/tmp/mahout-work-nan/reuters-out], --keyPrefix=[], --method=[sequential], --output=[/tmp/mahout-work-nan/reuters-out-seqdir], --startPhase=[0], --tempDir=[temp]}
16/10/25 20:39:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/25 20:39:21 INFO MahoutDriver: Program took 2268 ms (Minutes: 0.0378)
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$
```

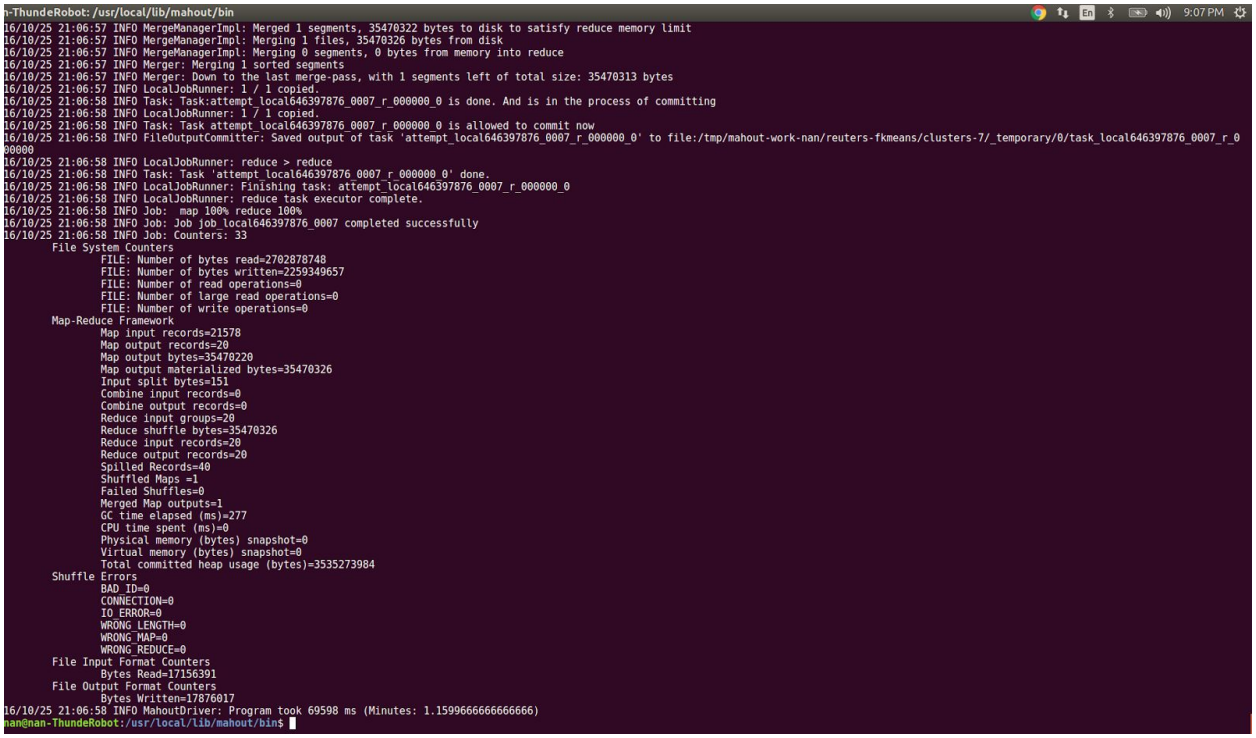
2.1.3 seq2sparse

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

