

Mahout - Kmeans

mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job --maxIter 15 --numClusters 10 --t1 425 --t2 200 --input random_number.txt --output kmeansRes

```
1.0 : [distance=143111.07389769657]: [83.0,254.0,193.0,154.0,203.0,489.0,298.0,401.0,68.0,377.0]
1.0 : [distance=151821.2395169402]: [92.0,268.0,147.0,467.0,350.0,449.0,383.0,311.0,106.0,130.0]
1.0 : [distance=77468.07427693577]: [214.0,84.0,337.0,366.0,262.0,369.0,150.0,103.0,34.0,229.0]
1.0 : [distance=213194.18691102415]: [464.0,450.0,453.0,256.0,91.0,482.0,371.0,229.0,231.0,349.0]
1.0 : [distance=137084.0307185871]: [247.0,402.0,340.0,395.0,88.0,255.0,287.0,410.0,85.0,51.0]
1.0 : [distance=102522.72044661827]: [54.0,366.0,267.0,459.0,259.0,493.0,42.0,178.0,199.0,261.0]
1.0 : [distance=168987.23886681534]: [8.0,428.0,471.0,342.0,40.0,463.0,344.0,396.0,302.0,240.0]
1.0 : [distance=151708.24590983242]: [368.0,89.0,460.0,255.0,228.0,337.0,272.0,324.0,40.0,455.0]
1.0 : [distance=194055.56647544052]: [244.0,469.0,243.0,481.0,23.0,426.0,392.0,441.0,104.0,215.0]
1.0 : [distance=243538.20283907675]: [389.0,49.0,285.0,493.0,85.0,331.0,413.0,34.0,26.0,376.0]
1.0 : [distance=157281.5328856695]: [26.0,221.0,364.0,346.0,2.0,354.0,148.0,447.0,247.0,468.0]
1.0 : [distance=208324.00574296666]: [78.0,188.0,499.0,282.0,3.0,172.0,481.0,163.0,56.0,415.0]
1.0 : [distance=190516.03023099387]: [49.0,13.0,280.0,315.0,311.0,486.0,170.0,344.0,292.0,81.0]
1.0 : [distance=114119.78979324317]: [94.0,241.0,475.0,442.0,273.0,333.0,41.0,399.0,103.0,123.0]
1.0 : [distance=110403.66263970523]: [192.0,262.0,411.0,369.0,68.0,194.0,173.0,214.0,15.0,54.0]
1.0 : [distance=172741.31401038473]: [32.0,340.0,221.0,407.0,362.0,499.0,126.0,41.0,212.0,102.0]
1.0 : [distance=124423.39121268131]: [235.0,347.0,477.0,235.0,84.0,315.0,412.0,58.0,129.0,306.0]
1.0 : [distance=163826.3696502156]: [29.0,303.0,374.0,155.0,40.0,395.0,67.0,6.0,81.0,323.0]
1.0 : [distance=232040.86553276004]: [12.0,79.0,334.0,491.0,19.0,135.0,53.0,402.0,186.0,420.0]
1.0 : [distance=162180.75663688802]: [11.0,202.0,321.0,406.0,443.0,477.0,310.0,364.0,258.0,149.0]
1.0 : [distance=110645.53461933462]: [129.0,180.0,119.0,429.0,221.0,289.0,189.0,168.0,225.0,122.0]
1.0 : [distance=121264.53510692855]: [58.0,218.0,345.0,490.0,359.0,277.0,370.0,367.0,37.0,193.0]
1.0 : [distance=168416.61263428768]: [232.0,296.0,413.0,96.0,364.0,346.0,199.0,315.0,15.0,483.0]
1.0 : [distance=124233.38541573752]: [223.0,247.0,211.0,407.0,60.0,302.0,241.0,471.0,34.0,345.0]
1.0 : [distance=144064.07633566344]: [280.0,473.0,260.0,239.0,80.0,314.0,294.0,65.0,46.0,194.0]
1.0 : [distance=134192.81059723068]: [59.0,258.0,89.0,420.0,120.0,318.0,308.0,391.0,111.0,351.0]
1.0 : [distance=131104.04545474472]: [159.0,455.0,380.0,320.0,308.0,427.0,290.0,448.0,28.0,163.0]
1.0 : [distance=121747.06268304703]: [158.0,391.0,473.0,246.0,314.0,369.0,34.0,385.0,39.0,223.0]
1.0 : [distance=154851.87604310783]: [266.0,207.0,464.0,342.0,354.0,208.0,60.0,419.0,74.0,421.0]
1.0 : [distance=83971.04475044343]: [345.0,333.0,232.0,425.0,125.0,291.0,256.0,301.0,49.0,275.0]
1.0 : [distance=154230.2289524153]: [90.0,284.0,480.0,212.0,289.0,499.0,268.0,342.0,289.0,55.0]
1.0 : [distance=188584.89570937702]: [21.0,449.0,493.0,396.0,99.0,289.0,425.0,171.0,349.0,388.0]
1.0 : [distance=65709.70717324084]: [204.0,380.0,241.0,357.0,171.0,470.0,225.0,308.0,173.0,393.0]
1.0 : [distance=133930.1394519275]: [132.0,133.0,377.0,274.0,115.0,277.0,131.0,125.0,316.0,491.0]
1.0 : [distance=148438.27289000317]: [69.0,375.0,459.0,234.0,135.0,285.0,87.0,184.0,350.0,85.0]
1.0 : [distance=152302.17520878115]: [224.0,301.0,472.0,184.0,453.0,405.0,184.0,258.0,120.0,462.0]
20/11/12 06:41:14 INFO ClusterDumper: Wrote 10 clusters
20/11/12 06:41:14 INFO MahoutDriver: Program took 426060 ms (Minutes: 7.101)
```

