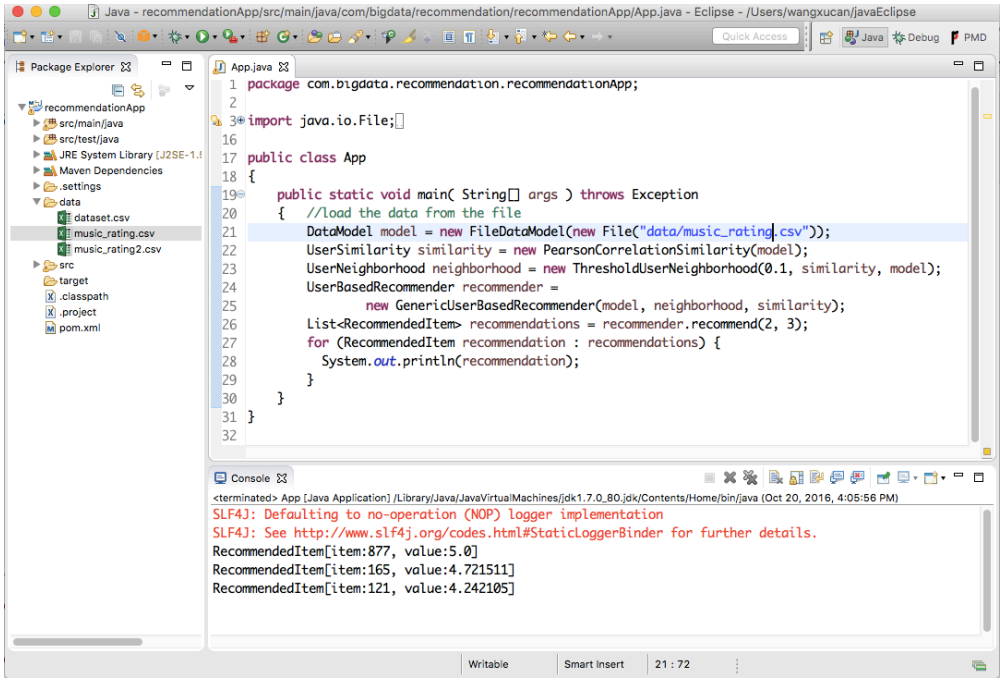
Big data analytics **Assignment 2**

Name: Xucan Wang **Uni: xw2401**

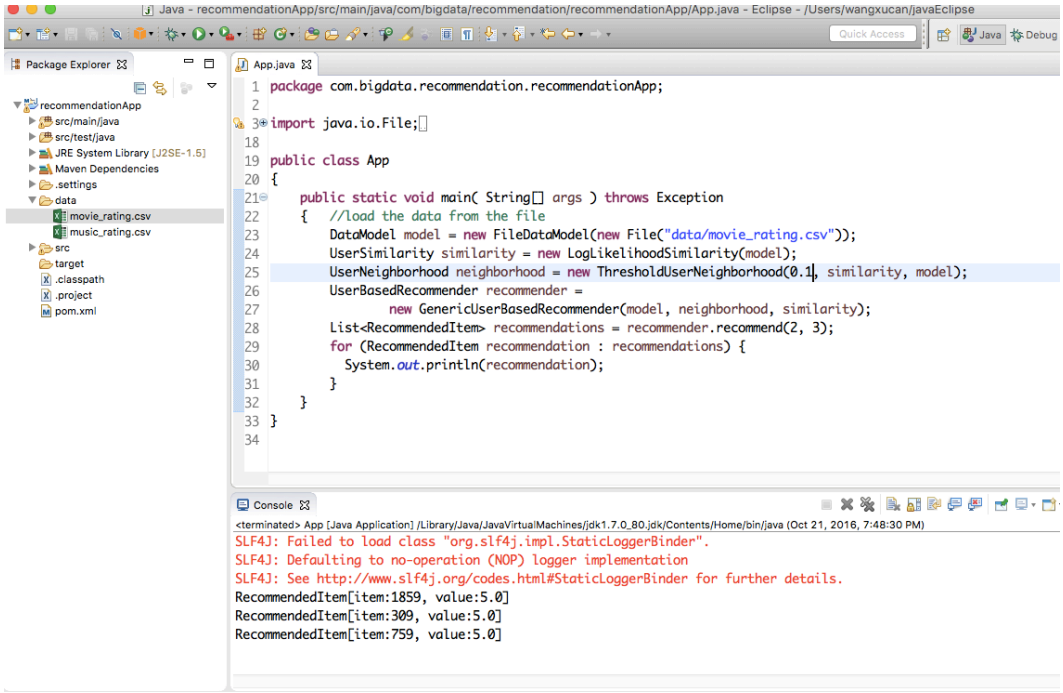
Part 1: Recommendation

There are three parts for the recommendation. First is the DataModel which is created from .csv file. The Csv file needs to contain the userId, itemId and the ratings for a specific item for the user. The UserSimilarity defines the similarity between users, which is used to get the neighborhoods of a specificuser. The UserNeighborhood computes the neighborhood of a user given a user. With DataModel, UserSimiliarity and UserNeighborhood we can create the UserBasedRecommender object, which is then used to make recommendations to the user.

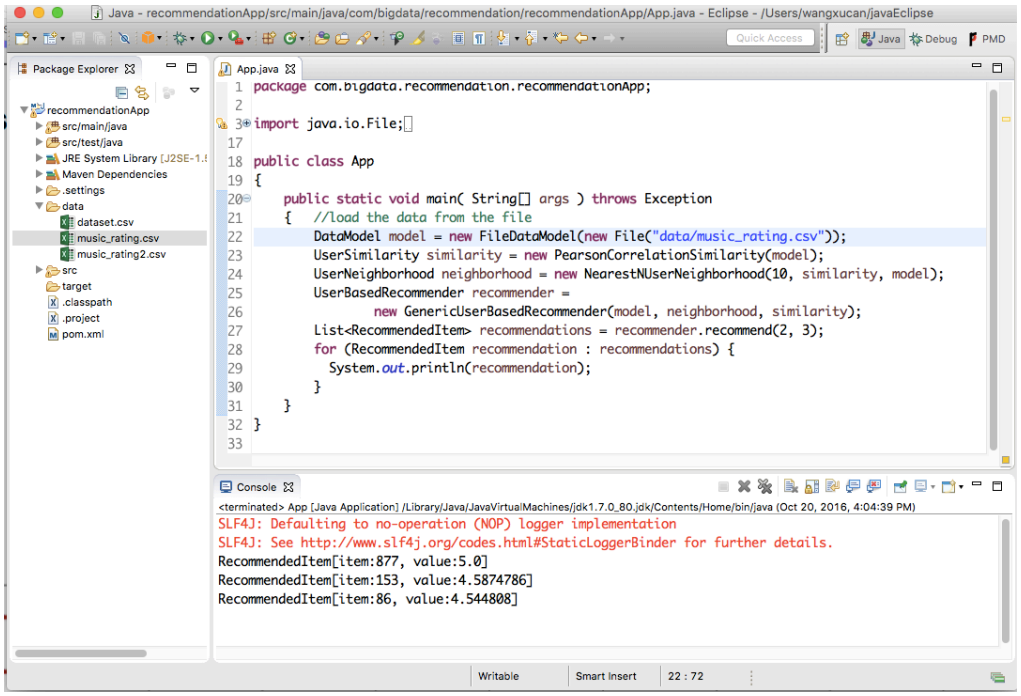
1) PearsonCorrelationSimilarity & ThresholdUserNeignborhood



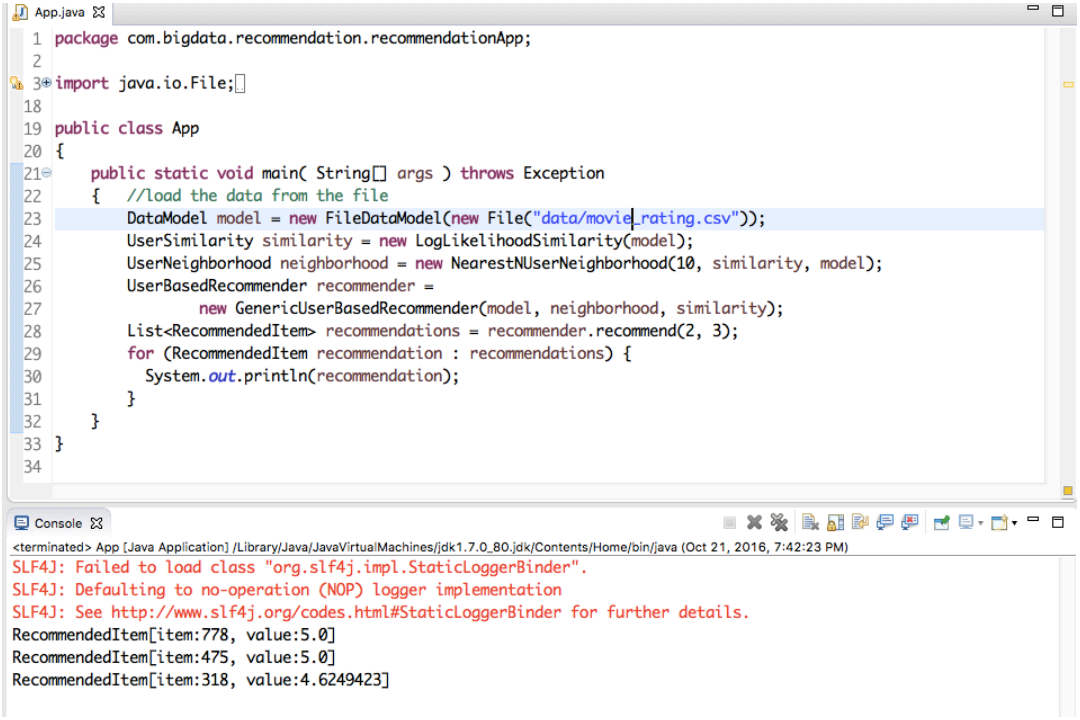
2) LogLikelihood & ThresholdUserNeignborhood



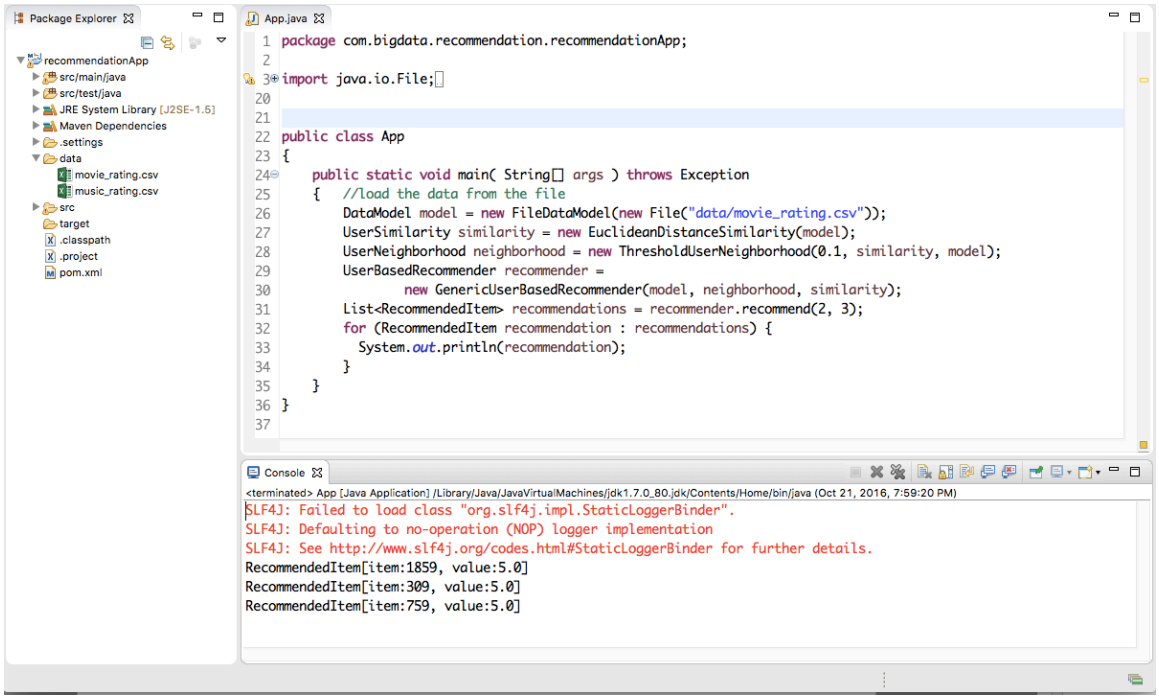
3) PearsonCorrelationSimilarity&NearestNUserNeighborhood