```
nan@nan-ThundeRobot: /usr/local/lib/mahout/bin
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 Task: 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_local319021801_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: Finishing task: attempt_local319021801_0009_r_000000_0
16/10/25 20:40:48 INFO LocalJobRunner: reduce task executor complete.
16/10/25 20:40:48 INFO Job: 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: Number of bytes read=2094072244
                FILE: Number of bytes written=2038591065
                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=43156
                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)=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: 0.33493333333333336)
nan@nan-ThundeRobot:/usr/local/lib/mahout/bin$
```

2.1.4 fkmeans

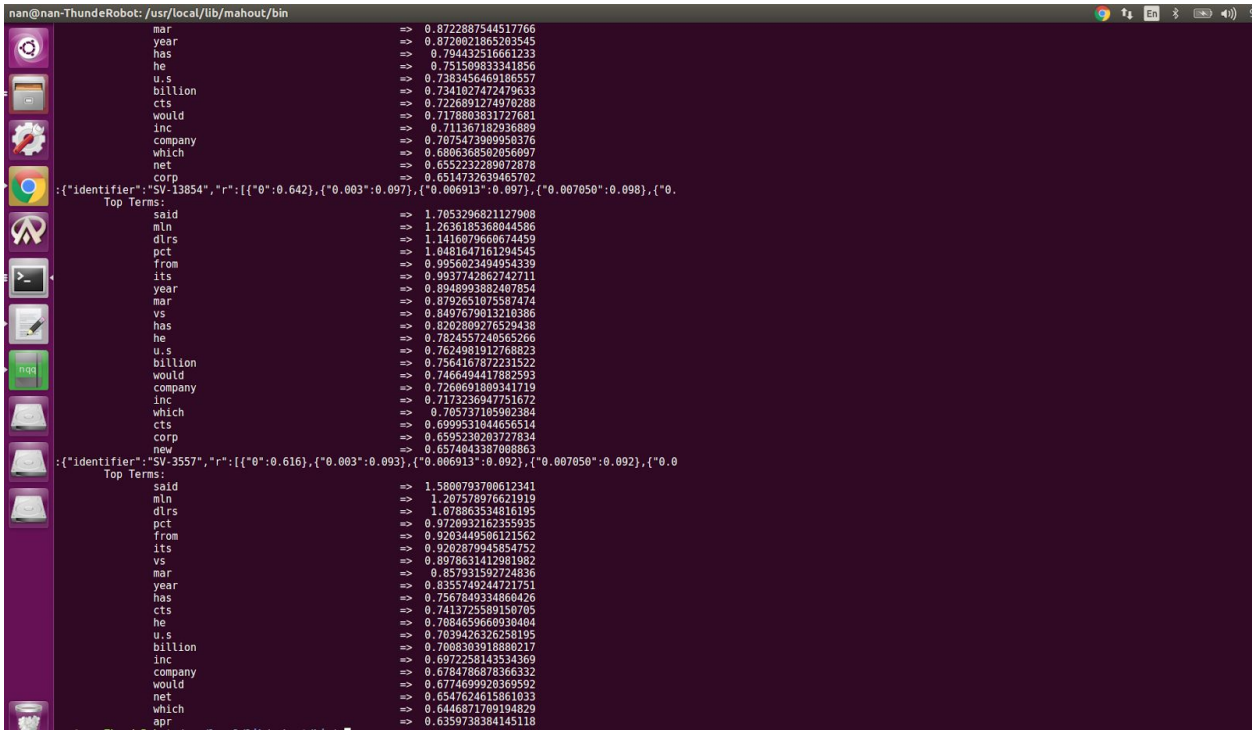
```
./mahout fkmeans \
  -i ${WORK_DIR}/reuters-out-seqdir-sparse-fkmeans/tfidf-vectors/ \
  -c ${WORK_DIR}/reuters-fkmeans-clusters \
  -o ${WORK_DIR}/reuters-fkmeans \
  -dm org.apache.mahout.common.distance.EuclideanDistanceMeasure \
  -x 10 -k 20 -ow -m 1.1
```



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



2.2 Wikipedia cluster

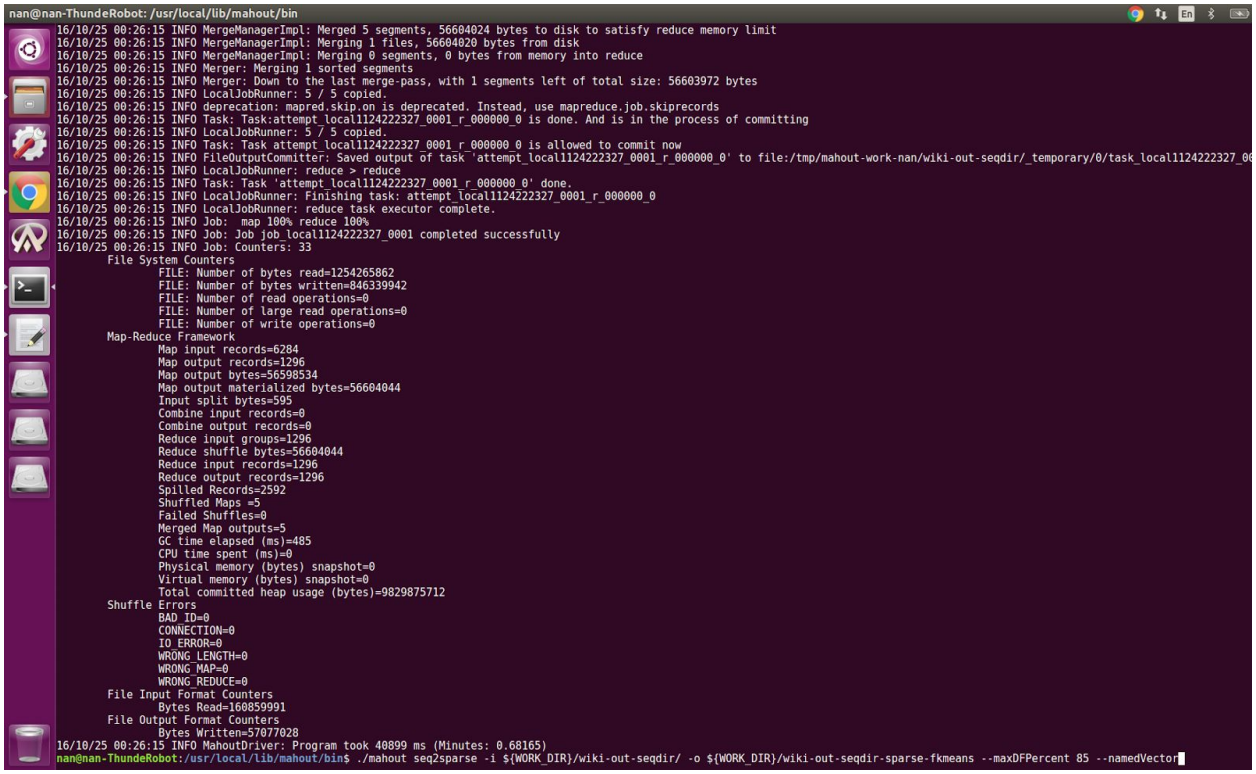
2.2.1 Build working directories

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                                                    => 0.6521729207225864
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-out ${WORK_DIR}/wiki-out
16/10/24 22:16:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
put: `/tmp/mahout-work-nan/wiki-out': No such file or directory
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs-mkdir ${WORK_DIR}/wiki-out
Error: Could not find or load main class dfs-mkdir
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ hdfs dfs -mkdir ${WORK_DIR}/wiki-out
16/10/24 22:16:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
mkdir: `/tmp/mahout-work-nan/wiki-out': File exists
nan@nan-ThundeRobot:/tmp/mahout-work-nan$ 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 ${WORK_DIR}/wiki-out
16/10/24 22:17:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
nan@nan-ThundeRobot:/tmp/mahout-work-nan$
```