hadoop fs -ls /user/ec2-user/kmeansRes

```
[ec2-user@ip-172-31-77-124 bin]$ hadoop fs -ls /user/ec2-user/kmeansRes
Found 20 items
-rw-r--r-- 2 ec2-user supergroup 194 2020-11-12 06:40 /user/ec2-user/kmeansRes/_policy
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:40 /user/ec2-user/kmeansRes/clusteredPoints
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:34 /user/ec2-user/kmeansRes/clusters-0
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:34 /user/ec2-user/kmeansRes/clusters-1
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:38 /user/ec2-user/kmeansRes/clusters-10
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:38 /user/ec2-user/kmeansRes/clusters-11
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:38 /user/ec2-user/kmeansRes/clusters-12
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:39 /user/ec2-user/kmeansRes/clusters-13
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:39 /user/ec2-user/kmeansRes/clusters-14
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:40 /user/ec2-user/kmeansRes/clusters-15-final
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:35 /user/ec2-user/kmeansRes/clusters-2
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:35 /user/ec2-user/kmeansRes/clusters-3
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:35 /user/ec2-user/kmeansRes/clusters-4
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:36 /user/ec2-user/kmeansRes/clusters-5
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:36 /user/ec2-user/kmeansRes/clusters-6
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:37 /user/ec2-user/kmeansRes/clusters-7
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:37 /user/ec2-user/kmeansRes/clusters-8
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:37 /user/ec2-user/kmeansRes/clusters-9
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:34 /user/ec2-user/kmeansRes/data
drwxr-xr-x - ec2-user supergroup 0 2020-11-12 06:34 /user/ec2-user/kmeansRes/random-seeds
```

Mahout – Matrix Factorization – Movie Lens data

bin/mahout splitDataset --input movielens/ratings.csv --output ml_dataset --trainingPercentage 0.9 -
-probePercentage 0.1 --tempDir dataset/tmp

```
20/11/12 06:50:55 INFO Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=107072
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=21770226
    HDFS: Number of bytes written=1158034
    HDFS: Number of read operations=6
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=4042
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=4042
    Total vcore-milliseconds taken by all map tasks=4042
    Total megabyte-milliseconds taken by all map tasks=4139008
  Map-Reduce Framework
    Map input records=1000209
    Map output records=100278
    Input split bytes=142
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=69
    CPU time spent (ms)=2620
    Physical memory (bytes) snapshot=174080000
    Virtual memory (bytes) snapshot=2133798912
    Total committed heap usage (bytes)=96993280
  File Input Format Counters
    Bytes Read=21770084
  File Output Format Counters
    Bytes Written=1158034
20/11/12 06:50:55 INFO MahoutDriver: Program took 46970 ms (Minutes: 0.7828333333333334)
```