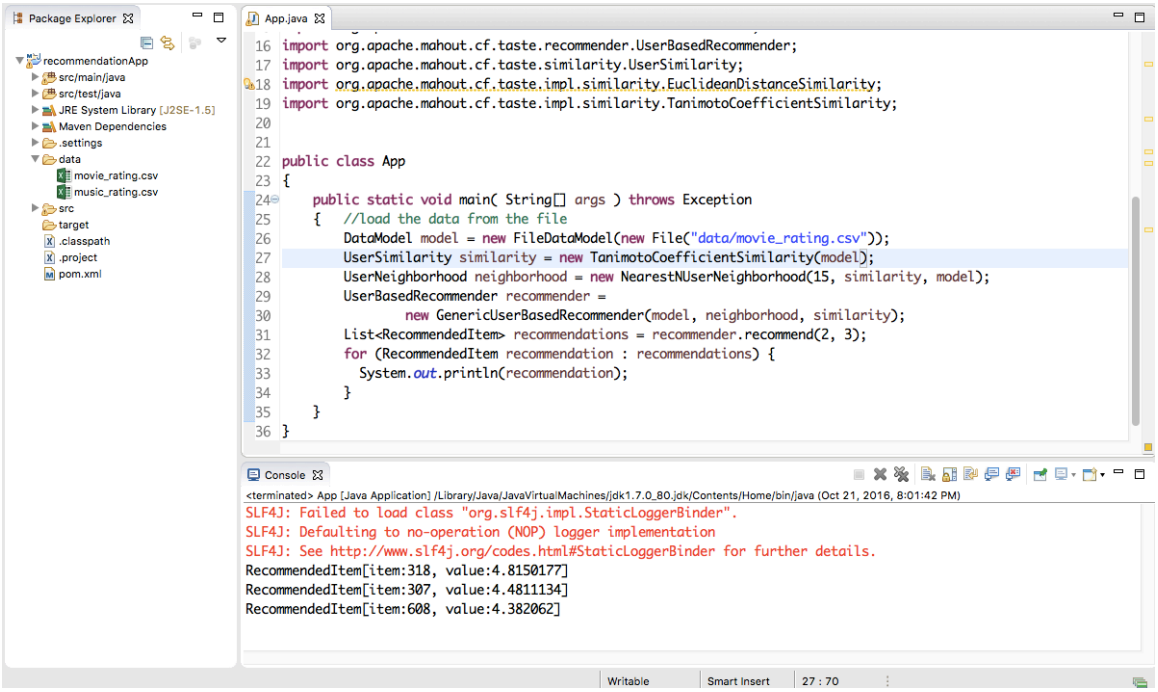
4) Loglikelihood & NearestNUserNeighborhood



5)EuclideanDistanceSimilarity & ThresholdUserNeignborhood



6) TanimotoCoefficientSimilarity & NearestNUserNeighborhood



As we can see, we use 2 datasets in the recommender. One is the music_rating.csv, which includes 425 users and their ratings to different songs. The dataset includes about 10102 user-supplied ratings. Another is movie_rating.csv, which includes 100,000 ratings from 700 users. And the algorithm generates top 3 recommended item for the user with ID 2.

Part 2: Clustering

In this part, I use the data from the example tutorial and from the reuters news

(1) Reuters News:

a. K-means

In this part, I use the "cluster-reuters.sh" to do the clustering job.

At first, when I run the command, one error occurred. Then I create the directory using hdfs dfs, then the problem disappeared.

```
16/10/20 23:08:09 INFO JobSubmitter: Cleaning up the staging area file:/tmp/hadoop-wangxucan/mapred/staging/wangxucan1501208986/.staging/job_local1501208986_0001
Exception in thread "main" org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hdfs://localhost:9000/tmp/mahout-work-wangxucan/reuters-out-seqdir
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFormat.java:323)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:265)
        at org.apache.hadoop.mapreduce.lib.input.SequenceFileInputFormat.listStatus(SequenceFileInputFormat.java:59)
```

```
[wangxucan (master) mahout-trunk $ ./examples/bin/cluster-reuters.sh                                                ]
Discovered Hadoop v2.
Setting dfs command to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs, dfs rm to /usr/local/Cellar/hadoop/2.7.3/bin/hd
fs dfs -rm -r -skipTrash.
Please select a number to choose the corresponding clustering algorithm
1. kmeans clustering (runs from this example script in cluster mode only)
2. fuzzykmeans clustering (may require increased heap space on yarn)
3. lda clustering
4. streamingkmeans clustering
5. clean -- cleans up the work area in /tmp/mahout-work-wangxucan
Enter your choice : 1
```

```
                Bytes Written=28869
16/10/21 11:53:57 INFO ClusterEvaluator: Scaled Inter-Cluster Density = 0.45931222023923834
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[384] = 0.6196735100806415
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[316] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[963] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[788] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[1237] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[794] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[734] = 0.5666666666666664
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[1455] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[1283] = 0.5666666666666668
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[932] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[389] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[452] = 0.6905822623826653
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[1142] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[108] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[950] = 0.5666666666666668
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[1229] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[1449] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[1000] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[1499] = 0.5906098266391152
16/10/21 11:53:57 INFO ClusterEvaluator: Intra-Cluster Density[19] = NaN
16/10/21 11:53:57 INFO ClusterEvaluator: Average Intra-Cluster Density = 0.6001442665170703
16/10/21 11:53:57 INFO ClusterDumper: Wrote 20 clusters
16/10/21 11:53:57 INFO MahoutDriver: Program took 10854 ms (Minutes: 0.1809)
:{"identifier":"CL-384","r":[{"0.2":0.513},{"0.5":0.692},{"0.6":0.768},{"0.7":0.543},{"0.75":0.765},{
        Top Terms:
                he                                      =>  3.9134248570128753
                said                                    =>  3.6165204931806016
                would                                   =>  3.5078525243105587
                have                                    =>   3.324054274525676
                last                                    =>  2.8937940247408993
                from                                    =>  2.5793220863475668
                some                                    =>  2.5299734569096066
                billion                                 =>  2.5207431616483036
                which                                   =>   2.497113834727894
                banks                                   =>  2.4121890968376105
                pct                                     =>  2.3752352791232663
                more                                    =>  2.3710836330493845
                its                                     =>  2.3143290189596324
                market                                  =>   2.308005894814338
                has                                     =>   2.281259171612613
                u.s                                     =>  2.2577419564440535
                year                                    =>   2.235641827950111
                about                                   =>   2.210096547653625
                we                                      =>   2.197323724106475
                been                                    =>  2.1645465997549205
        Weight : [props - optional]:  Point:

:{"identifier":"VL-316","r":[],"c":[{"08":4.031},{"2":1.951},{"29.53":7.363},{"55":4.724},{"action":4
        Top Terms:
                ilc                                     =>  13.455424308776855
                montoya                                 =>  10.986308097839355
                disbursements                           =>  10.986308097839355
```

```
:{"identifier":"VL-1499","r":[{"0":0.438},{"0.01":0.263},{"0.077":0.278},{"0.1":0.456},{"0.2":0.492},
        Top Terms:
                said                                    =>  1.4483192144182537
                mln                                     =>   1.278848085702802
                dlrs                                    =>  1.1386757567477181
                vs                                      =>   1.073048673174001
                pct                                     =>  0.9123013656464698
                cts                                     =>  0.9051320905642659
                its                                     =>  0.8622784409898737
                from                                    =>   0.847106046053479
                year                                    =>  0.8291803765648226
                net                                     =>  0.7893499478860058
                inc                                     =>  0.7836996392513376
                2                                       =>  0.7777500656611792
                corp                                    =>  0.7185749607464963
                company                                 =>  0.7084198830449986
                4                                       =>  0.6160361142131016
                has                                     =>  0.6152814222710932
                shr                                     =>  0.5927909036082842
                u.s                                     =>  0.5921665951657341
                1986                                    =>   0.569558723548679
                loss                                    =>  0.5630149809234345
        Weight : [props - optional]:  Point:

:{"identifier":"VL-19","r":[],"c":[{"16":2.834},{"26":2.816},{"48":4.724},{"buy":5.512},{"common":3.5
        Top Terms:
                sterling                                =>   8.452377319335938
                software                                =>    7.59517240524292
                discretion                              =>   7.363028049468994
                consent                                 =>   7.075345993041992
                ok                                      =>   7.075345993041992
                holders                                 =>  6.6149139404296875
                depending                               =>   5.822583198547363
                buy                                     =>   5.511603355407715
                majority                                =>   5.242764472961426
                required                                =>   5.060442924499512
                debentures                              =>   4.965132713317871
                subordinated                            =>   4.935279846191406
                conditions                              =>   4.798078536987305
                48                                      =>   4.723970890045166
                convertible                             =>   4.590439319610596
                received                                =>   4.334506034851074
                eight                                   =>    4.13090705871582
                purchase                                =>    4.11783504486084
                its                                     =>  3.9509592056274414
                common                                  =>   3.571291208267212
        Weight : [props - optional]:  Point:


Inter-Cluster Density: 0.45931222023923834
Intra-Cluster Density: 0.6001442665170703
CDbw Inter-Cluster Density: 0.0
CDbw Intra-Cluster Density: 18.84288640124127
CDbw Separation: 28011.919431283415
wangxucan (master) mahout-trunk $
```

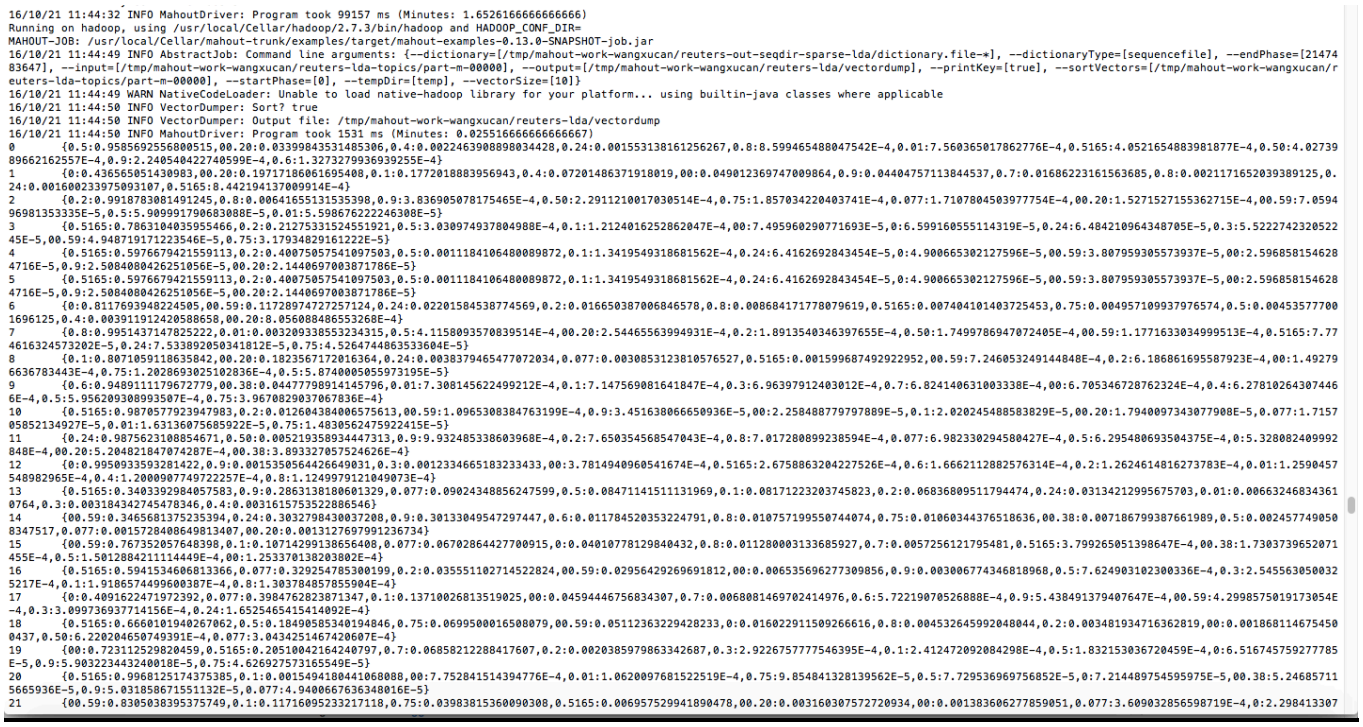

b. Lda clustering

It takes more time than the k-means clustering method. The K-means takes about 1 minute and the Lda clustering takes about 3 minutes. So the Lda clustering takes more time

result:

/tmp/mahout-work-wangxucan Go!