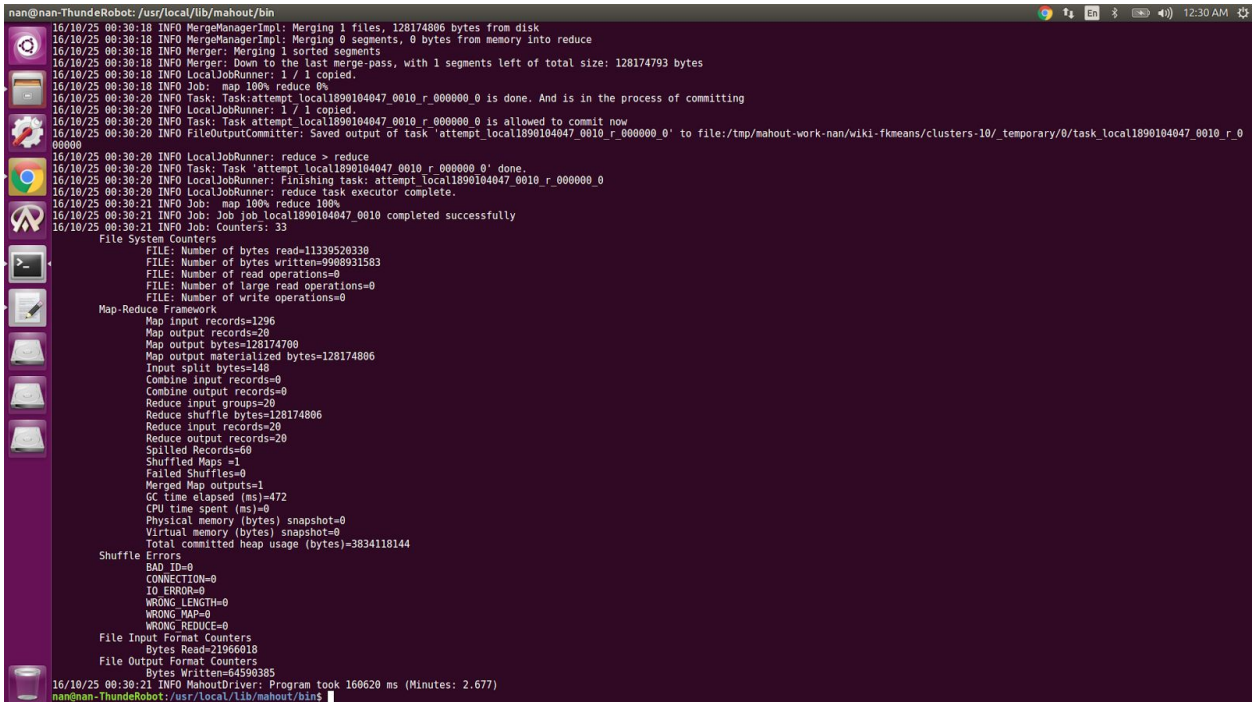
2.2.2 seqwiki

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



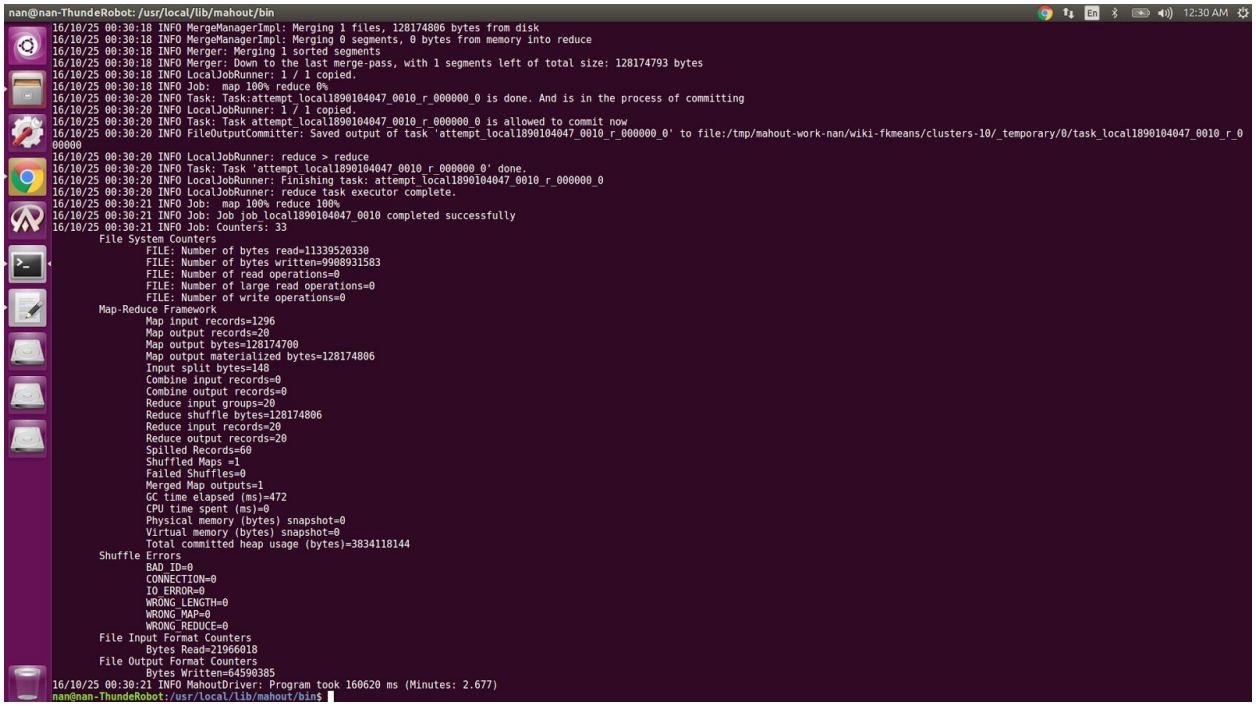
2.2.3 seq2sparse

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



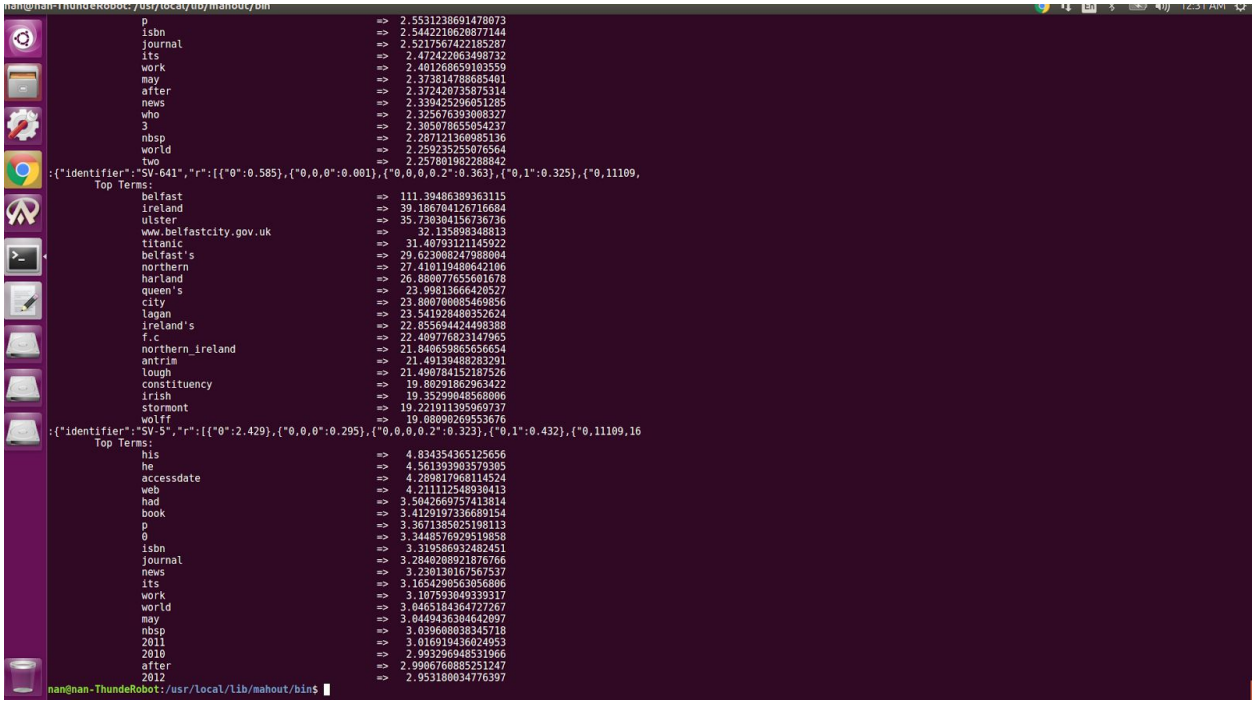
2.2.4 fkmeans

