Azat Dovgeldiyev

CSC 555

Assignment 6

Problem 6.

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

Figure 9.8: A utility matrix for exercises

Exercise 9.3.1: Figure 9.8 is a utility matrix, representing the ratings, on a 1–5 star scale, of eight items, a through h, by three users A, B, and C. Compute the following from the data of this matrix.

(a) Treating the utility matrix as boolean, compute the Jaccard distance between each pair of users.

a)

No rating => 1,2

1 => 3,4,5

	a	b	c	d	e	f	g	h
A	**1**	**1**		**1**			**1**	
B		**1**	**1**	**1**				
C				**1**		**1**	**1**	**1**

$A \cap B = \{b, d\}$

$A \cup B = \{a, b, d, g, c\}$

Jaccard distance between A and B is :

$d(A,B) = 1 - SIM(A,B) => 1 - \frac{2}{5} = \frac{3}{5} = \mathbf{0.6}$

$B \cap C = \{d\}$
$B \cup C = \{b, c, d, f, g, h\}$

Jaccard distance between B and C is :

$$d(B,C) = 1 - SIM(B,C) => 1 - \frac{1}{6} = \frac{5}{6} = \mathbf{0.833}$$

$A \cap C = \{d, g\}$
$A \cup C = \{a, b, d, g, f, h\}$

Jaccard distance between A and C is :

$$d(A,C) = 1 - SIM(A,C) => 1 - \frac{2}{6} = \frac{2}{3} = \mathbf{0.666}$$

(e) Normalize the matrix by subtracting from each nonblank entry the average value for its user.

	a	b	c	d	e	f	g	h
A	**4-20/6**	**5-20/6**		**5-20/6**	**1-20/6**		**3-20/6**	**2-20/6**
B		**3 – 14/6**	**4-14/6**	**3-14/6**	**1-14/6**	**2-14/6**	**1-14/6**	
C	**2-3**		**1-3**	**3-3**		**4-3**	**5-3**	**3-3**

	a	b	c	d	e	f	g	h
A	**0.67**	**1.67**		**1.67**	**-2.33**		**-0.33**	**-0.67**
B		**0.67**	**1.67**	**-1.33**	**-0.33**	**-1.33**		**2.33**
C	**-1**		**-2**	**0**		**1**	**2**	**0**

b)

Clustering items based on a distance measure, such as items can be clustered into a smaller subset and assigned values would be the average ratings for each user for the columns in each cluster.

Problem 2.

a) How does Spark ensure that data is not lost when a failure occurs?
Input data comes form hdfs, and stored in hdfs, preventing data from loss during failure provided sufficient replication defined for HDFS.

b) From a resource managing perspective, which Hadoop nodes should be chosen to run Spark tasks?
Yarn configs are used to write to HDFS and connect to the YARN ResourceManager.

Problem 3.

a)

Randompy.py

```
 GNU nano 2.9.8                                          randompy.py

#!usr/bin/python
import sys
f = open("random_txt.txt",'w')
for i in range(1,3500001):
    if i%10==0:
        f.write(str(i) + '\n')
    else:
        f.write(str(i)+' ')
f.close()
```

Random_txt.txt file

```
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$ wc -l random_txt.
txt
350000 random_txt.txt
```

hadoop fs -cat testdata3/random_txt.txt | less

```
1 2 3 4 5 6 7 8 9 10
11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30
31 32 33 34 35 36 37 38 39 40
41 42 43 44 45 46 47 48 49 50
51 52 53 54 55 56 57 58 59 60
61 62 63 64 65 66 67 68 69 70
71 72 73 74 75 76 77 78 79 80
81 82 83 84 85 86 87 88 89 90
91 92 93 94 95 96 97 98 99 100
101 102 103 104 105 106 107 108 109 110
111 112 113 114 115 116 117 118 119 120
121 122 123 124 125 126 127 128 129 130
131 132 133 134 135 136 137 138 139 140
141 142 143 144 145 146 147 148 149 150
151 152 153 154 155 156 157 158 159 160
161 162 163 164 165 166 167 168 169 170
171 172 173 174 175 176 177 178 179 180
181 182 183 184 185 186 187 188 189 190
191 192 193 194 195 196 197 198 199 200
201 202 203 204 205 206 207 208 209 210
211 212 213 214 215 216 217 218 219 220
221 222 223 224 225 226 227 228 229 230
```