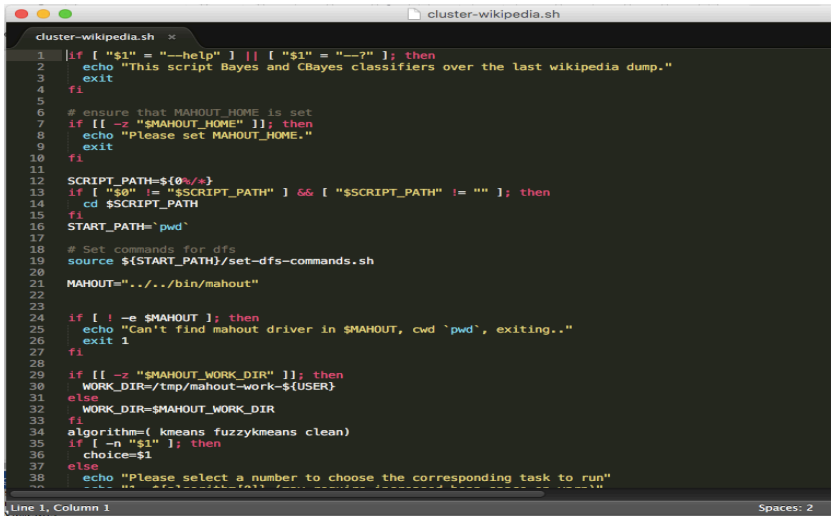
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 11:44:27 AM EDT	0	0 B	reuters-lda
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 11:44:22 AM EDT	0	0 B	reuters-lda-model
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 11:44:31 AM EDT	0	0 B	reuters-lda-topics
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 11:37:54 AM EDT	0	0 B	reuters-out
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 11:42:39 AM EDT	0	0 B	reuters-out-matrix
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 11:41:40 AM EDT	0	0 B	reuters-out-seqdir
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 11:42:24 AM EDT	0	0 B	reuters-out-seqdir-sparse-lda
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 11:37:51 AM EDT	0	0 B	reuters-sgm

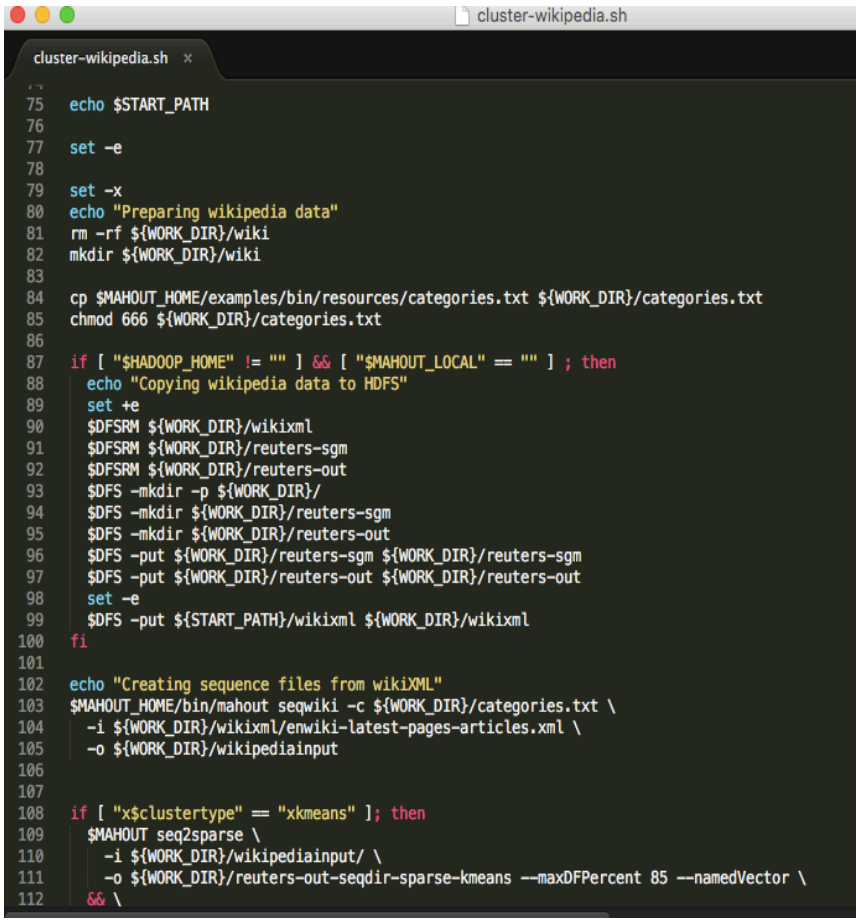


```
16/10/21 11:44:32 INFO MahoutDriver: Program took 99157 ms (Minutes: 1.6526166666666666)
Running on hadoop, using /usr/local/Cellar/hadoop/2.7.3/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/Cellar/mahout-trunk/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar
16/10/21 11:44:49 INFO AbstractJob: Command line arguments: {--dictionary=[/tmp/mahout-work-wangxucan/reuters-out-seqdir-sparse-lda/dictionary.file-*], --dictionaryType=[sequencefile], --endPhase=[2147483647], --input=[/tmp/mahout-work-wangxucan/reuters-lda-topics/part-m-00000], --output=[/tmp/mahout-work-wangxucan/reuters-lda/vectordump], --printKey=[true], --sortVectors=[/tmp/mahout-work-wangxucan/reuters-lda-topics/part-m-00000], --startPhase=[0], --tempDir=[temp], --vectorSize=[10]}
16/10/21 11:44:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/21 11:44:50 INFO VectorDumper: Sort? true
16/10/21 11:44:50 INFO VectorDumper: Output file: /tmp/mahout-work-wangxucan/reuters-lda/vectordump
16/10/21 11:44:50 INFO MahoutDriver: Program took 1531 ms (Minutes: 0.025516666666666667)
0       {0.5:0.9585692556800515,00.20:0.03399843531485306,0.4:0.0022463908898034428,0.24:0.0015531381612562 67,0.8:8.599465488047542E-4,0.01:7.5603650178627766E-4,0.5165:4.052165488398187 7E-4,0.50:4.02739896621625 57E-4,0.9:2.2405404227405 99E-4,0.6:1.3273279936939255E-4}
1       {0:0.4365650514309 83,00.20:0.19717186061695408,0.1:0.17720188839569 43,0.4:0.07201486371918019,00:0.04901236974709864,0.9:0.04404757113844537,0.7:0.01686223161563685,0.8:0.00211716520393891 25,0.24:0.00160023397509 3107,0.5165:8.442194137009914E-4}
2       {0.2:0.9918783081491245,0.8:0.006461551313535398,0.9:3.836905078175465E-4,0.50:2.2911210017030514E-4,0.75:1.857034220403741E-4,0.077:1.7107804503977754E-4,00.20:1.5271527155362715E-4,00.59:7.0594969813533355E-5,0.5:5.90999179068308 8E-5,0.01:5.5986762222463 08E-5}
3       {0.5165:0.7863104035955466,0.2:0.21275331524551921,0.5:3.030974937804988E-4,0.1:1.2124016252862047E-4,00:7.495960290771693E-5,0:6.599160555114319E-5,0.24:6.484210964348705E-5,0.3:5.522274232052245E-5,00.59:4.948719171223546E-5,0.75:3.17934829161222E-5}
4       {0.5165:0.5976679421559113,0.2:0.40075057541097503,0.5:0.001118410648008 9872,0.1:1.3419549318681562E-4,0.24:6.4162692843454E-5,0:4.900665302127596E-5,00.59:3.8079593055739 37E-5,00:2.5968581546284716E-5,0.9:2.5084080426251056E-5,00.20:2.144069700387178 6E-5}
5       {0.5165:0.5976679421559113,0.2:0.40075057541097503,0.5:0.001118410648008 9872,0.1:1.3419549318681562E-4,0.24:6.4162692843454E-5,0:4.900665302127596E-5,00.59:3.8079593055739 37E-5,00:2.5968581546284716E-5,0.9:2.5084080426251056E-5,00.20:2.144069700387178 6E-5}
6       {0:0.8117693948224505,00.59:0.11720974727257124,0.24:0.022015845 38774569,0.2:0.016650387006846578,0.8:0.008684171778079619,0.5165:0.007404101403725453,0.75:0.004957109937976574,0.5:0.004535777001696125,0.4:0.0039119124205 88658,00.20:8.056088486553268E-4}
7       {0.8:0.9951437147825222,0.01:0.003209338553234315,0.5:4.11580935708395 14E-4,00.20:2.544655639949 31E-4,0.2:1.8913540346397655E-4,0.50:1.7499786947072405E-4,00.59:1.1771633034999513E-4,0.5165:7.774616324573202E-5,0.24:7.533892050341812E-5,0.75:4.526474486353360 4E-5}
8       {0.1:0.8071059118635842,00.20:0.1823567172016364,0.24:0.003837946547707 2034,0.077:0.0030853123810576527,0.5165:0.001599687492922952,00.59:7.246053249144848E-4,0.2:6.186861695587923E-4,00:1.4927966367834 3E-4,0.75:1.2028693025102836E-4,0.5:5.8740005055973195E-5}
9       {0.6:0.9489111179672779,00.38:0.044777989141457 96,0.01:7.308145622499212E-4,0.1:7.1475690816418 47E-4,0.3:6.963979124030212E-4,0.7:6.824140631003338E-4,00:6.705346728762324E-4,0.4:6.278102643074466E-4,0.5:5.956209308993507E-4,0.75:3.967082903706783 6E-4}
10      {0.5165:0.9870577923947983,0.2:0.012604384006575613,00.59:1.0965308384763199E-4,0.9:3.4516380666506 36E-5,00:2.258488779797889E-5,0.1:2.0202454885838 29E-5,00.20:1.7940097343077908E-5,0.077:1.71570585213492 7E-5,0.01:1.631360756859 22E-5,0.75:1.4830562475922415E-5}
11      {0.24:0.9875623108854671,0.50:0.005219358934447313,0.9:9.9324853386039 68E-4,0.2:7.650354568547043E-4,0.8:7.017280899238594E-4,0.077:6.982330294580427E-4,0.5:6.295480693504375E-4,0.5:5.328082409992848E-4,00.20:5.204821847074287E-4,00.38:3.893327057524626E-4}
12      {0:0.9950933593281422,0.9:0.0015350564426649031,0.3:0.0012334665183233433,00:3.781494096054167 4E-4,0.5165:2.675886320427752 6E-4,0.6:1.6662112882576314E-4,0.2:1.2624614816273783E-4,0.01:1.2590457548982965E-4,0.4:1.2000907749722257E-4,0.8:1.1249979121049073E-4}
13      {0.5165:0.3403392984057583,0.9:0.2863138180601329,0.077:0.0902434885624 7599,0.5:0.08471141511131969,0.1:0.08171223203745823,0.2:0.06836809511794474,0.24:0.03134212995675703,0.01:0.00663246834361076 4,0.3:0.003184342745478346,0.4:0.0031615753522886546}
14      {00.59:0.34656813752353 94,0.24:0.3032798430037208,0.9:0.30133049547297447,0.6:0.011784520353224791,0.8:0.010757199550744074,0.75:0.01060344376518636,00.38:0.007186799387661989,0.5:0.0024577490508347517,0.077:0.0015728408649813407,00.20:0.0013127697991236734}
15      {00.59:0.7673520576483 98,0.1:0.10714299138656408,0.077:0.06702864427700915,0:0.040107781298404 32,0.8:0.011280003133685927,0.7:0.0057256121795481,0.5165:3.799265051398647E-4,00.38:1.7303739652071455E-4,0.5:1.5012884211114449E-4,00:1.2533701382038 02E-4}
16      {0.5165:0.5941534606813366,0.077:0.329254785300199,0.2:0.035551102714522824,00.59:0.02956429269691812,00:0.0065356962773 09856,0.9:0.003006774346818968,0.5:7.624903102300336E-4,0.3:2.5455630500325217E-4,0.1:1.9186574499600387E-4,0.8:1.3037848578559 04E-4}
17      {0:0.4091622471972392,0.077:0.3984762823871347,0.1:0.13710026813519025,00:0.045944467568343 07,0.7:0.006808146970241 4976,0.6:5.722190705268 88E-4,0.9:5.438491379407647E-4,00.59:4.299857501917305 4E-4,0.3:3.099736937714156E-4,0.24:1.6525465415414092E-4}
18      {0.5165:0.6660101940267062,0.5:0.1849058534019484 6,0.75:0.0699500016508079,00.59:0.05112363229428233,0:0.016022911509266616,0.8:0.004532645992048044,0.2:0.0034819347163628 19,00:0.0018681146754500437,0.50:6.220204650749391E-4,0.077:3.043425146742060 7E-4}
19      {00:0.723112529820459,0.5165:0.20510042164240797,0.7:0.06858212288417607,0.2:0.0020385979863342687,0.3:2.9226757777546395E-4,0.1:2.4124720920842 98E-4,0.5:1.8321530367204 59E-4,0:6.516745759277785E-5,0.9:5.903223443240018E-5,0.75:4.626927573165549E-5}
20      {0.5165:0.9968125174375385,0.1:0.0015494180441068088,00:7.752841514394776E-4,0.01:1.0620097681522519E-4,0.75:9.854841328139562E-5,0.5:7.729536969756852E-5,0:7.214489754595975E-5,00.38:5.246857115665936E-5,0.9:5.031858671551132E-5,0.077:4.940066763634806E-5}
21      {00.59:0.8305038395375749,0.1:0.1171609523321711 8,0.75:0.0398381536009030 8,0.5165:0.006957529941890478,00.20:0.0031603075727209 34,00:0.001383606277859051,0.077:3.609032856598719E-4,0:2.298413307
```