./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

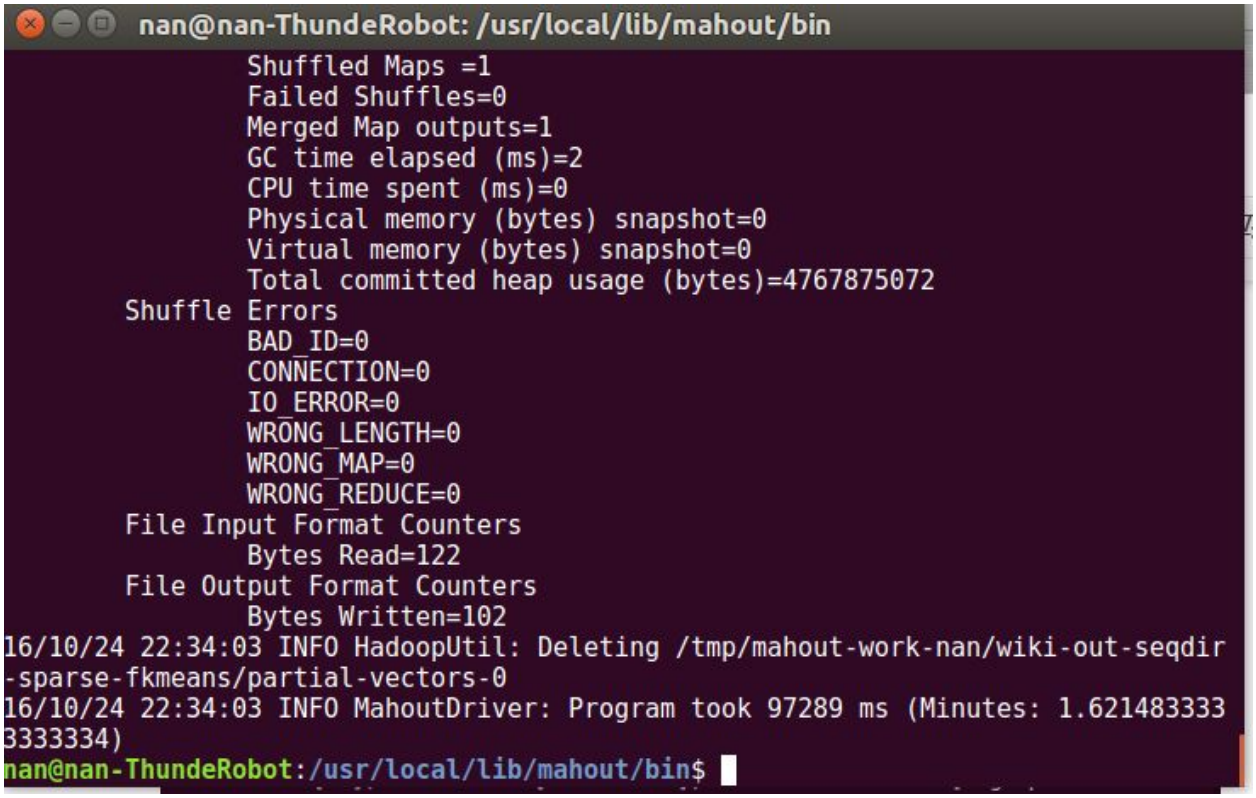


2.2.5 show data

```
./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
```



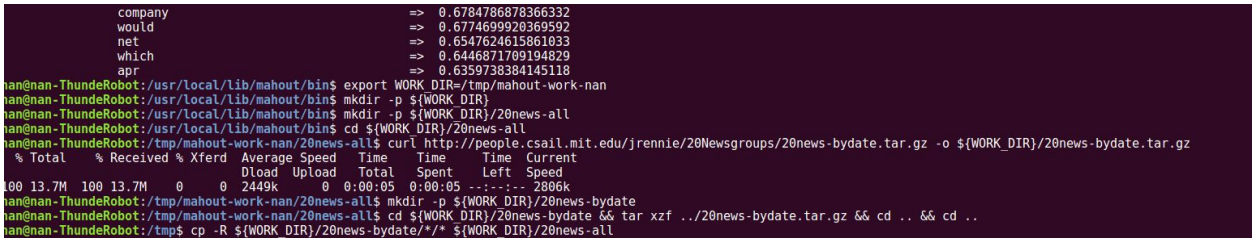
3. Classification with Mahout



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.