Size – ratings.csv

cd /home/ec2-user/MovieLens/ml-1m

ls -lrt

```
[ec2-user@ip-172-31-77-124 ml-1m]$ ls -lrt
total 35612
-rw-r----- 1 ec2-user ec2-user 134368 Feb 28 2003 users.dat
-rw-r----- 1 ec2-user ec2-user 24594131 Feb 28 2003 ratings.dat
-rw-r----- 1 ec2-user ec2-user 171308 Mar 26 2003 movies.dat
-rw-r----- 1 ec2-user ec2-user 5189 Aug 24 2011 README
-rw-rw-r-- 1 ec2-user ec2-user 11553456 Nov 12 06:48 ratings.csv
```

Two sampled files size (HDFS) –

/user/ec2-user/ml_dataset/trainingSet/

/user/ec2-user/ml_dataset/probeSet

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -ls /user/ec2-user/ml_dataset/trainingSet/
Found 2 items
-rw-r--r-- 2 ec2-user supergroup 0 2020-11-12 06:50 /user/ec2-user/ml_dataset/trainingSet/_SUCCESS
-rw-r--r-- 2 ec2-user supergroup 10395422 2020-11-12 06:50 /user/ec2-user/ml_dataset/trainingSet/part-m-00000
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -ls /user/ec2-user/ml_dataset/probeSet/
Found 2 items
-rw-r--r-- 2 ec2-user supergroup 0 2020-11-12 06:50 /user/ec2-user/ml_dataset/probeSet/_SUCCESS
-rw-r--r-- 2 ec2-user supergroup 1158034 2020-11-12 06:50 /user/ec2-user/ml_dataset/probeSet/part-m-00000
```

The two sampled files and ratings.csv files are of same size.

time bin/mahout parallelALS --input ml_dataset/trainingSet/ --output als/out --tempDir als/tmp --numFeatures 20 --numIterations 3 --lambda 0.065

```
20/11/12 07:08:25 INFO Job: map 100% reduce 0%
20/11/12 07:08:25 INFO Job: Job job_1605160849550_0064 completed successfully
20/11/12 07:08:25 INFO Job: Counters: 30
    File System Counters
        FILE: Number of bytes read=1061117
        FILE: Number of bytes written=108954
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=8230040
        HDFS: Number of bytes written=648993
        HDFS: Number of read operations=6
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=4288
        Total time spent by all reduces in occupied slots (ms)=0
        Total time spent by all map tasks (ms)=4288
        Total vcore-milliseconds taken by all map tasks=4288
        Total megabyte-milliseconds taken by all map tasks=4390912
    Map-Reduce Framework
        Map input records=3694
        Map output records=3694
        Input split bytes=132
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=77
        CPU time spent (ms)=3290
        Physical memory (bytes) snapshot=176410624
        Virtual memory (bytes) snapshot=2133405696
        Total committed heap usage (bytes)=98041856
    File Input Format Counters
        Bytes Read=8229908
    File Output Format Counters
        Bytes Written=648993
20/11/12 07:08:25 INFO MahoutDriver: Program took 162105 ms (Minutes: 2.70175)

real    2m48.388s
user    0m13.689s
sys     0m3.966s
```

bin/mahout evaluateFactorization --input ml_dataset/probeSet/ --output als/rmse/ --userFeatures als/out/U/ --itemFeatures als/out/M/ --tempDir als/tmp