(2) Wikipedia article

I changed some of the content in the cluster-reuters.sh and made the cluster-wikipedia.sh to do clustering on the Wikipedia articles, here is the screen shot of my script. At first I messed up with the path thing, so I just make some change to the cluster-reuters.sh without changing the path name but the file we used is the wiki file. You can see from the script.


cluster-wikipedia.sh
cluster-wikipedia.sh
if ["$1" = "--help"] || ["$1" = "--?"]; then
echo "This script Bayes and CBayes classifiers over the last wikipedia dump."
exit
fi
# ensure that MAHOUT_HOME is set
if [[-z "$MAHOUT_HOME"]]; then
echo "Please set MAHOUT_HOME."
exit
fi
SCRIPT_PATH=${0%/*}
if ["$0" != "$SCRIPT_PATH"] && ["$SCRIPT_PATH" != ""]; then
cd $SCRIPT_PATH
fi
START_PATH=`pwd`
# Set commands for dfs
source ${START_PATH}/set-dfs-commands.sh
MAHOUT="../../bin/mahout"
if [! -e $MAHOUT]; then
echo "Can't find mahout driver in $MAHOUT, cwd `pwd`, exiting.."
exit 1
fi
if [[-z "$MAHOUT_WORK_DIR"]]; then
WORK_DIR=/tmp/mahout-work-${USER}
else
WORK_DIR=$MAHOUT_WORK_DIR
fi
algorithm=(kmeans fuzzykmeans clean)
if [-n "$1"]; then
choice=$1
else
echo "Please select a number to choose the corresponding task to run"
Line 1, Column 1
Spaces: 2

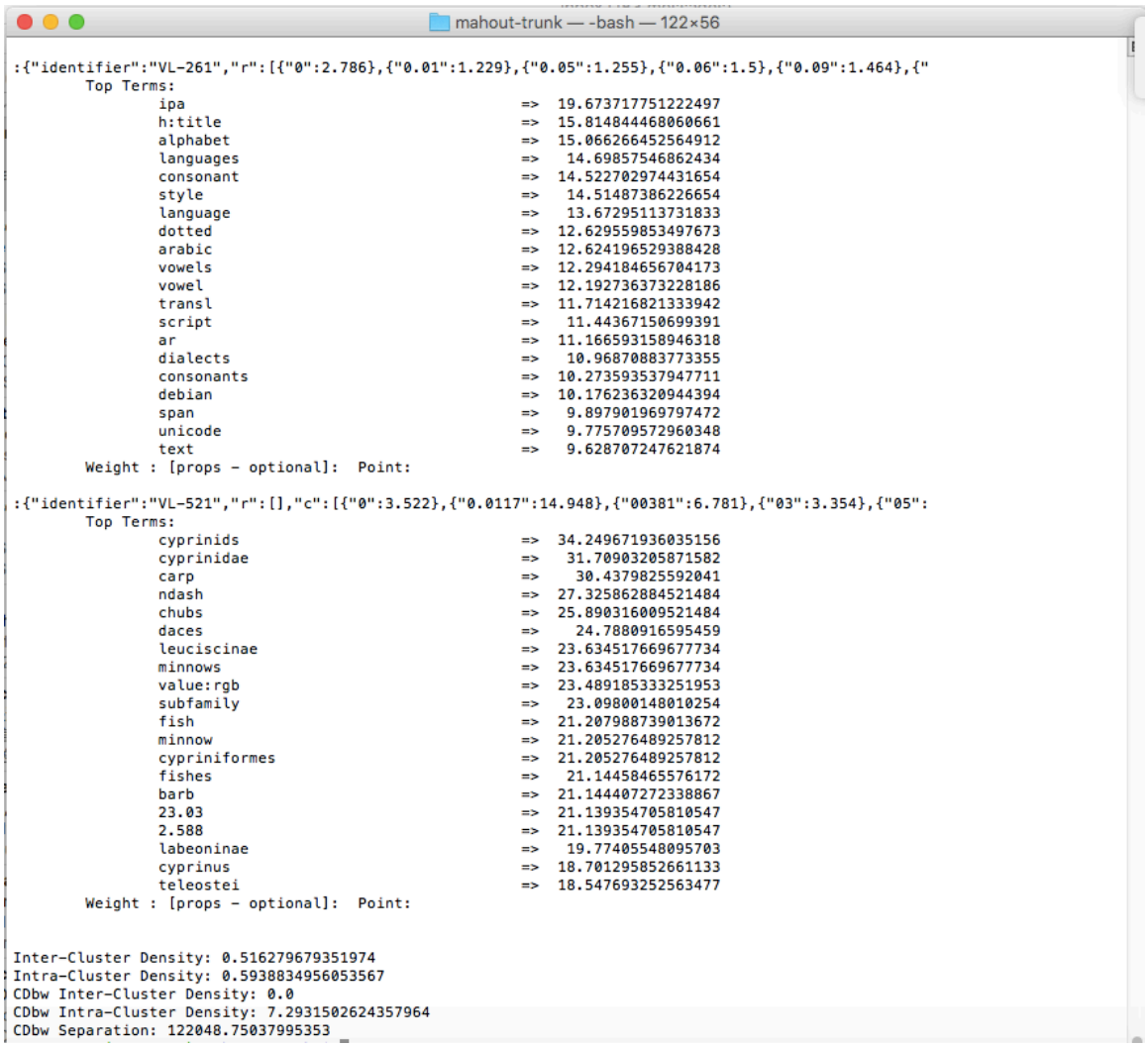

cluster-wikipedia.sh
cluster-wikipedia.sh
echo $START_PATH
set -e
set -x
echo "Preparing wikipedia data"
rm -rf ${WORK_DIR}/wiki
mkdir ${WORK_DIR}/wiki
cp $MAHOUT_HOME/examples/bin/resources/categories.txt ${WORK_DIR}/categories.txt
chmod 666 ${WORK_DIR}/categories.txt
if ["$HADOOP_HOME" != ""] && ["$MAHOUT_LOCAL" == ""] ; then
echo "Copying wikipedia data to HDFS"
set +e
$DFSRM ${WORK_DIR}/wikixml
$DFSRM ${WORK_DIR}/reuters-sgm
$DFSRM ${WORK_DIR}/reuters-out
$DFS -mkdir -p ${WORK_DIR}/
$DFS -mkdir ${WORK_DIR}/reuters-sgm
$DFS -mkdir ${WORK_DIR}/reuters-out
$DFS -put ${WORK_DIR}/reuters-sgm ${WORK_DIR}/reuters-sgm
$DFS -put ${WORK_DIR}/reuters-out ${WORK_DIR}/reuters-out
set -e
$DFS -put ${START_PATH}/wikixml ${WORK_DIR}/wikixml
fi
echo "Creating sequence files from wikiXML"
$MAHOUT_HOME/bin/mahout seqwiki -c ${WORK_DIR}/categories.txt \
-i ${WORK_DIR}/wikixml/enwiki-latest-pages-articles.xml \
-o ${WORK_DIR}/wikipediainput
if ["x$clustertype" == "xkmeans"]; then
$MAHOUT seq2sparse \
-i ${WORK_DIR}/wikipediainput/ \
-o ${WORK_DIR}/reuters-out-seqdir-sparse-kmeans --maxDFPercent 85 --namedVector \
&& \

cluster-wikipedia.sh

```
  && \
  $MAHOUT kmeans \
    -i ${WORK_DIR}/reuters-out-seqdir-sparse-kmeans/tfidf-vectors/ \
    -c ${WORK_DIR}/reuters-kmeans-clusters \
    -o ${WORK_DIR}/reuters-kmeans \
    -dm org.apache.mahout.common.distance.EuclideanDistanceMeasure \
    -x 10 -k 20 -ow --clustering \
  && \
  $MAHOUT clusterdump \
    -i `$DFS -ls -d ${WORK_DIR}/reuters-kmeans/clusters-*-final | awk '{print $8}'` \
    -o ${WORK_DIR}/reuters-kmeans/clusterdump \
    -d ${WORK_DIR}/reuters-out-seqdir-sparse-kmeans/dictionary.file-0 \
    -dt sequencefile -b 100 -n 20 --evaluate -dm org.apache.mahout.common.distance.EuclideanDistanc
    --pointsDir ${WORK_DIR}/reuters-kmeans/clusteredPoints \
    && \
  cat ${WORK_DIR}/reuters-kmeans/clusterdump

elif [ "x$clustertype" == "xfuzzykmeans" ]; then
  $MAHOUT seq2sparse \
    -i ${WORK_DIR}/wikipediainput/ \
    -o ${WORK_DIR}/reuters-out-seqdir-sparse-fkmeans --maxDFPercent 85 --namedVector \
  && \
  $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 \
  && \
  $MAHOUT clusterdump \
    -i `$DFS -ls -d ${WORK_DIR}/reuters-fkmeans/clusters-*-final | awk '{print $8}'` \
    -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
fi
```