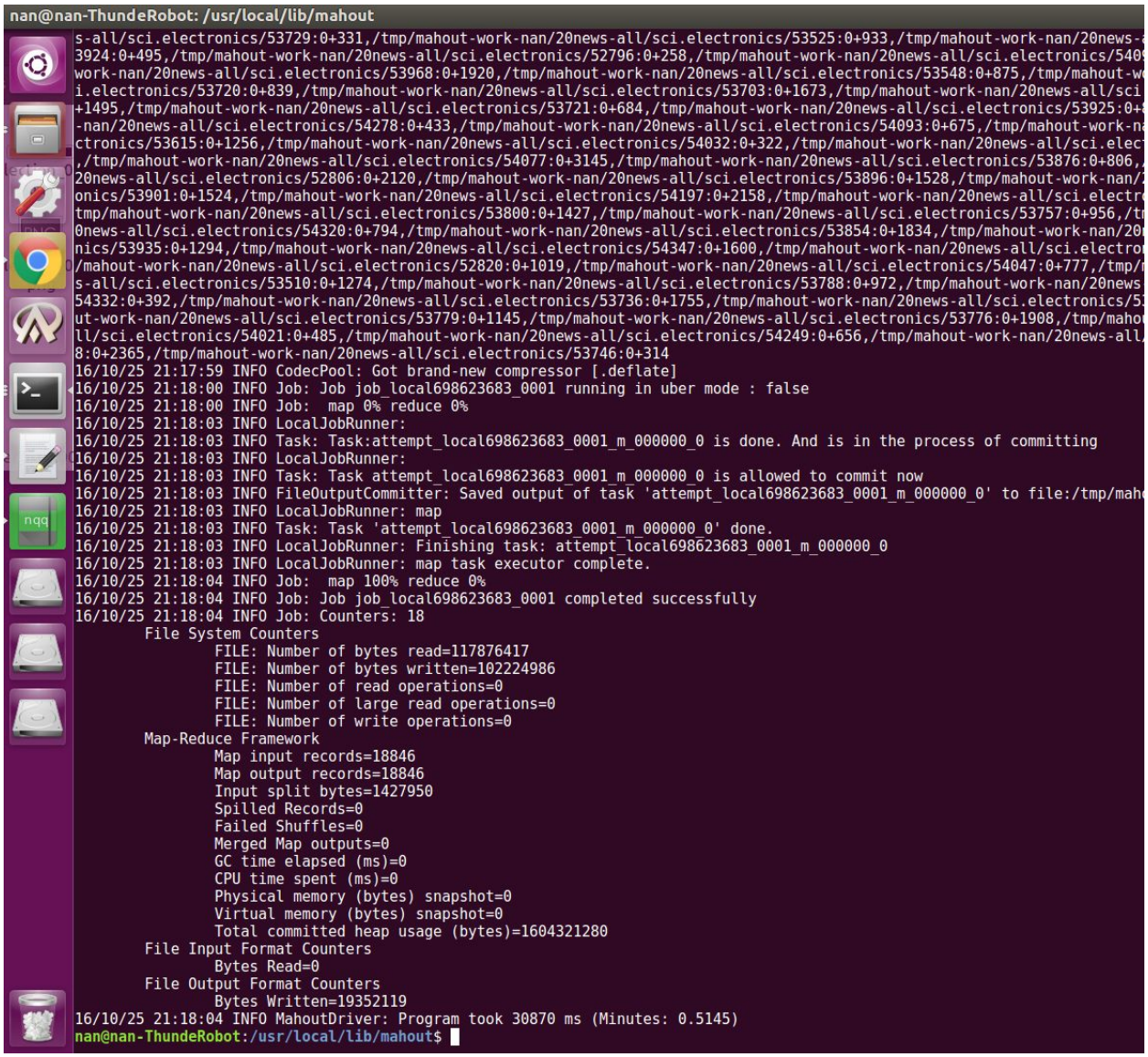
3.1.1 Prepare data



3.1.2 seqdirectory

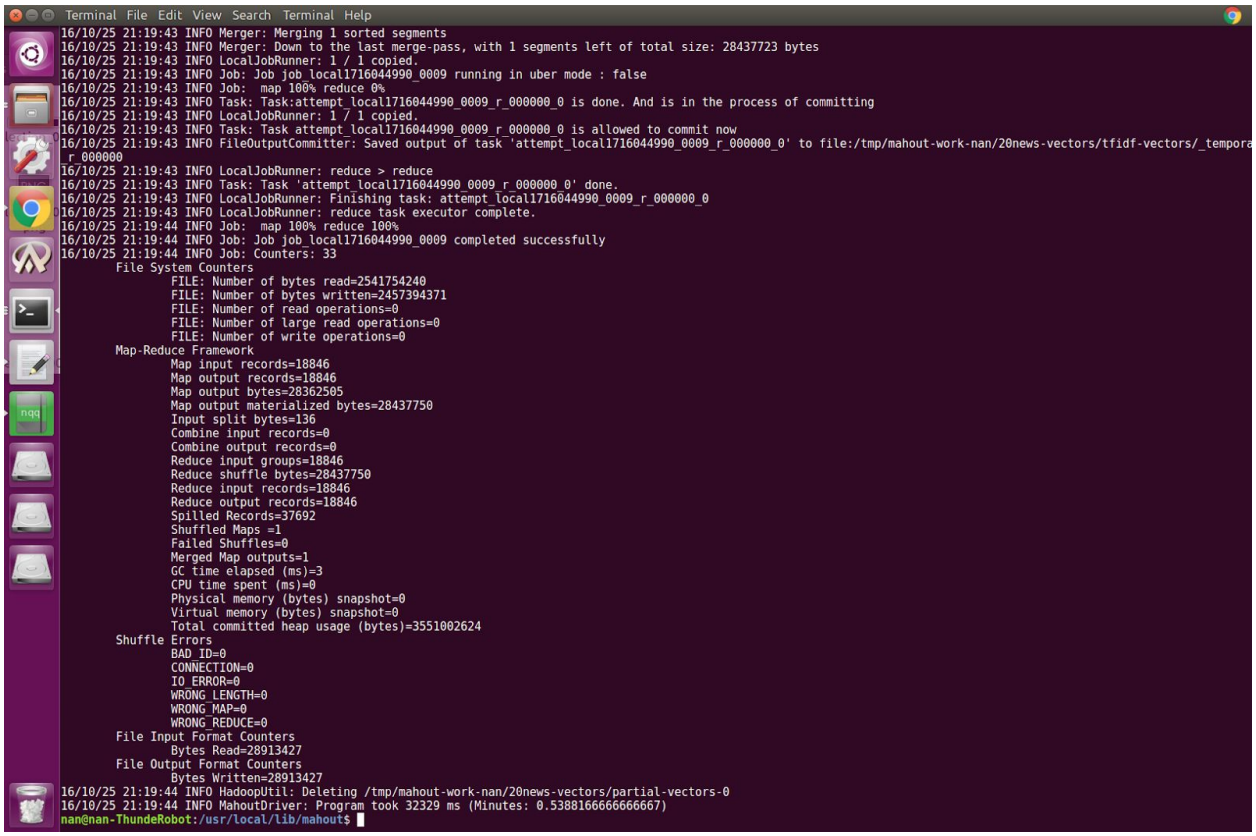
```
./bin/mahout seqdirectory \
  -i ${WORK_DIR}/20news-all \
```

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



3.1.3 seq2sparse

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



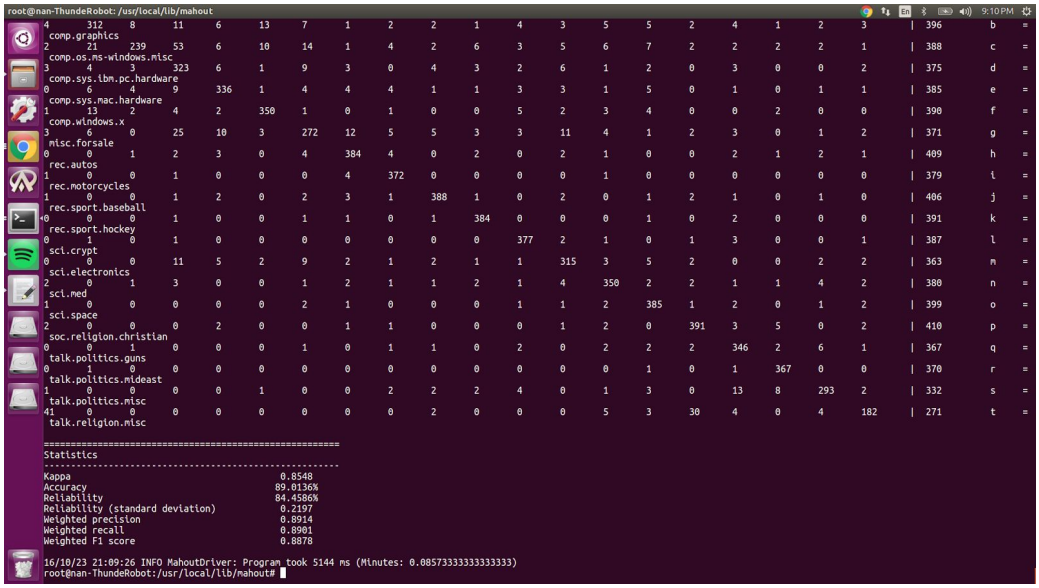
3.1.4 split

```
./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
```

