**CSC 555 Assignment 5**

**Yiyang Yang**

**Part 1**

(a)

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

Jaccard distance:

A - B: 1 – 2/5 = 3/5

B – C: 1 – 1/6 = 5/6

A – C: 1 – 1/3 = 2/3

(e)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h | Avg |
| A | 4 | 5 |  | 5 | 1 |  | 3 | 2 | 3.33 |
| B |  | 3 | 4 | 3 | 1 | 2 | 1 |  | 2.33 |
| C | 2 |  | 1 | 3 |  | 4 | 5 | 3 | 3 |

Substract average

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h | Avg |
| A | 0.67 | 1.67 |  | 1.67 | -2.33 |  | -0.33 | -0.67 | 3.33 |
| B |  | 0.67 | 1.67 | 0.67 | -1.33 | -0.33 | -1.33 |  | 2.33 |
| C | -1 |  | -2 | 0 |  | 1 | 2 | 0 | 3 |

**Part 2**

a)

Spark read the data from HDFS by default.

b)

Content based:

System will analyze user’s history and then recommends items.

Collaborative filtering based:

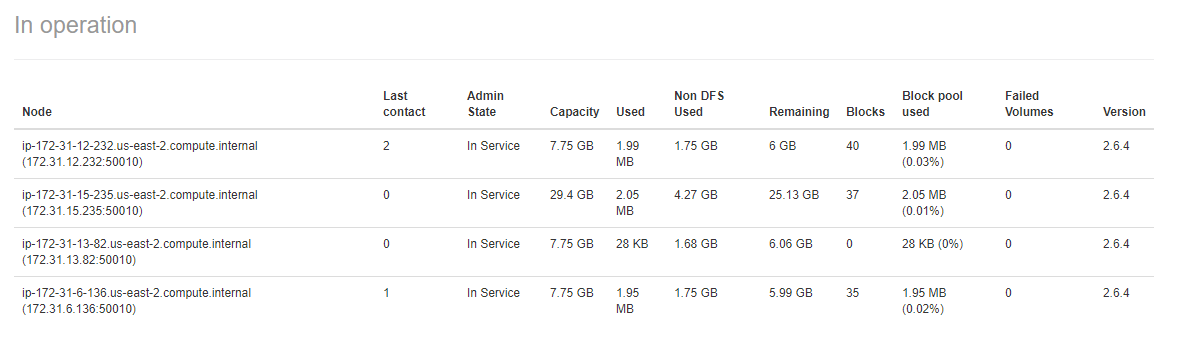
System will find similarity between users and then recommend items by users’ preferences.

c)

Cluster item and users. By selecting distance measures, and use it to perform a clustering.

**Part 3**

a)



b)



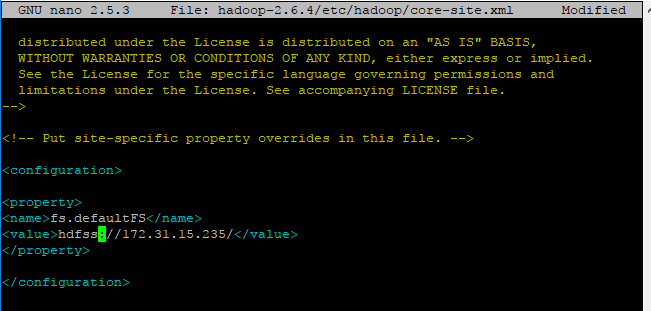
c)



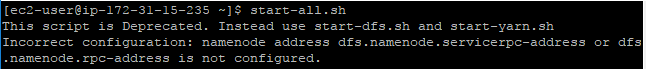
It looks like remove one node the run time will be a little bit slower.

d)

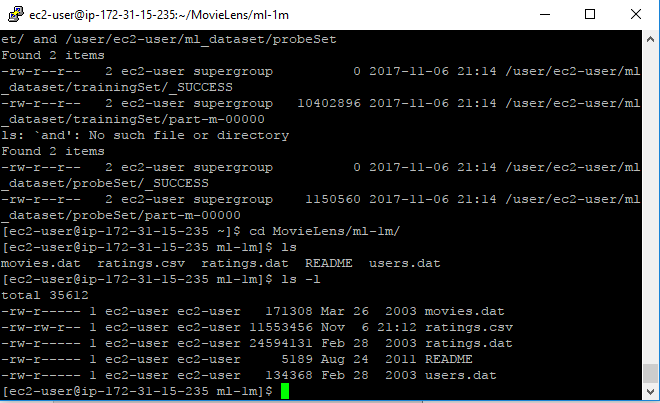
Typo



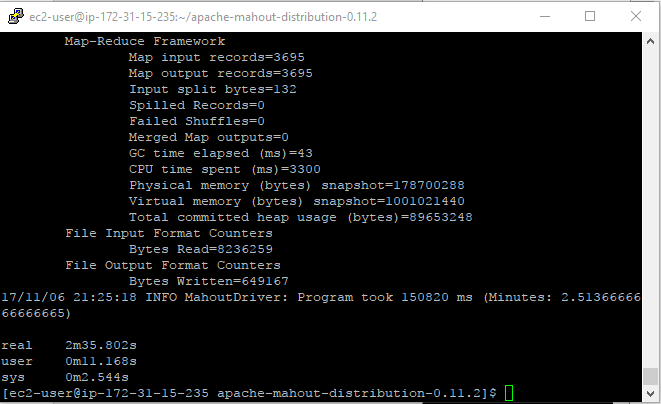
Error message

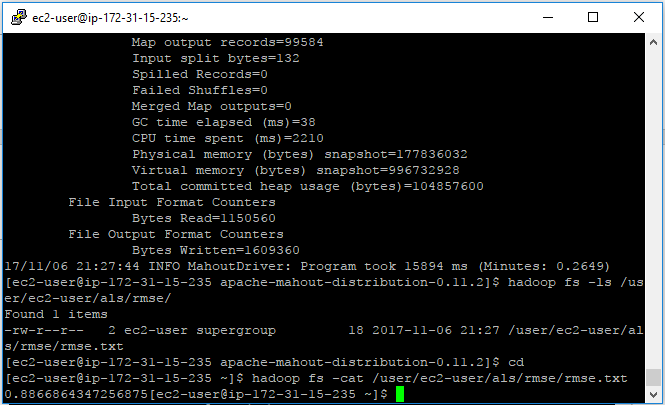


**Part 4**

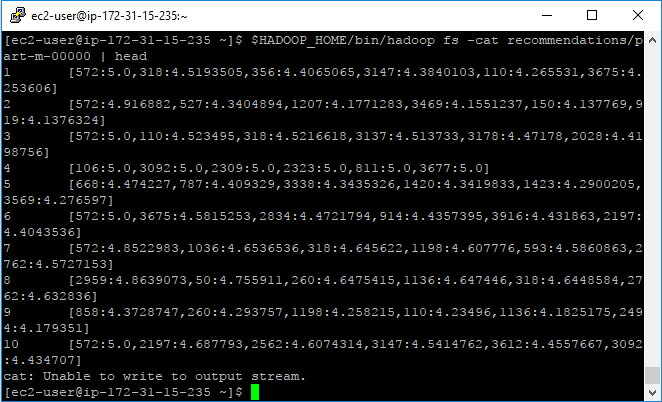


The size of rating.csv is 11553456 bytes, the two separate files size are 10402896 bytes and 1150560 bytes, 10402896 + 1150560 = 11553456. Sampled file sizes add up to the original input file size.





Result of RMSE value is 0.8866864347256875



2: 572:4.916882

3: 572:5.0

4: 106:5.0