K-means clustering

```
[wangxucan (master *) mahout-trunk $ ./examples/bin/cluster-wikipedia.sh
Discovered Hadoop v2.
Setting dfs command to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs, dfs rm to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs
-rm -r -skipTrash.
Please select a number to choose the corresponding task to run
1. kmeans (may require increased heap space on yarn)
2. fuzzykmeans
3. clean -- cleans up the work area in /tmp/mahout-work-wangxucan
Enter your choice : 1
```



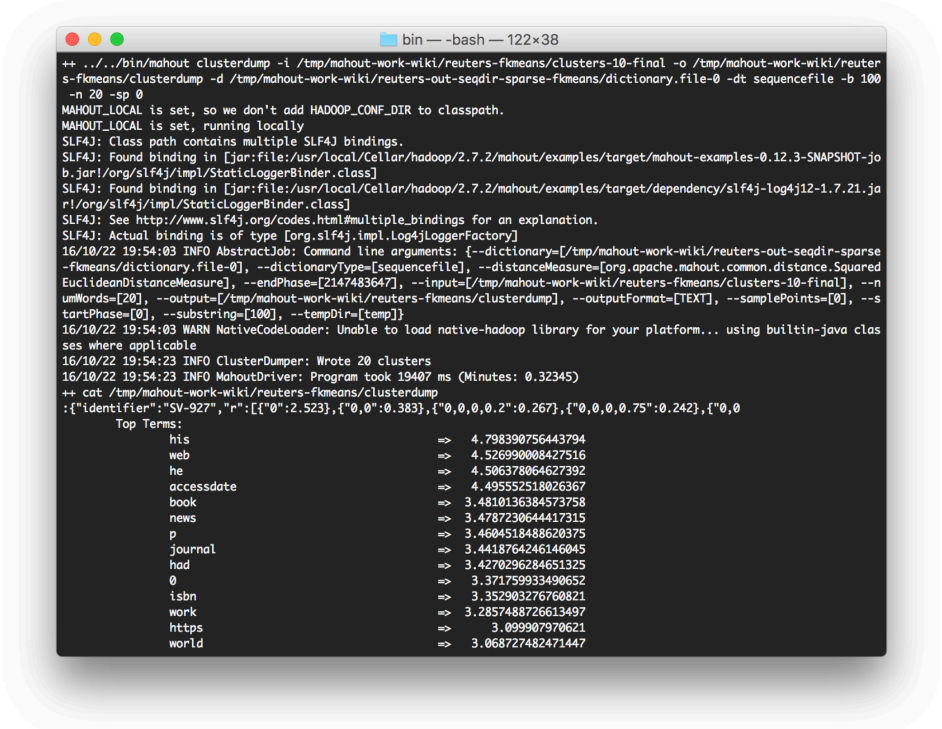
fuzzykmeans:

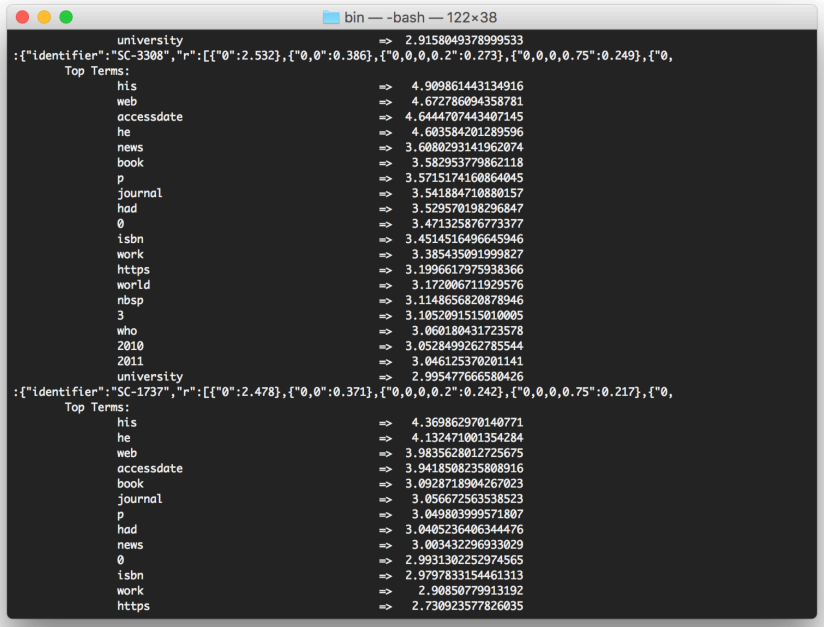
The fuzzykmeans took more time than the k-means and at last it caused a running out of memory error like the following. It took more heap space and resources. So in this part I only included the eroor it caused.

```
[wangxucan (master *) mahout-trunk $ ./examples/bin/cluster-wikipedia.sh
Discovered Hadoop v2.
Setting dfs command to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs, dfs rm to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs
-rm -r -skipTrash.
Please select a number to choose the corresponding task to run
1. kmeans (may require increased heap space on yarn)
2. fuzzykmeans
3. clean -- cleans up the work area in /tmp/mahout-work-wangxucan
Enter your choice : 2
```

```
java.lang.Exception: java.lang.OutOfMemoryError: Java heap space
        at org.apache.hadoop.mapred.LocalJobRunner$Job.runTasks(LocalJobRunner.java:462)
        at org.apache.hadoop.mapred.LocalJobRunner$Job.run(LocalJobRunner.java:522)
Caused by: java.lang.OutOfMemoryError: Java heap space
        at it.unimi.dsi.fastutil.ints.Int2DoubleOpenHashMap.rehash(Int2DoubleOpenHashMap.java:1059)
        at it.unimi.dsi.fastutil.ints.Int2DoubleOpenHashMap.insert(Int2DoubleOpenHashMap.java:295)
        at it.unimi.dsi.fastutil.ints.Int2DoubleOpenHashMap.put(Int2DoubleOpenHashMap.java:301)
        at org.apache.mahout.math.RandomAccessSparseVector.setQuick(RandomAccessSparseVector.java:130)
        at org.apache.mahout.math.VectorBinaryAssign$AssignNonzerosIterateThatLookupThisInplaceUpdates.assign(VectorBinary
Assign.java:174)
        at org.apache.mahout.math.VectorBinaryAssign.assignBest(VectorBinaryAssign.java:125)
        at org.apache.mahout.math.AbstractVector.assign(AbstractVector.java:558)
        at org.apache.mahout.clustering.AbstractCluster.observe(AbstractCluster.java:247)
        at org.apache.mahout.clustering.AbstractCluster.observe(AbstractCluster.java:235)
        at org.apache.mahout.clustering.AbstractCluster.observe(AbstractCluster.java:41)
        at org.apache.mahout.clustering.classify.ClusterClassifier.train(ClusterClassifier.java:153)
        at org.apache.mahout.clustering.iterator.CIMapper.map(CIMapper.java:56)
        at org.apache.mahout.clustering.iterator.CIMapper.map(CIMapper.java:34)
        at org.apache.hadoop.mapreduce.Mapper.run(Mapper.java:146)
        at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:787)
        at org.apache.hadoop.mapred.MapTask.run(MapTask.java:341)
        at org.apache.hadoop.mapred.LocalJobRunner$Job$MapTaskRunnable.run(LocalJobRunner.java:243)
        at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
        at java.lang.Thread.run(Thread.java:745)
16/10/22 19:16:10 INFO Job: Job job_local549920152_0002 failed with state FAILED due to: NA
16/10/22 19:16:10 INFO Job: Counters: 0
Exception in thread "main" java.lang.InterruptedException: Cluster Iteration 2 failed processing /tmp/mahout-work-wangxuca
n/reuters-fkmeans/clusters-2
        at org.apache.mahout.clustering.iterator.ClusterIterator.iterateMR(ClusterIterator.java:183)
        at org.apache.mahout.clustering.fuzzykmeans.FuzzyKMeansDriver.buildClusters(FuzzyKMeansDriver.java:287)
        at org.apache.mahout.clustering.fuzzykmeans.FuzzyKMeansDriver.run(FuzzyKMeansDriver.java:223)
        at org.apache.mahout.clustering.fuzzykmeans.FuzzyKMeansDriver.run(FuzzyKMeansDriver.java:115)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
        at org.apache.mahout.clustering.fuzzykmeans.FuzzyKMeansDriver.main(FuzzyKMeansDriver.java:51)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:497)
        at org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(ProgramDriver.java:71)
        at org.apache.hadoop.util.ProgramDriver.run(ProgramDriver.java:144)
        at org.apache.hadoop.util.ProgramDriver.driver(ProgramDriver.java:152)
        at org.apache.mahout.driver.MahoutDriver.main(MahoutDriver.java:195)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:497)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
```

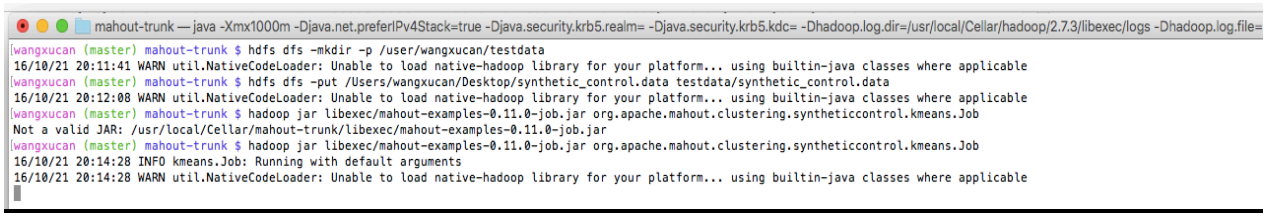
Since I didn't get my result, I referenced my teammates' result. It looks like this



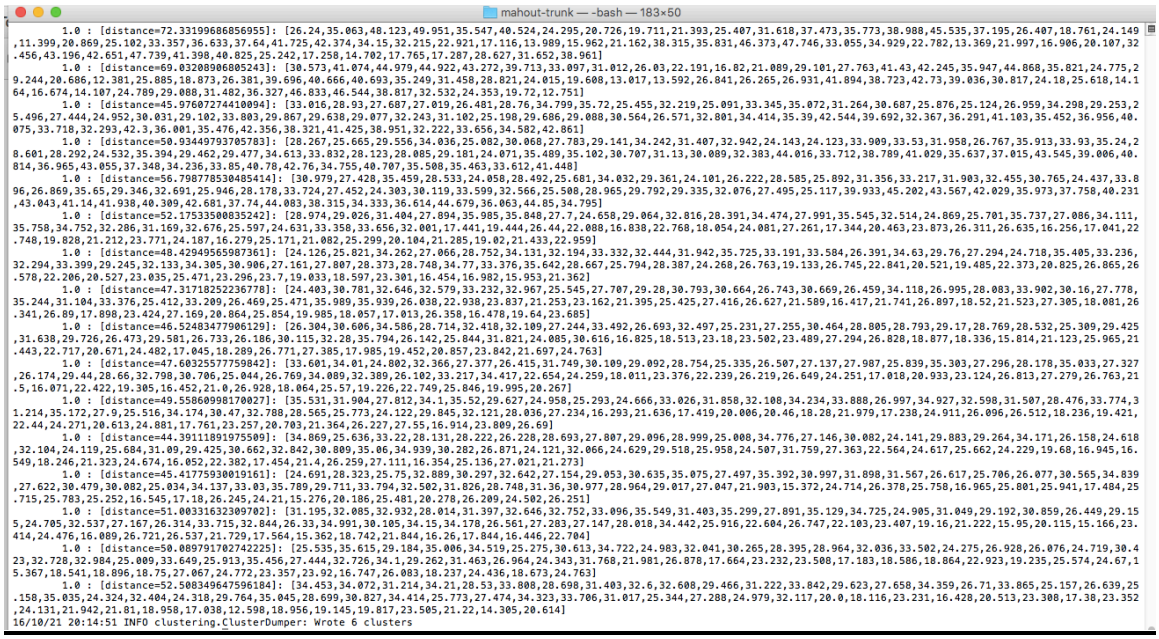


(3)synthetic_control.data

K-means clustering

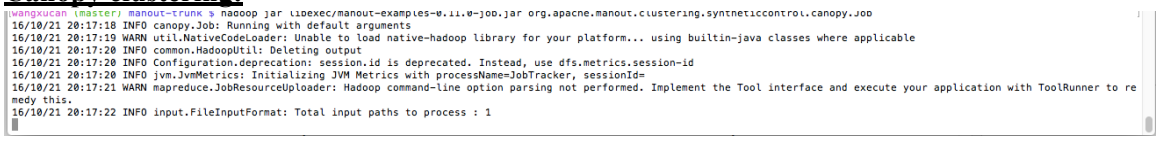


result:

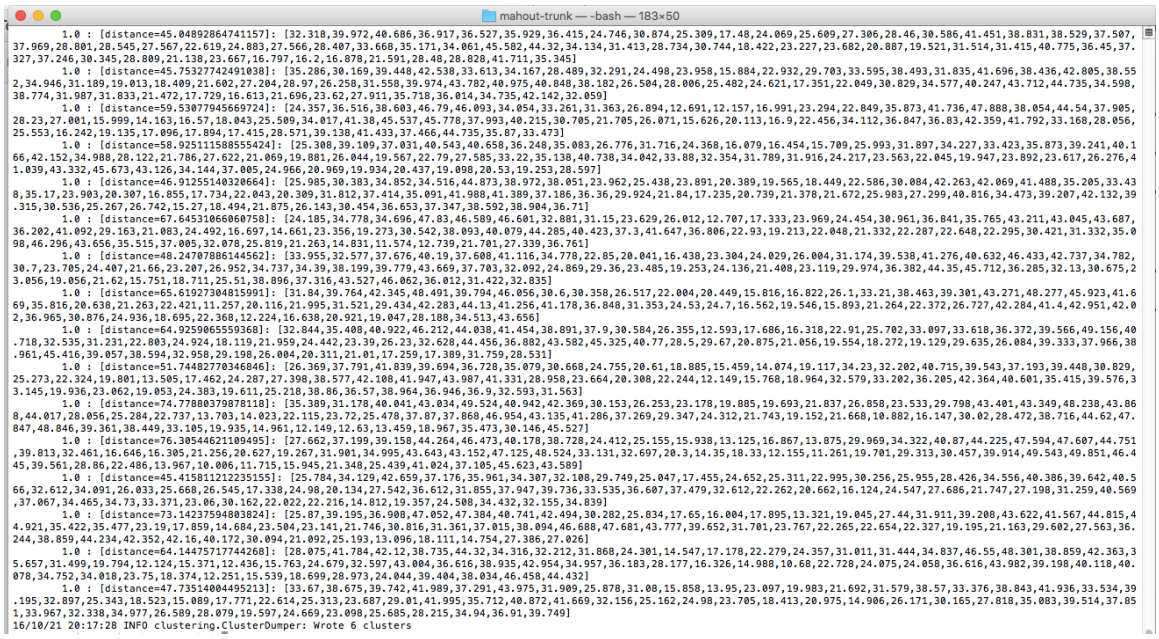


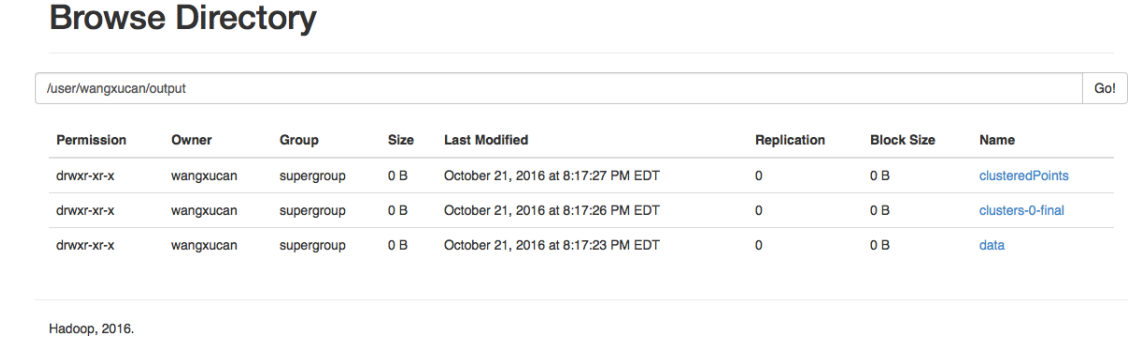
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	wangxucan	supergroup	194 B	October 21, 2016 at 8:14:49 PM EDT	1	128 MB	_policy
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:50 PM EDT	0	0 B	clusteredPoints
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:35 PM EDT	0	0 B	clusters-0
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:37 PM EDT	0	0 B	clusters-1
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:38 PM EDT	0	0 B	clusters-2
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:40 PM EDT	0	0 B	clusters-3
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:41 PM EDT	0	0 B	clusters-4
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:43 PM EDT	0	0 B	clusters-5
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:44 PM EDT	0	0 B	clusters-6
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:46 PM EDT	0	0 B	clusters-7
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:48 PM EDT	0	0 B	clusters-8
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:49 PM EDT	0	0 B	clusters-9-final
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:33 PM EDT	0	0 B	data
drwxr-xr-x	wangxucan	supergroup	0 B	October 21, 2016 at 8:14:34 PM EDT	0	0 B	random-seeds

Canopy clustering:



result:





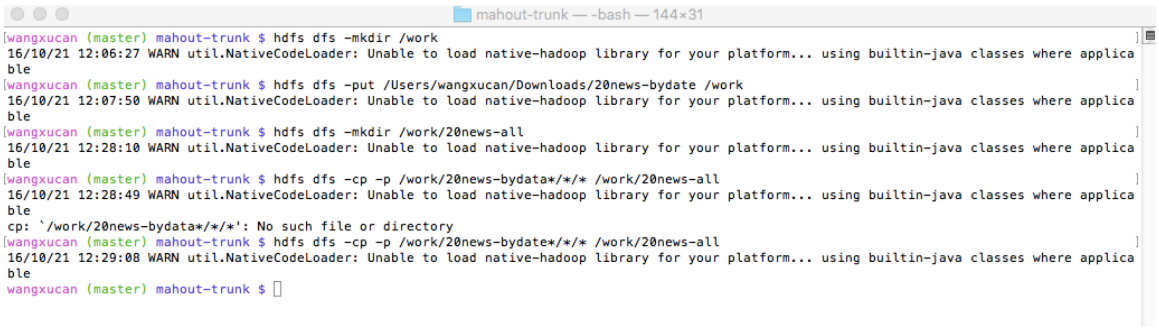
Part 3: Classification

Use datasets from online news or Wikipedia articles and do clustering to find related documents.

1) 20 news by date

In this part, I use the news dataset to train a Cnaive Bayesian model to do classification

a. Upload file and make directory in HDFS



b. Convert the dataset into <Text,Text> sequence file

```
[wangxucan (master) mahout-trunk $ ./bin/mahout seqdirectory -i /work/20news-all -o /work/20news-seq -ow
Running on hadoop, using /usr/local/Cellar/hadoop/2.7.3/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/Cellar/mahout-trunk/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar
16/10/21 12:46:37 INFO AbstractJob: Command line arguments: {--charset=[UTF-8], --chunkSize=[64], --endPhase=[2147483647], --fileFilterClass=[or
g.apache.mahout.text.PrefixAdditionFilter], --input=[/work/20news-all], --keyPrefix=[], --method=[mapreduce], --output=[/work/20news-seq], --ove
rwrite=null, --startPhase=[0], --tempDir=[temp]}
16/10/21 12:46:38 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/21 12:46:38 INFO deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
16/10/21 12:46:38 INFO deprecation: mapred.compress.map.output is deprecated. Instead, use mapreduce.map.output.compress
16/10/21 12:46:38 INFO deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
16/10/21 12:46:39 INFO deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
```

c. Convert the dataset into a <Text, VectorWrtable> SequenceFile Containing term frequencies for each document

```
mahout-trunk — java -Xmx1000m -Djava.net.preferIPv4Stack=true -Djava.security.krb5.realm= -Djava.security.krb5.kdc= -Dhadoop.log.dir=/usr/local/Cellar/h...
[wangxucan (master) mahout-trunk $ ./bin/mahout seq2sparse -i /work/20news-seq -o /work/20news-vectors -lnorm -nv -wt tfidf
Running on hadoop, using /usr/local/Cellar/hadoop/2.7.3/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/Cellar/mahout-trunk/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar
```

d. Split the preprocessed data into training set and testing set

```
16/10/21 14:02:53 INFO MahoutDriver: Program took 52953 ms (Minutes: 0.88255)
[wangxucan (master) mahout-trunk $ ./bin/mahout split -i /work/20news-vectors/tfidf-vectors --trainingOutput /work/20news-train-vectors --testOut
put /work/20news-test-vectors --randomSelectionPct 40 --overwrite --sequenceFiles -xm sequential
Running on hadoop, using /usr/local/Cellar/hadoop/2.7.3/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/Cellar/mahout-trunk/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar
```

e. Train the classifier

```
[wangxucan (master) mahout-trunk $ ./bin/mahout trainnb -i /work/20news-train-vectors -o /work/model -li /work/labelindex -ow -o /work/20news-testing -c
Running on hadoop, using /usr/local/Cellar/hadoop/2.7.3/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/Cellar/mahout-trunk/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar
16/10/21 14:10:39 WARN MahoutDriver: No trainnb.props found on classpath, will use command-line arguments only
16/10/21 14:10:39 INFO AbstractJob: Command line arguments: {--alphaI=[1.0], --endPhase=[2147483647], --input=[/work/20news-train-vectors], --labelIndex=[/work/labelindex], --output=[
/work/20news-testing], --overwrite=null, --startPhase=[0], --tempDir=[temp], --trainComplementary=null}
16/10/21 14:10:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/21 14:10:41 INFO CodecPool: Got brand-new compressor [.deflate]
16/10/21 14:10:41 INFO CodecPool: Got brand-new decompressor [.deflate]
16/10/21 14:10:44 INFO deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
16/10/21 14:10:44 INFO deprecation: mapred.compress.map.output is deprecated. Instead, use mapreduce.map.output.compress
```

f. Test the classifier

```
16/10/21 14:10:57 INFO MahoutDriver: Program took 18088 ms (Minutes: 0.30146666666666666)
[wangxucan (master) mahout-trunk $ ./bin/mahout testnb -i /work/20news-test-vectors -m /work/model -l /work/labelindex -ow -o /work/20news-testing -c
Running on hadoop, using /usr/local/Cellar/hadoop/2.7.3/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/Cellar/mahout-trunk/examples/target/mahout-examples-0.13.0-SNAPSHOT-job.jar
```