3.1.5 train different model and test model and show result

C-NB result

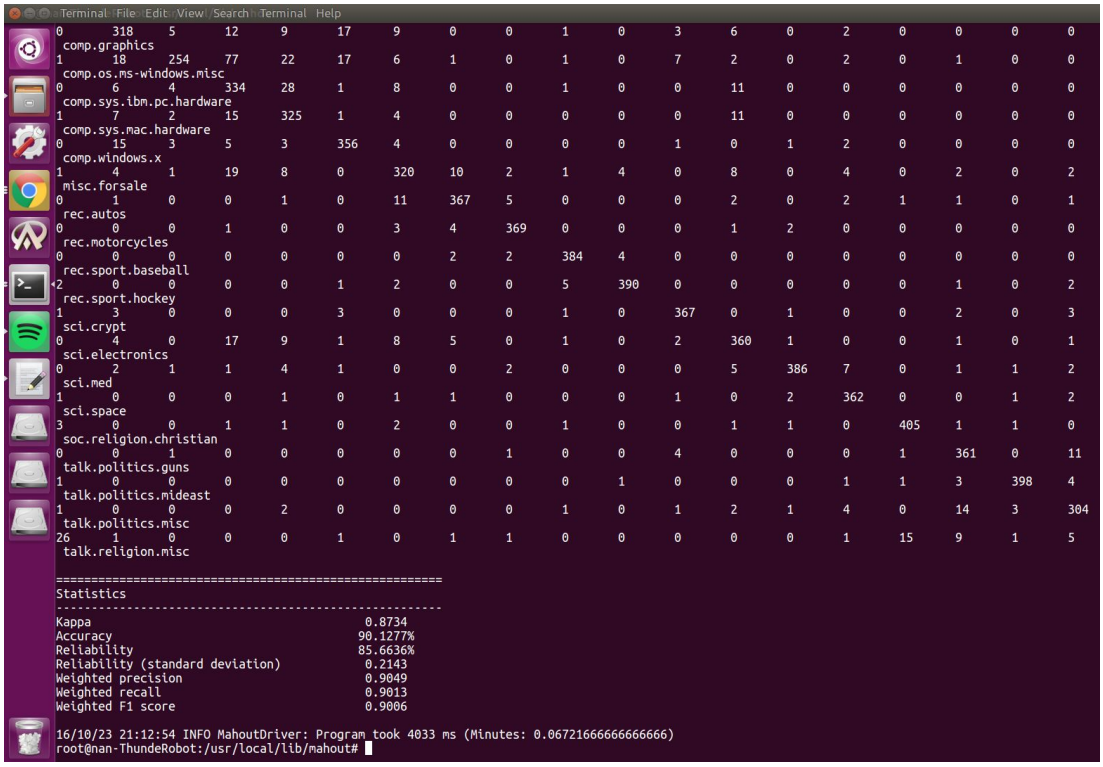
```
./bin/mahout trainnb \
  -i ${WORK_DIR}/20news-train-vectors \
  -o ${WORK_DIR}/model \
  -li ${WORK_DIR}/labelindex \
  -ow $c

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

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



NB result

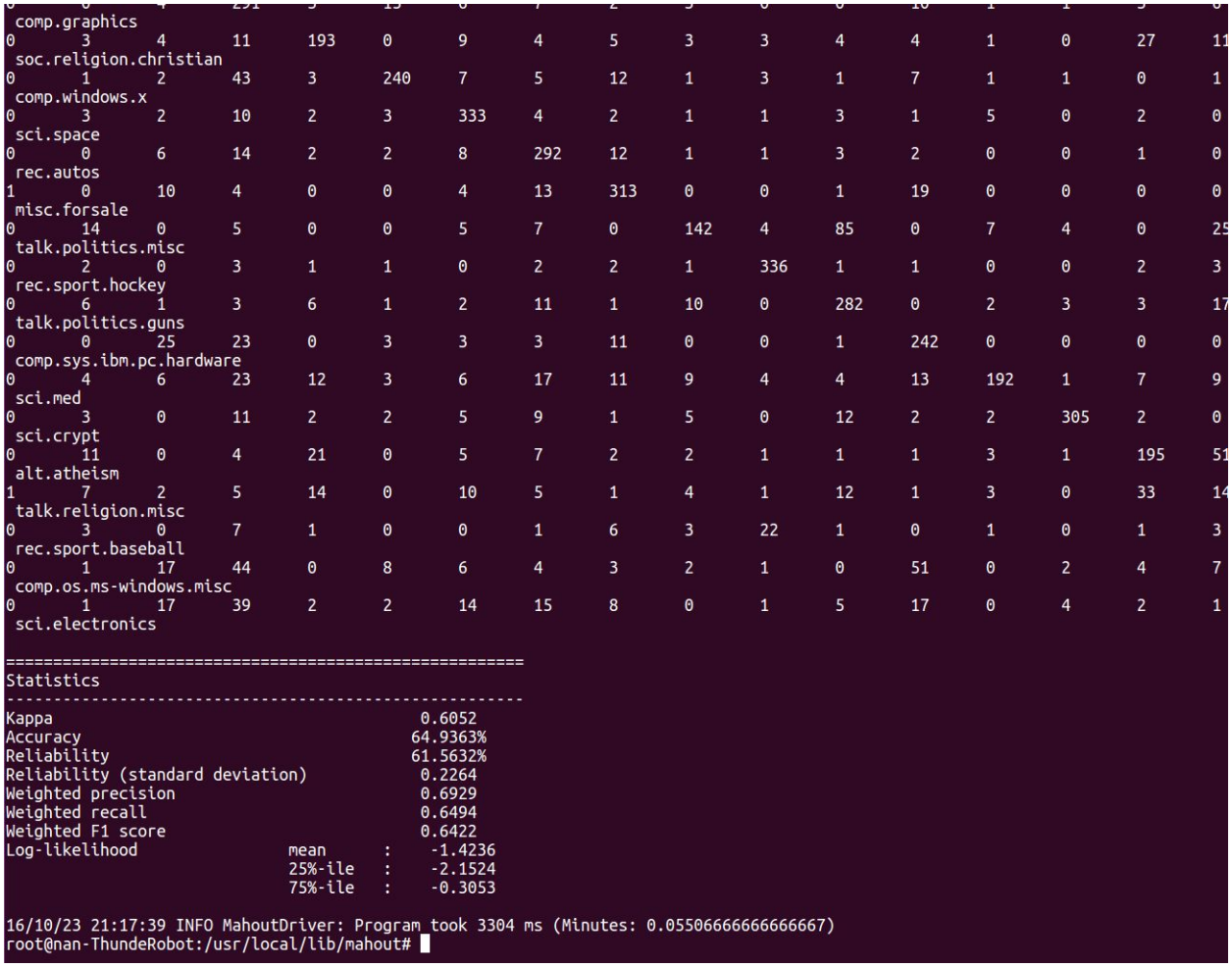
```
./bin/mahout trainnb \
  -i ${WORK_DIR}/20news-train-vectors \
  -o ${WORK_DIR}/model \
```

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

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

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



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



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.

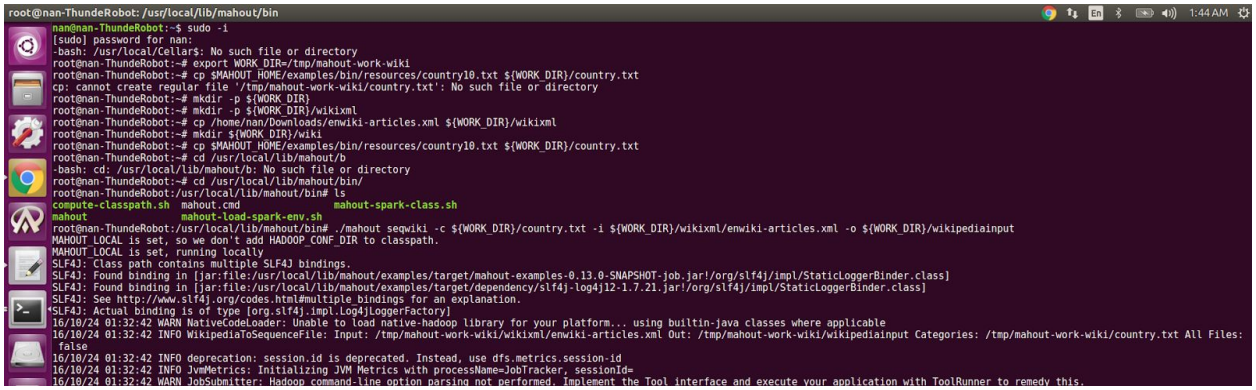
3-2 Classification on wiki data

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

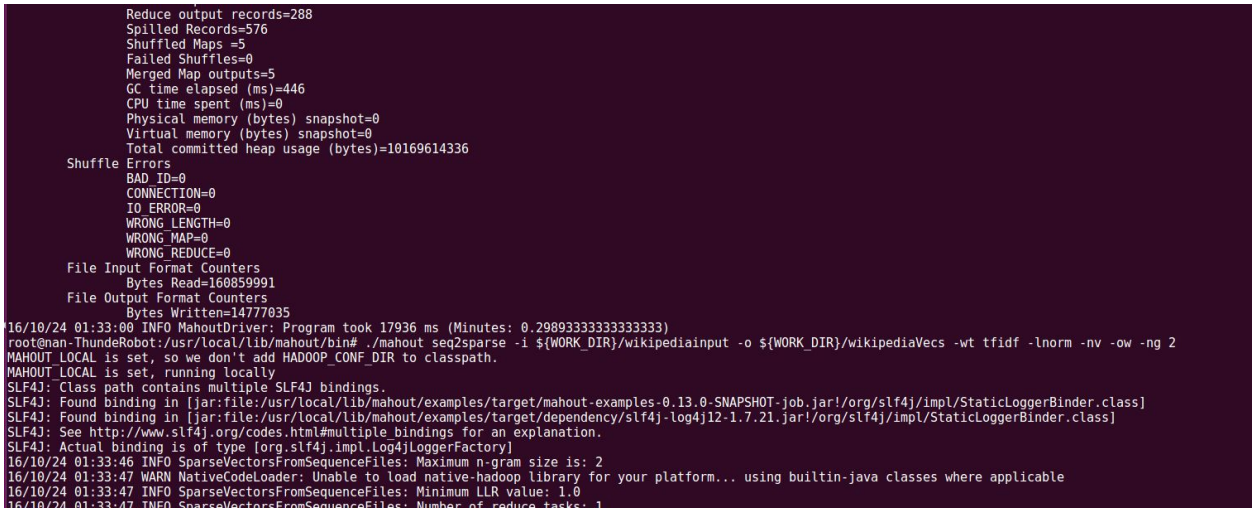
mkdir ${WORK_DIR}/wiki
root@nan-ThundeRobot:/tmp/mahout-work-wiki# cp $MAHOUT_HOME/examples/bin/resources/country10.txt ${WORK_DIR}/country.txt

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



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



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