```
20/11/12 07:15:11 INFO Job: Job Job_1000100010000_0000 Completed Successfully
20/11/12 07:15:11 INFO Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=107291
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=2868276
    HDFS: Number of bytes written=1620392
    HDFS: Number of read operations=11
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=3209
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=3209
    Total vcore-milliseconds taken by all map tasks=3209
    Total megabyte-milliseconds taken by all map tasks=3286016
  Map-Reduce Framework
    Map input records=100278
    Map output records=100266
    Input split bytes=132
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=65
    CPU time spent (ms)=1670
    Physical memory (bytes) snapshot=183230464
    Virtual memory (bytes) snapshot=2136690688
    Total committed heap usage (bytes)=97517568
  File Input Format Counters
    Bytes Read=1158034
  File Output Format Counters
    Bytes Written=1620392
20/11/12 07:15:11 INFO MahoutDriver: Program took 13583 ms (Minutes: 0.22638333333333333)
```

```
bin/mahout recommendfactorized --input als/out/userRatings/ --output recommendations/ --
userFeatures als/out/U/ --itemFeatures als/out/M/ --numRecommendations 6 --maxRating 5
```

```
20/11/12 07:16:40 INFO Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=108028
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=10007054
    HDFS: Number of bytes written=522720
    HDFS: Number of read operations=12
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=1
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=4738
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=4738
    Total vcore-milliseconds taken by all map tasks=4738
    Total megabyte-milliseconds taken by all map tasks=4851712
  Map-Reduce Framework
    Map input records=6040
    Map output records=6040
    Input split bytes=132
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=55
    CPU time spent (ms)=3610
    Physical memory (bytes) snapshot=186871808
    Virtual memory (bytes) snapshot=2140422144
    Total committed heap usage (bytes)=99614720
  File Input Format Counters
    Bytes Read=8296812
  File Output Format Counters
    Bytes Written=522720
20/11/12 07:16:40 INFO MahoutDriver: Program took 14729 ms (Minutes: 0.24548333333333333)
```

```
$HADOOP_HOME/bin/hadoop fs -cat recommendations/part-m-00000 | head
```

```
[ec2-user@ip-172-31-77-124 apache-mahout-distribution-0.11.2]$ $HADOOP_HOME/bin/hadoop fs -cat re
1      [572:4.68172,1198:4.447456,3147:4.364343,318:4.3070664,1226:4.231943,1250:4.221608]
2      [527:4.5096736,953:4.3980846,2762:4.294965,572:4.2847795,260:4.258563,914:4.21368]
3      [932:4.7042904,110:4.627271,3233:4.625551,2571:4.57553,527:4.535577,1246:4.5260067]
4      [128:5.0,3849:5.0,2931:5.0,602:5.0,2673:5.0,2203:5.0]
5      [3569:4.928562,1002:4.6734457,1423:4.393467,3171:4.3929787,503:4.3558617,2503:4.3341203]
6      [572:5.0,3233:4.612861,2197:4.604468,2084:4.3968053,2501:4.3958707,73:4.3644757]
7      [572:4.982803,260:4.760239,1198:4.7268867,858:4.6119757,318:4.5756745,1287:4.5683265]
8      [858:4.726266,260:4.7034802,1198:4.683688,318:4.6793427,50:4.614435,1221:4.6123137]
9      [858:4.379681,1198:4.359496,260:4.343517,110:4.325867,2905:4.2913766,1196:4.2215014]
10     [2197:5.0,572:4.9503646,2156:4.883605,831:4.703894,2562:4.6821556,497:4.5904346]
```

What is the top movie recommendation (movie ID) for users 4 and 5?

The top movie recommendation for user 4 is 128 and for user 5 is 3569

Problem 4

wget <http://dbgroup.cdm.depaul.edu/Courses/CSC555/spark-2.1.0-bin-hadoop2.6.tar>

tar xvf spark-2.1.0-bin-hadoop2.6.tar

cp slaves.template slaves

cat slaves

```
[ec2-user@ip-172-31-77-124 conf]$ cat slaves
#
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements. See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License. You may obtain a copy of the License at
#
# http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#
# A Spark Worker will be started on each of the machines listed below.
172.31.77.124
172.31.67.199
172.31.67.145
```

scp -r spark-2.1.0-bin-hadoop2.6 172.31.67.199:/home/ec2-user/

scp -r spark-2.1.0-bin-hadoop2.6 172.31.67.145:/home/ec2-user/