result:

```
16/10/21 14:18:03 INFO TestNaiveBayesDriver: Complementary Results:
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :       6700       89.1194%
Incorrectly Classified Instances        :        818       10.8806%
Total Classified Instances              :       7518

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p       q       r       s       t       <--Classified as
295     0       0       1       0       0       0       0       0       0       0       0       3       7       2       1       0       9       | 318         a     =
 alt.atheism
1       296     3       8       4       14      3       1       1       3       4       12      3       0       2       1       2       1       1       1       | 361         b     =
 comp.graphics
4       8       239     63      4       6       8       3       5       3       5       6       6       5       4       0       0       1       7       1       | 378         c     =
 comp.os.ms-windows.misc
0       3       4       316     15      4       14      1       2       1       1       2       9       3       2       0       6       1       1       0       | 385         d     =
 comp.sys.ibm.pc.hardware
3       4       2       8       308     1       3       2       2       2       0       2       6       0       1       1       2       0       0       0       | 347         e     =
 comp.sys.mac.hardware
3       12      2       3       6       355     0       1       0       0       0       8       1       2       2       1       0       1       1       1       | 399         f     =
 comp.windows.x
2       10      1       19      16      2       300     16      11      3       8       1       13      0       3       3       2       0       1       3       | 414         g     =
 misc.forsale
1       0       0       2       4       0       3       376     5       0       2       1       4       1       0       0       1       3       1       1       | 405         h     =
 rec.autos
1       0       0       2       0       1       0       2       379     1       0       1       3       1       1       0       1       0       0       0       | 393         i     =
 rec.motorcycles
0       0       0       1       2       2       2       0       0       385     5       0       1       1       0       0       0       0       0       0       | 399         j     =
 rec.sport.baseball
0       0       1       1       3       0       0       0       0       1       396     0       0       0       0       1       3       0       1       0       | 407         k     =
 rec.sport.hockey
1       0       0       0       0       0       1       0       0       0       0       386     1       0       1       0       0       0       0       1       | 391         l     =
 sci.crypt
0       3       1       10      4       2       7       6       3       0       0       5       358     3       3       2       2       0       5       2       | 416         m     =
 sci.electronics
2       1       0       1       0       1       1       1       0       0       1       1       5       377     3       2       4       1       1       2       | 404         n     =
 sci.med
0       2       0       0       2       2       0       1       0       1       1       0       0       0       375     0       3       0       1       0       | 388         o     =
 sci.space
3       1       0       1       0       0       0       2       1       1       0       0       1       4       0       387     2       0       0       1       | 404         p     =
 soc.religion.christian
0       0       1       0       0       1       0       0       1       0       0       4       0       0       1       1       327     0       4       2       | 342         q     =
 talk.politics.guns
0       1       0       0       0       0       0       0       0       0       0       0       0       0       0       0       1       387     1       0       | 390         r     =
 talk.politics.mideast
0       0       0       0       0       0       1       0       0       3       0       0       0       4       2       2       17      8       281     4       | 322         s     =
```

```
=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.8568
Accuracy                                   89.1194%
Reliability                                84.5424%
Reliability (standard deviation)            0.2181
Weighted precision                          0.8928
Weighted recall                             0.8912
Weighted F1 score                           0.8891

16/10/21 14:18:03 INFO MahoutDriver: Program took 9805 ms (Minutes: 0.16341666666666665)
wangxucan (master) mahout-trunk $
```

Then I used the existed model in classify-20newsgroups.sh and try different algorithms

C Naivebayes

```
mahout-trunk — classify-20newsgroups.sh — 183×51
[wangxucan (master) mahout-trunk $ ./examples/bin/classify-20newsgroups.sh
Discovered Hadoop v2.
Setting dfs command to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs, dfs rm to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs -rm -r -skipTrash.
Please select a number to choose the corresponding task to run
1. cnaivebayes-MapReduce
2. naivebayes-MapReduce
3. cnaivebayes-Spark
4. naivebayes-Spark
5. sgd
6. clean-- cleans up the work area in /tmp/mahout-work-wangxucan
Enter your choice : 1
```

The result is:

```
Summary
-------------------------------------------------------
Correctly Classified Instances          :       6740       89.0592%
Incorrectly Classified Instances        :        828       10.9408%
Total Classified Instances              :       7568

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p       q       r       s       t       <--Classified as
288     0       0       0       0       0       0       0       0       0       0       2       0       1       1       4       0       0       0       11      | 307       a
 alt.atheism
3       279     3       11      3       16      9       2       1       2       1       8       4       2       1       3       3       3       2       3       | 359       b
 comp.graphics
3       18      278     58      12      9       11      6       3       2       3       7       8       3       7       2       1       0       3       1       | 435       c
 comp.os.ms-windows.misc
1       9       4       341     12      5       9       3       2       0       2       5       7       4       3       2       2       0       3       2       | 416       d
 comp.sys.ibm.pc.hardware
0       2       7       8       349     4       3       3       0       2       3       0       0       1       1       2       0       1       1       0       | 387       e
 comp.sys.mac.hardware
0       12      0       1       1       334     2       1       3       1       0       4       0       2       3       1       0       0       2       1       | 368       f
 comp.windows.x
1       10      1       27      17      4       280     12      8       3       5       3       11      3       2       3       2       2       3       1       | 398       g
 misc.forsale
1       0       0       1       2       0       3       366     4       1       2       1       2       2       1       0       4       0       2       0       | 392       h
 rec.autos
1       0       0       0       0       0       1       3       408     0       0       0       1       1       0       1       1       0       0       0       | 417       i
 rec.motorcycles
0       0       0       1       0       0       1       0       1       371     9       0       2       4       0       0       0       0       0       0       | 389       j
 rec.sport.baseball
0       0       1       0       1       0       2       0       1       3       412     0       0       0       0       1       1       0       3       0       | 425       k
 rec.sport.hockey
0       0       0       1       0       0       0       0       0       1       0       408     1       1       0       1       2       0       1       0       | 416       l
 sci.crypt
0       7       0       9       12      1       6       4       2       2       2       2       344     2       5       4       2       1       2       1       | 408       m
 sci.electronics
3       1       1       1       2       0       1       2       0       2       1       0       5       359     1       1       0       0       1       2       | 383       n
 sci.med
1       0       0       0       1       0       1       1       0       0       1       0       2       2       378     1       2       0       0       2       | 392       o
 sci.space
4       0       0       1       0       1       0       0       1       0       0       0       0       3       1       369     1       3       1       4       | 389       p
 soc.religion.christian
0       0       1       0       0       1       1       0       0       0       0       4       0       0       2       0       360     0       3       0       | 372       q
 talk.politics.guns
0       1       0       0       0       0       0       0       0       2       0       0       0       0       1       2       0       361     0       1       | 368       r
 talk.politics.mideast
2       0       0       0       0       0       0       2       0       0       2       2       0       0       0       0       17      3       276     0       | 304       s
 talk.politics.misc
21      0       0       0       0       0       1       0       0       1       0       2       0       1       1       18      8       3       8       179     | 243       t
 talk.religion.misc
```

```
=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.856
Accuracy                                  89.0592%
Reliability                               84.7123%
Reliability (standard deviation)           0.2181
Weighted precision                         0.8913
Weighted recall                            0.8906
Weighted F1 score                          0.8883

16/10/21 14:34:33 INFO MahoutDriver: Program took 10292 ms (Minutes: 0.17153333333333334)
```

Naivebayes

```
(wangxucan (master) mahout-trunk $ ./examples/bin/classify-20newsgroups.sh
Discovered Hadoop v2.
Setting dfs command to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs, dfs rm to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs -rm -r -skipTrash.
Please select a number to choose the corresponding task to run
1. cnaivebayes-MapReduce
2. naivebayes-MapReduce
3. cnaivebayes-Spark
4. naivebayes-Spark
5. sgd
6. clean-- cleans up the work area in /tmp/mahout-work-wangxucan
Enter your choice : 2
```

mahout-trunk — -bash — 183×50

```
16/10/21 14:50:23 INFO TestNaiveBayesDriver: Standard NB Results:
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :       6857       90.5334%
Incorrectly Classified Instances        :        717        9.4666%
Total Classified Instances              :       7574

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p       q       r       s       t       <--Classified as
297     0       0       0       0       0       0       0       0       0       0       0       0       0       2       3       0       0       1       16      | 319          a     =
 alt.atheism
2       331     3       19      3       12      6       0       0       2       0       2       3       1       4       2       1       0       1       0       | 392          b     =
 comp.graphics
0       35      247     70      21      14      6       0       0       0       0       0       2       0       0       0       0       0       2       0       | 397          c     =
 comp.os.ms-windows.misc
1       8       5       328     20      2       7       0       0       0       1       1       6       0       0       0       0       0       0       0       | 379          d     =
 comp.sys.ibm.pc.hardware
0       2       3       10      365     2       3       2       0       0       0       1       6       0       0       0       0       0       0       0       | 394          e     =
 comp.sys.mac.hardware
0       23      1       4       5       352     0       1       0       2       1       1       1       1       1       0       1       0       0       0       | 394          f     =
 comp.windows.x
0       1       0       18      13      0       336     10      0       3       2       2       9       1       1       0       1       0       2       0       | 399          g     =
 misc.forsale
0       0       0       1       1       1       4       384     6       0       0       0       5       0       2       0       3       0       2       0       | 409          h     =
 rec.autos
0       0       0       0       2       0       3       6       384     0       0       0       2       0       0       0       0       0       0       1       | 398          i     =
 rec.motorcycles
0       0       0       1       1       0       3       1       2       403     4       0       0       1       0       0       0       0       0       0       | 416          j     =
 rec.sport.baseball
1       0       0       2       1       0       1       1       0       3       408     0       0       0       0       1       0       0       3       0       | 421          k     =
 rec.sport.hockey
0       5       0       2       1       1       4       1       0       0       0       380     0       2       0       0       0       0       0       0       | 396          l     =
 sci.crypt
0       6       0       20      8       1       6       3       0       0       0       2       321     0       1       0       1       0       2       0       | 371          m     =
 sci.electronics
0       3       0       2       0       2       3       0       0       0       0       0       4       422     2       0       2       1       1       2       | 444          n     =
 sci.med
0       6       0       0       1       1       0       0       0       0       0       0       0       2       382     0       0       0       1       0       | 393          o     =
 sci.space
2       1       0       0       0       0       1       0       0       0       0       0       0       3       0       358     2       2       1       3       | 373          p     =
 soc.religion.christian
0       0       1       0       0       0       0       0       2       1       0       5       0       0       0       0       326     1       7       1       | 344          q     =
 talk.politics.guns
2       0       0       0       0       0       0       0       0       0       0       0       0       0       0       2       0       389     1       0       | 394          r     =
 talk.politics.mideast
0       1       0       2       1       0       0       1       0       2       0       1       0       0       2       0       13      1       271     2       | 297          s     =
 talk.politics.misc
```

```
=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.8775
Accuracy                                  90.5334%
Reliability                               85.8573%
Reliability (standard deviation)           0.2166
Weighted precision                           0.91
Weighted recall                            0.9053
Weighted F1 score                          0.9047

16/10/21 14:50:23 INFO MahoutDriver: Program took 10648 ms (Minutes: 0.17746666666666666)
wangxucan (master) mahout-trunk $
```

Comparison: The accuracy of naïvebayes is 90.5334%, reliability is 85.8573%.

The accuracy of Cnaivebayes is 89.0592%, the reliability is 84.7123%

And the accuracy of my own trained model is 89.1194% and the reliability is 84.5424% which is a little lower than the existed one.

The running time of are both 5 minutes. So the naivebayes is better.