```
                Reduce 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
                WRONG_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 50299 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 s
equential
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, --randomSe
lectionPct=[20], --sequenceFiles=null, --startPhase=[0], --tempDir=[temp], --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: File /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], --overw
ionPct=[20], --sequenceFiles=null, --startPhase=[0], --tempDir=[temp], --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 SplitInput: file: part-r-00000, input: 130607 train: 227, test: 61 starting at 0
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: {--alphaI=[1.0], --endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/training], --labelIndex=[/tmp/mahout-work-wi
/tmp/mahout-work-wiki/model], --overwrite=null, --startPhase=[0], --tempDir=[temp], --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

```
                Combine input records=1
                Combine 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: 0.10528333333333334)
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
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 01:38:55 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.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
16/10/24 01:38:56 INFO deprecation: mapred.compress.map.output is deprecated. Instead, use mapreduce.map.output.compress
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: session.id is deprecated. Instead, use dfs.metrics.session-id
```

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                                90.9091%
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: 0.04901666666666667)
```

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, 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 01:56:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/24 01:56:49 INFO WikipediaToSequenceFile: Input: /tmp/mahout-work-wiki/wikixml/enwiki-articles.xml Out: /tmp/mahout-work-wiki/wikipediainput Categories: /tmp/mahout-work-wiki/country.txt All Fil
 false
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=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=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 01:57:40 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/wikipediaVecs
16/10/24 01:57:41 INFO SparseVectorsFromSequenceFiles: Minimum LLR value: 1.0
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=
16/10/24 01:57:41 INFO FileInputFormat: Total input paths to process : 1
```