$MAHOUT_HOME/bin/mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job --maxIter 15 --numClusters 10 --t1 5 --t2 3 --input testdata3 --output kmeansRes1

```
0]
       1.0 : [distance=3.3489E11]: [749501.0,749502.0,749503.0,749504.0,749505.0,749506.0,749507.0,749508.0,749509.0,749510.0]
       1.0 : [distance=3.34926601E11]: [749511.0,749512.0,749513.0,749514.0,749515.0,749516.0,749517.0,749518.0,749519.0,749520.
0]
       1.0 : [distance=3.34963204E11]: [749521.0,749522.0,749523.0,749524.0,749525.0,749526.0,749527.0,749528.0,749529.0,749530.
0]
       1.0 : [distance=3.34999809E11]: [749531.0,749532.0,749533.0,749534.0,749535.0,749536.0,749537.0,749538.0,749539.0,749540.
0]
       1.0 : [distance=3.35036416E11]: [749541.0,749542.0,749543.0,749544.0,749545.0,749546.0,749547.0,749548.0,749549.0,749550.
0]
       1.0 : [distance=3.35073025E11]: [749551.0,749552.0,749553.0,749554.0,749555.0,749556.0,749557.0,749558.0,749559.0,749560.
0]
       1.0 : [distance=3.35109636E11]: [749561.0,749562.0,749563.0,749564.0,749565.0,749566.0,749567.0,749568.0,749569.0,749570.
0]
       1.0 : [distance=3.35146249E11]: [749571.0,749572.0,749573.0,749574.0,749575.0,749576.0,749577.0,749578.0,749579.0,749580.
0]
       1.0 : [distance=3.35182864E11]: [749581.0,749582.0,749583.0,749584.0,749585.0,749586.0,749587.0,749588.0,749589.0,749590.
0]
       1.0 : [distance=3.35219481E11]: [749591.0,749592.0,749593.0,749594.0,749595.0,749596.0,749597.0,749598.0,749599.0,749600.
0]
       1.0 : [distance=3.352561E11]: [749601.0,749602.0,749603.0,749604.0,749605.0,749606.0,749607.0,749608.0,749609.0,749610.0]
20/11/15 17:15:30 INFO ClusterDumper: Wrote 10 clusters
20/11/15 17:15:30 INFO MahoutDriver: Program took 1266917 ms (Minutes: 21.115283333333334)
```

hadoop fs -ls kmeansRes1

```
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$ hadoop fs -ls kmeansRes1
Found 20 items
-rw-r--r--   2 ec2-user supergroup        194 2020-11-15 16:59 kmeansRes1/_policy
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 17:00 kmeansRes1/clusteredPoints
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:54 kmeansRes1/clusters-0
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:55 kmeansRes1/clusters-1
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:58 kmeansRes1/clusters-10
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:58 kmeansRes1/clusters-11
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:58 kmeansRes1/clusters-12
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:59 kmeansRes1/clusters-13
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:59 kmeansRes1/clusters-14
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:59 kmeansRes1/clusters-15-final
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:55 kmeansRes1/clusters-2
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:55 kmeansRes1/clusters-3
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:56 kmeansRes1/clusters-4
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:56 kmeansRes1/clusters-5
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:56 kmeansRes1/clusters-6
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:57 kmeansRes1/clusters-7
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:57 kmeansRes1/clusters-8
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:57 kmeansRes1/clusters-9
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:54 kmeansRes1/data
drwxr-xr-x   - ec2-user supergroup          0 2020-11-15 16:54 kmeansRes1/random-seeds
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$
```

b)

more Movielens/ml-1m/ratings.dat

```
1::48::5::978824351
1::1097::4::978301953
1::1721::4::978300055
1::1545::4::978824139
1::745::3::978824268
1::2294::4::978824291
1::3186::4::978300019
1::1566::4::978824330
1::588::4::978824268
1::1907::4::978824330
1::783::4::978824291
1::1836::5::978300172
1::1022::5::978300055
1::2762::4::978302091
1::150::5::978301777
1::1::5::978824268
1::1961::5::978301590
1::1962::4::978301753
1::2692::4::978301570
1::260::4::978300760
1::1028::5::978301777
1::1029::5::978302205
1::1207::4::978300719
1::2028::5::978301619
1::531::4::978302149
1::3114::4::978302174
1::608::4::978301398
1::1246::4::978302091
2::1357::5::978298709
--More--(0%)
```

```
        Map-Reduce Framework
                Map input records=1000209
                Map output records=100212
                Input split bytes=142
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=70
                CPU time spent (ms)=3200
                Physical memory (bytes) snapshot=191127552
                Virtual memory (bytes) snapshot=2148175872
                Total committed heap usage (bytes)=97517568
        File Input Format Counters
                Bytes Read=21770084
        File Output Format Counters
                Bytes Written=1157666
20/11/15 17:28:47 INFO MahoutDriver: Program took 40326 ms (Minutes: 0.6721)
```