2) Wikipedia

I use the existed classify-wikipedia.sh in the examples

CBayes:

```
(wangxucan (master) mahout-trunk $ ./examples/bin/classify-wikipedia.sh
Discovered Hadoop v2.
Setting dfs command to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs, dfs rm to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs -rm -r -skipTrash.
Please select a number to choose the corresponding task to run
1. CBayes (may require increased heap space on yarn)
2. BinaryCBayes
3. clean -- cleans up the work area in /tmp/mahout-work-wiki
Enter your choice : 1
```

result:

```
Testing on holdout set: CBayes
+ /usr/local/Cellar/mahout/bin/mahout testnb -i /tmp/mahout-work-wiki/testing -m /tmp/mahout-work-wiki/model -l /tmp/mahout-work-wiki/labelindex -ow -o /tmp/mahout-work-wiki/output -c
 -seq
Running on hadoop, using /usr/local/Cellar/hadoop/2.7.3/bin/hadoop and HADOOP_CONF_DIR=
MAHOUT-JOB: /usr/local/Cellar/mahout/examples/target/mahout-examples-0.11.0-job.jar
16/10/21 17:59:19 WARN MahoutDriver: No testnb.props found on classpath, will use command-line arguments only
16/10/21 17:59:19 INFO AbstractJob: Command line arguments: {--endPhase=[2147483647], --input=[/tmp/mahout-work-wiki/testing], --labelIndex=[/tmp/mahout-work-wiki/labelindex], --model
=[/tmp/mahout-work-wiki/model], --output=[/tmp/mahout-work-wiki/output], --overwrite=null, --runSequential=null, --startPhase=[0], --tempDir=[temp], --testComplementary=null}
16/10/21 17:59:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/10/21 17:59:20 INFO HadoopUtil: Deleting /tmp/mahout-work-wiki/output
16/10/21 17:59:25 INFO CodecPool: Got brand-new compressor [.deflate]
16/10/21 17:59:25 INFO CodecPool: Got brand-new decompressor [.deflate]
16/10/21 17:59:34 INFO TestNaiveBayesDriver: Complementary Results:
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :       2013       85.6231%
Incorrectly Classified Instances        :        338       14.3769%
Total Classified Instances              :       2351

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       <--Classified as
759     8       44      6       6       6       4       8       9       9       |  859         a     = australia
0       107     2       2       1       1       6       0       2       0       |  121         b     = austria
0       0       9       1       2       1       0       0       0       0       |  13          c     = bahamas
4       7       20      441     8       7       2       11      7       2       |  509         d     = canada
0       0       0       0       34      0       0       0       0       0       |  34          e     = colombia
0       0       1       2       1       44      0       0       0       2       |  50          f     = cuba
0       1       1       0       5       0       80      0       0       0       |  87          g     = pakistan
0       1       0       3       3       2       0       6       0       0       |  15          h     = panama
6       17      51      4       6       5       11      14      482     6       |  602         i     = united kingdom
2       2       2       0       2       1       0       1       0       51      |  61          j     = vietnam

=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.7892
Accuracy                                   85.6231%
Reliability                                74.2079%
Reliability (standard deviation)            0.2924
Weighted precision                          0.9232
Weighted recall                             0.8562
Weighted F1 score                           0.8823

16/10/21 17:59:34 INFO MahoutDriver: Program took 15185 ms (Minutes: 0.2530833333333333)
```

The accuracy is 85.6231%, the reliability is 74.2079%

Binary-CBayes:

```
(wangxucan (master) mahout-trunk $ ./examples/bin/classify-wikipedia.sh
Discovered Hadoop v2.
Setting dfs command to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs, dfs rm to /usr/local/Cellar/hadoop/2.7.3/bin/hdfs dfs -rm -r -skipTrash.
Please select a number to choose the corresponding task to run
1. CBayes (may require increased heap space on yarn)
2. BinaryCBayes
3. clean -- cleans up the work area in /tmp/mahout-work-wiki
Enter your choice : 2
```

Result:

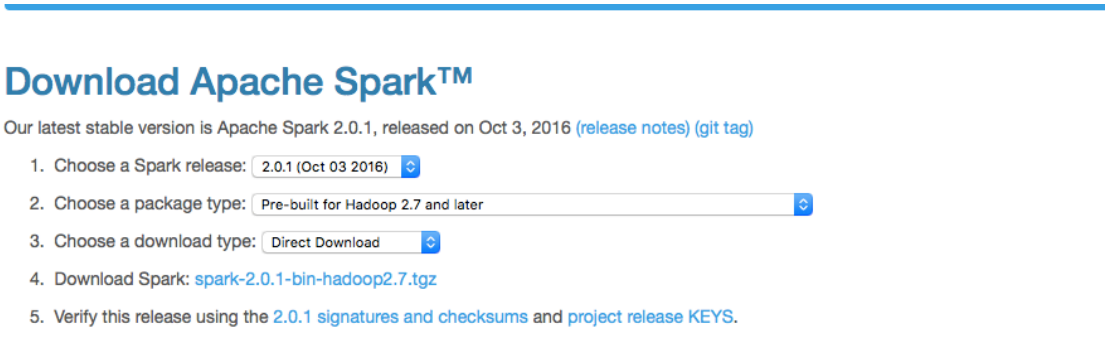
```
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :       2397       87.4498%
Incorrectly Classified Instances        :        344       12.5502%
Total Classified Instances              :       2741

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       <--Classified as
609     11       |  620         a     = united kingdom
333     1788     |  2121        b     = united states

=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.6953
Accuracy                                   87.4498%
Reliability                                60.8419%
Reliability (standard deviation)            0.5315
Weighted precision                          0.9153
Weighted recall                             0.8745
Weighted F1 score                           0.8823

16/10/21 18:13:59 INFO MahoutDriver: Program took 4846 ms (Minutes: 0.08076666666666667)
wangxucan (master) mahout-trunk $
```

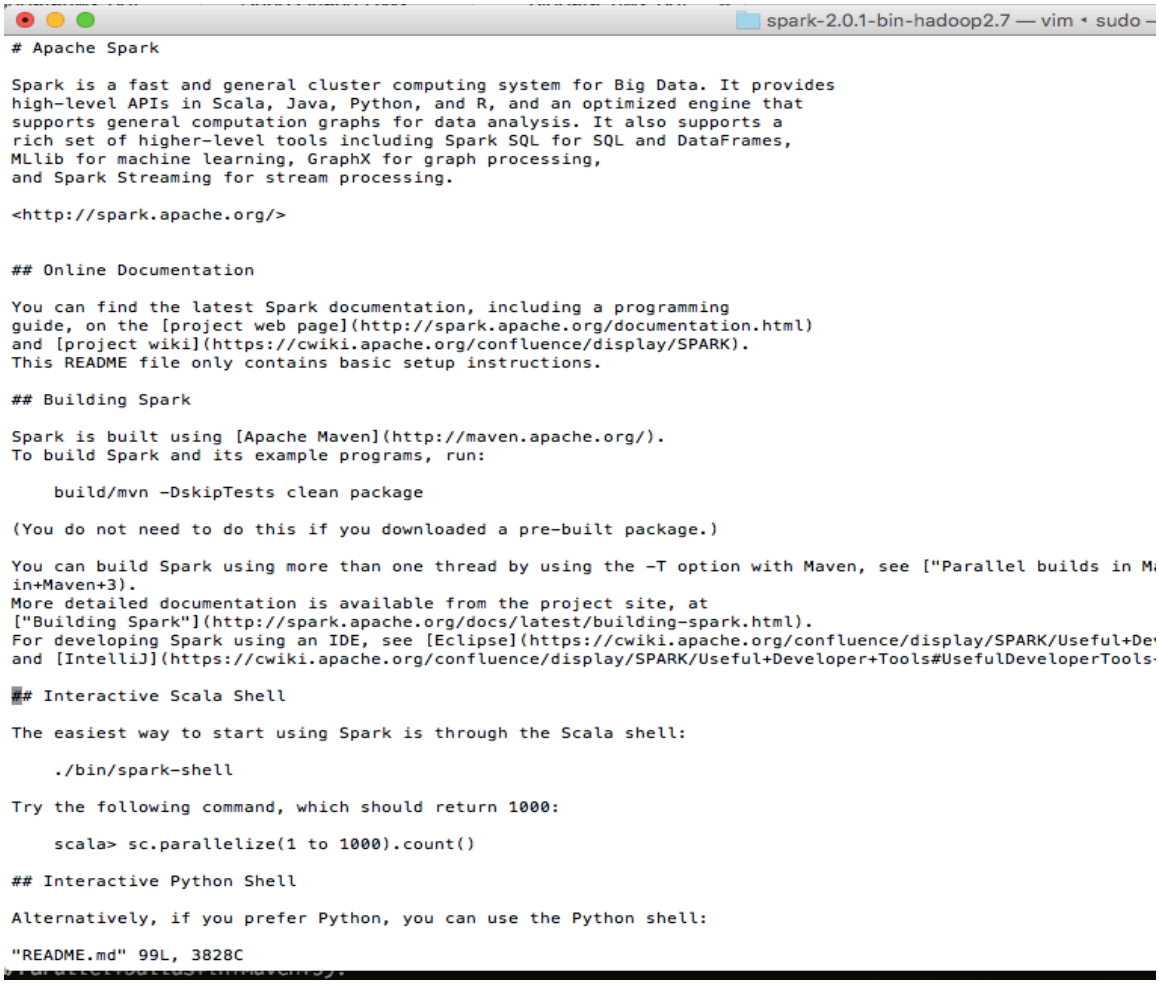
The accuracy is 87.4498%, reliability is 60.8419%

Compared to the cBayes model, the binaryCBayes model has higher accuracy but is less reliable.

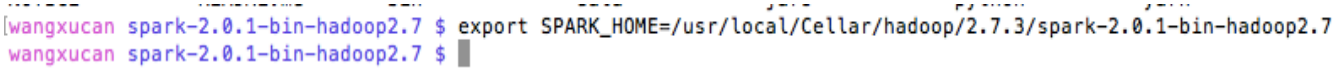
Part 4: Install Spark and run word count algorithm

a. download the pre-built spark

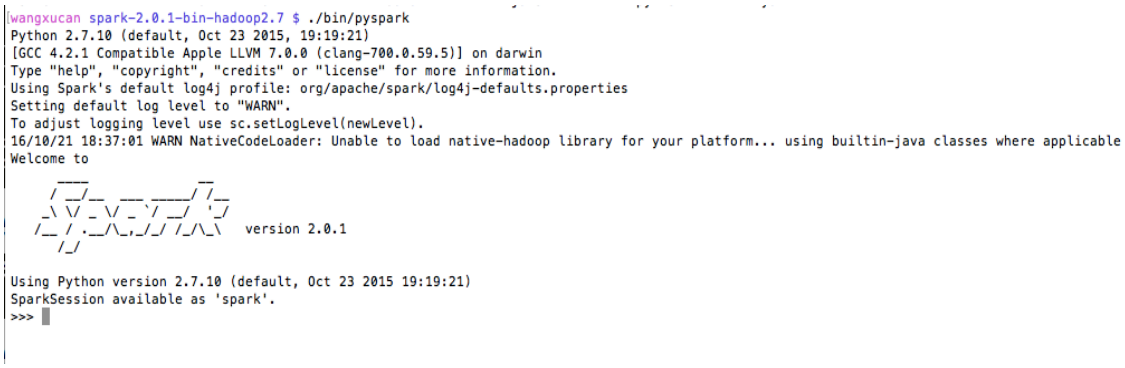




b. set the environment variable for spark



c. use the ./bin/pyspark to start it



d1: read the source file License.txt into a resilient distributed dataset

d2: count the number of items in the RDD, in my case, 1562 lines.
d3: First item in RDD which is empty in my case
d4: Count how many lines contains “License”
d5: Count how many lines contains “Apache”

```
16/10/21 18:37:01 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... u
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.0.1
      /_/

Using Python version 2.7.10 (default, Oct 23 2015 19:19:21)
SparkSession available as 'spark'.
>>> textFile = sc.textFile("LICENSE.txt")
>>> textFile.count()
1562
>>> textFile.first()
u''
>>> textFile.filter(lambda line:"License" in line).count()
112
>>> linesWithApache = textFile.filter(lambda line: "Apache" in line)
>>> linesWithApache.count()
5
>>>
```

The file looks like this

LICENSE.txt

```
                                 Apache License
                           Version 2.0, January 2004
                        http://www.apache.org/licenses/

   TERMS AND CONDITIONS FOR USE, REPRODUCTION, AND DISTRIBUTION

   1. Definitions.

      "License" shall mean the terms and conditions for use, reproduction,
      and distribution as defined by Sections 1 through 9 of this document.

      "Licensor" shall mean the copyright owner or entity authorized by
      the copyright owner that is granting the License.

      "Legal Entity" shall mean the union of the acting entity and all
      other entities that control, are controlled by, or are under common
      control with that entity. For the purposes of this definition,
      "control" means (i) the power, direct or indirect, to cause the
      direction or management of such entity, whether by contract or
      otherwise, or (ii) ownership of fifty percent (50%) or more of the
      outstanding shares, or (iii) beneficial ownership of such entity.

      "You" (or "Your") shall mean an individual or Legal Entity
      exercising permissions granted by this License.

      "Source" form shall mean the preferred form for making modifications,
      including but not limited to software source code, documentation
      source, and configuration files.
```