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=0
                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=52268806
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 158009 ms (Minutes: 2.6334833333333334)
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 seque
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 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], --overwrite=null, --randomSelect
ionPct=[20], --sequenceFiles=null, --startPhase=[0], --tempDir=[temp], --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
```

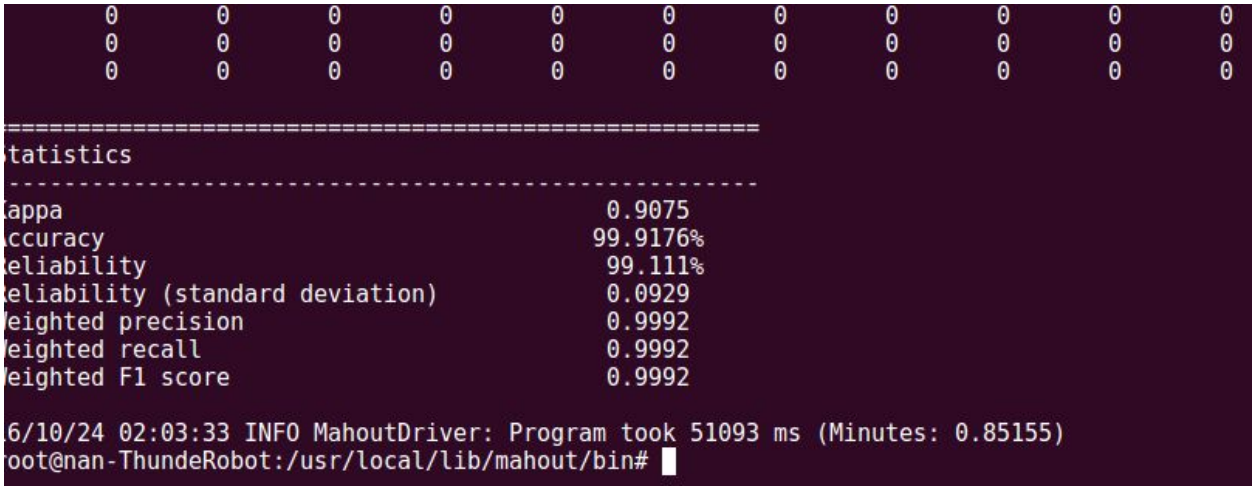
root@nan-ThundeRobot:/usr/local/lib/mahout/bin# ./mahout split -i ${WORK_DIR}/wikipediaVecs/tfidf-vectors/ --trainingOutput ${WORK_DIR}/training --testOutput ${WORK_DIR}/testing -rp 20 -ow -seq -xm sequential

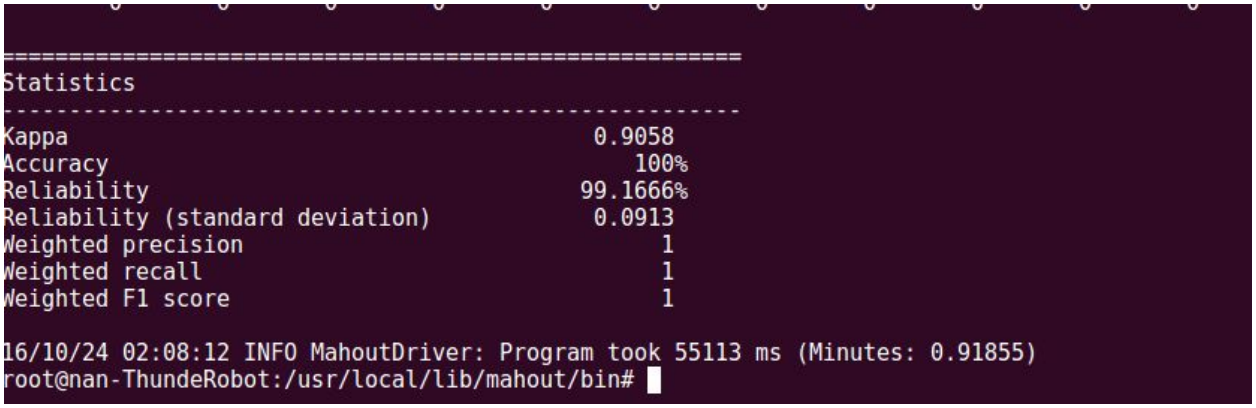
```
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/10/24 02:01:13 WARN MahoutDriver: No split.props found on classpath, will use command-line arguments only
16/10/24 02:01:13 INFO AbstractJob: Command line arguments: {--endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/wikipediaVecs/tfidf-vectors/], --method=[sequential], --o
ionPct=[20], --sequenceFiles=null, --startPhase=[0], --tempDir=[temp], --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: 0.09376666666666666)
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: {--alphaI=[1.0], --endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/training], --labelIndex=[/tmp/mahout-work
/tmp/mahout-work-wiki/model], --overwrite=null, --startPhase=[0], --tempDir=[temp], --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





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

```
nan@nan-ThundeRobot:/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-ThundeRobot 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

```
>>> textFile = sc.textFile("/home/nan/Downloads/enwiki-articles.xml")
>>> textFile.filter(lambda line:"great" in line).count()
6582
>>> textFile.count()
1226682
>>>
```

```
>>> word_Counts = textFile.flatMap(lambda line:line.split()).map(lambda word:(word,1)).reduceByKey(lambda a, b:a+b)
>>> word_Counts.collect()
```


```
,(u'Cain/Gay&amp;nbsp;\u2013', 1), (u'|pages=275\u201388', 1), (u'South)]', 46), (u'*[http://www.sbarnabas.com', 1), (u"''[[wicket", 1), (u'book|last=Boyer|first=Ed.', 1), (u'&quot;Cynthiae', 1), (u'Xandre', 1), (u'ligase.&lt;ref', 1), (u'2013|pages=40\u201343|url=http://www.newscientist.com/articleimages/mg21929350.800/1-whats-brown-and-soggy-and-could-save-the-world.html|author=Rachel', 1), (u'order]]s,', 2), (u'municipality.&lt;ref&gt;[http://www.ajaccio.fr', 1), (u'beyond,&quot;', 2), (u'[http://www.census.gov/Press-Release/www/2000/cb00-151.html', 1), (u"{''strictı", 1), (u'Robert]].', 8), (u"Telecommunications''.", 1), (u'Topography.png|thumb|left|Topography', 1), (u'Norri', 2), (u'|page=1407', 1), (u'\xc6thelberht}}', 1), (u'[[Celine', 2), (u'{[Scripting', 4), (u'(CBN).', 1), (u'Ostindische', 1), (u'name=&quot;bind10-design&quot;&gt;{{cite', 1), (u'[[Glycogenolysis|breakdown', 1), (u'Court|UPC', 5), (u'lautaroj', 2), (u'Isles]];', 1), (u'25,000,000', 1), (u'Suburban,', 1), (u'Gelenberg', 1), (u'{{commons|Category:Celestine', 1), (u'[[war]]fare', 1), (u'States&lt;br&gt;China', 1), (u'camelCase', 2), (u'0-387-98431-3', 1), (u'Delporte|Eug\xe8ne', 9), (u'(manager)|Joe', 5), (u'al-B\u012br\u016bn\u012b}}&lt;/ref&gt;', 1), (u'|url=http://www.icrc.org/customary-ihl/eng/docs/v2_rul_rule76', 2), (u'|url=http://geo.international.gc.ca/cip-pic/geo/ethiopia-en.aspx', 1), (u'[http://freemarketmojo.com/?p=5539', 2), (u'|first16=M.', 1), (u'America|work=Greater', 1), (u'|align=&quot;left&quot;|[[Fuerteventura]]', 1), (u'[[Saeed', 10), (u'*(1788)', 1), (u'1032-1048:', 1), (u'{{convert|30|in|cm|abbr=on}}.&lt;ref&gt;[http://www.extremetech.com/computing/130605-where-are-all-the-high-resolution-desktop-displays', 1), (u'Serbian]].', 1), (u'[[Versoix]].', 1), (u'thought.&lt;ref&gt;Richard', 1), (u'[[File:Item-ceb-4-1296', 2), (u'|work=Euromoney', 1), (u'used&quot;.&lt;ref', 1), (u'|page=581203}}&lt;/ref&gt;', 1), (u'&quot;ice', 4), (u"{{math|&amp;nbsp;''f&lt;sub&gt;i&lt;/sub&gt;''", 2), (u'EurasiaNet.org.', 1), (u'[[Khalifa', 1)]
```