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

Sampled file sizes add up to the original input file size

```
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$ hadoop fs -ls /user/ec2-user/movielens
Found 1 items
-rw-r--r--   2 ec2-user supergroup   11553456 2020-11-15 17:27 /user/ec2-user/movielens/ratings.csv
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$ hadoop fs -ls /user/ec2-user/ml_dataset/probeSet
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2020-11-15 17:28 /user/ec2-user/ml_dataset/probeSet/_SUCCESS
-rw-r--r--   2 ec2-user supergroup    1157666 2020-11-15 17:28 /user/ec2-user/ml_dataset/probeSet/part-m-00000
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$ hadoop fs -ls /user/ec2-user/ml_dataset/trainingSet
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2020-11-15 17:28 /user/ec2-user/ml_dataset/trainingSet/_SUCCESS
-rw-r--r--   2 ec2-user supergroup   10395790 2020-11-15 17:28 /user/ec2-user/ml_dataset/trainingSet/part-m-00000
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$
```

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

```
        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)=75
                CPU time spent (ms)=3730
                Physical memory (bytes) snapshot=190914560
                Virtual memory (bytes) snapshot=2146394112
                Total committed heap usage (bytes)=100139008
        File Input Format Counters
                Bytes Read=8230728
        File Output Format Counters
                Bytes Written=648993
20/11/15 17:37:09 INFO MahoutDriver: Program took 147333 ms (Minutes: 2.45555)

real    2m33.693s
user    0m12.679s
sys     0m4.173s
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$
```

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

```
20/11/15 17:38:36 INFO MahoutDriver: Program took 13148 ms (Minutes: 0.21913333333333335)
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$ hadoop fs -ls /user/ec2-user/als/rmse/rmse.txt
-rw-r--r--   2 ec2-user supergroup         18 2020-11-15 17:38 /user/ec2-user/als/rmse/rmse.txt
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$ hadoop fs -cat /user/ec2-user/als/rmse/rmse.txt
0.8849254480772761[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$
```

RMSE = 0.8849254480772761

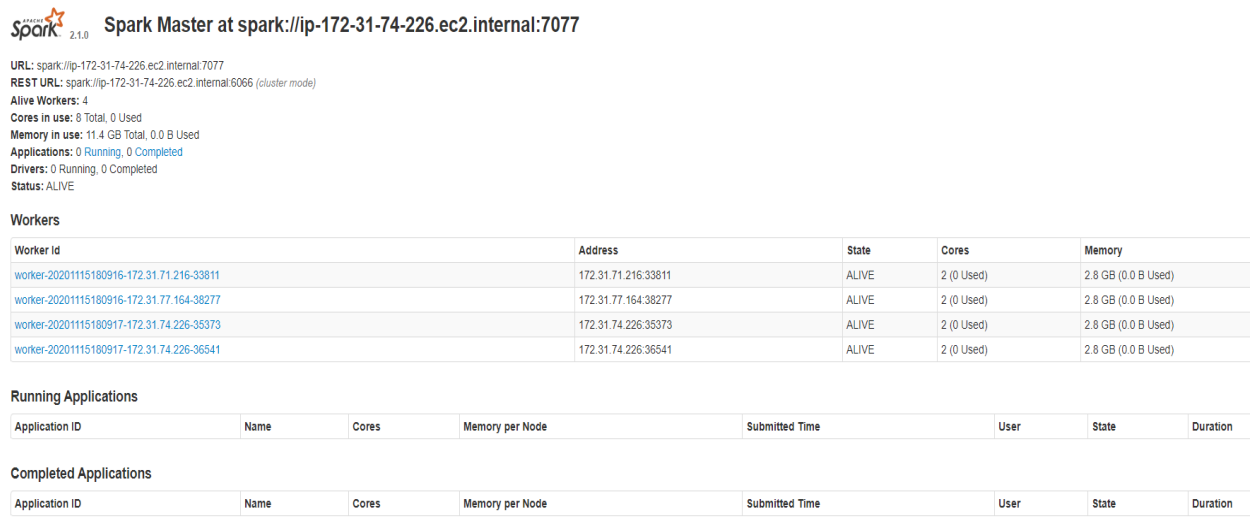
Finally, let's generate some predictions:

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

```
[ec2-user@ip-172-31-74-226 apache-mahout-distribution-0.11.2]$ $HADOOP_HOME/bin/hadoop fs -cat recommen
d
1       [572:5.0,1226:4.56718,1380:4.421373,593:4.4206724,932:4.380784,953:4.341562]
2       [572:4.7326922,527:4.4004936,3092:4.287736,318:4.237336,919:4.18811,953:4.186276]
3       [572:4.4184737,110:4.398523,2762:4.3529034,2571:4.340958,318:4.2635036,1036:4.219557]
4       [1221:5.0,1193:5.0,750:5.0,615:5.0,858:5.0,912:5.0]
5       [2503:4.5701146,2981:4.4554214,106:4.403509,1420:4.3336763,3134:4.3222466,668:4.249948]
6       [572:5.0,3314:4.681011,985:4.640288,2156:4.537192,3637:4.4986386,2332:4.4648986]
7       [1198:4.76922,1240:4.6984468,718:4.695146,1036:4.6298122,260:4.598336,2762:4.5182023]
8       [50:4.7097163,858:4.6860614,1218:4.666805,745:4.6521616,2762:4.6506844,1148:4.6440187]
9       [2905:4.3795056,110:4.283271,296:4.2771235,858:4.266666,260:4.216727,1198:4.2023897]
10      [572:5.0,2197:5.0,2156:4.7616115,2426:4.446826,598:4.441046,3092:4.4366703]
```

USER	MOVIE_ID
User 4	**1221** (all movies are rated 5)
User 5	**2503**

Problem 4.



Spark Master at spark://ip-172-31-74-226.ec2.internal:7077

URL: spark://ip-172-31-74-226.ec2.internal:7077
REST URL: spark://ip-172-31-74-226.ec2.internal:6066 *(cluster mode)*
Alive Workers: 4
Cores in use: 8 Total, 0 Used
Memory in use: 11.4 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-20201115180916-172.31.71.216-33811	172.31.71.216:33811	ALIVE	2 (0 Used)	2.8 GB (0.0 B Used)
worker-20201115180916-172.31.77.164-38277	172.31.77.164:38277	ALIVE	2 (0 Used)	2.8 GB (0.0 B Used)
worker-20201115180917-172.31.74.226-35373	172.31.74.226:35373	ALIVE	2 (0 Used)	2.8 GB (0.0 B Used)
worker-20201115180917-172.31.74.226-36541	172.31.74.226:36541	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

text_file = sc.textFile("hdfs://ec2-3-238-126-42.compute-1.amazonaws.com/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://ec2-3-238-126-42.compute-1.amazonaws.com/data/countOutput")

```
>>> counts.saveAsTextFile("hdfs://ec2-3-238-126-42.compute-1.amazonaws.com/data/countOutput")
>>> exit()
[ec2-user@ip-172-31-74-226 spark-2.1.0-bin-hadoop2.6]$ hadoop fs -ls /data/countOutput
Found 3 items
-rw-r--r--   3 ec2-user supergroup          0 2020-11-16 01:17 /data/countOutput/_SUCCESS
-rw-r--r--   3 ec2-user supergroup   14542025 2020-11-16 01:17 /data/countOutput/part-00000
-rw-r--r--   3 ec2-user supergroup   14376208 2020-11-16 01:17 /data/countOutput/part-00001
[ec2-user@ip-172-31-74-226 spark-2.1.0-bin-hadoop2.6]$ hadoop fs -cat /data/countOutput/part-00001 | less
(u'species="188144"', 7)
(u'species="40681"', 2)
(u'ndr1-1', 6)
(u'ER15_174_BHI7</OrganismName>', 1)
(u'Palb2+/+;p53-/-', 3)
(u'DP4', 1)
(u'ChIP-Chip</Description>', 1)
(u'accession="PRJNA177610"', 1)
(u'nanhaiticus</OrganismName>', 1)
(u'id="49937"/>', 1)
(u'spiders', 2)
(u'species="337382"', 2)
(u'woody', 57)
(u'\t\t\t\t\t\t<Reference>23951555</Reference>', 1)
(u'\t\t\t\t\t\t\t\t\t\t<Name>pK1S1</Name>', 1)
(u'canes', 1)
(u'(TMX)', 1)
(u'Fragility</Name>', 2)
```