```
[ec2-user@ip-172-31-77-124 spark-2.1.0-bin-hadoop2.6]$ pwd
/home/ec2-user/spark-2.1.0-bin-hadoop2.6
sbin/start-all.sh
```

[ec2-user@ip-172-31-77-124 spark-2.1.0-bin-hadoop2.6]\$ jps

```
[ec2-user@ip-172-31-77-124 spark-2.1.0-bin-hadoop2.6]$ jps
5074 Worker
4691 JobHistoryServer
4213 ResourceManager
4357 NodeManager
4951 Master
5127 Jps
3706 NameNode
3868 DataNode
4063 SecondaryNameNode
```


← → ↻ 🔍 Not secure | 3.235.226.57:8080 ☆ ⚙️ 👤 ...

SPARK 2.1.0 **Spark Master at spark://ip-172-31-77-124.ec2.internal:7077**

URL: spark://ip-172-31-77-124.ec2.internal:7077
 REST URL: spark://ip-172-31-77-124.ec2.internal:6066 (cluster mode)

Alive Workers: 3
 Cores in use: 6 Total, 0 Used
 Memory in use: 8.5 GB Total, 0.0 B Used
 Applications: 0 Running, 0 Completed
 Drivers: 0 Running, 0 Completed
 Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-20201114030335-172.31.67.145-42329	172.31.67.145:42329	ALIVE	2 (0 Used)	2.8 GB (0.0 B Used)
worker-20201114030335-172.31.67.199-41335	172.31.67.199:41335	ALIVE	2 (0 Used)	2.8 GB (0.0 B Used)
worker-20201114030335-172.31.77.124-36677	172.31.77.124:36677	ALIVE	2 (0 Used)	2.8 GB (0.0 B Used)

Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
----------------	------	-------	-----------------	----------------	------	-------	----------

Completed Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
----------------	------	-------	-----------------	----------------	------	-------	----------

wget <http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/bioproject.xml>

hadoop fs -put bioproject.xml /data/

```
text_file = sc.textFile("hdfs://ip-172-31-77-124.ec2.internal/data/bioproject.xml")
counts = text_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word,
1)).reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://ip-172-31-77-124.ec2.internal/data/output")
```

hadoop fs -ls /data/output/

```
[ec2-user@ip-172-31-77-124 ~]$ hadoop fs -ls /data/output/
Found 3 items
-rw-r--r--  3 ec2-user supergroup          0 2020-11-14 03:21 /data/output/_SUCCESS
-rw-r--r--  3 ec2-user supergroup 14542025 2020-11-14 03:21 /data/output/part-00000
-rw-r--r--  3 ec2-user supergroup 14376208 2020-11-14 03:21 /data/output/part-00001
```

hadoop fs -cat /data/output/part-00000


```
(u'Ambion\xae,', 1)
(u'accession="PRJNA216174"', 1)
(u'units="Kb">5690</GenomeSize>', 2)
(u'species="385492"', 2)
(u'eucalypti</Name>', 1)
(u'accession="PRJNA47607"', 1)
(u'accession="PRJNA16845"', 1)
(u'rejection-associated', 2)
(u'\u2018ON\u2019;', 1)
(u'id="79913"/>', 1)
(u'id="155461"/>', 1)
(u'(WT|Nestin)', 1)
(u'Oooverexpression', 1)
(u'taxID="3681">', 1)
(u'accession="PRJNA112247"', 1)
(u'id="189918"/>', 1)
(u'\t\t\t\t\t<Title>Haloplanus', 2)
(u'(4-5851025)</Strain>', 1)
(u'between-species', 5)
(u'\t\t\t\t\t\t\t<ID>GSE34798</ID>', 1)
(u'MLL-AF9/Ras', 2)
(u'DFCI', 4)
(u'accession="PRJNA218446"', 1)
```