Homework 3

Technology : Python: 2.7 Scala: 2.11 Spark: 2.2.1

Directory:

Surbhi_Batra_hw2

- Surbhi_Batra
  - Surbhi_Batra_Description
- OutputFiles
  - Surbhi_Batra_ModelBasedCF_Small.txt
  - Surbhi_Batra_ModelBasedCF_Big.txt
  - Surbhi_Batra_SimilarMovies_Jaccard.txt
  - Surbhi_Batra_UserBasedCF.txt
  - Surbhi_Batra_ItemBasedCF.txt
- Solution
  - Surbhi_Batra_task1_Jaccard.py
  - Surbhi_Batra_task2_ModelBasedCF.py
  - Surbhi_Batra_task2_UserBasedCF.py
  - Surbhi_Batra_task2_ItemBasedCF.py

Algorithm:

>>>Execute these Commands after : cd/to/Surbhi_Batra_hw3/Solution

Python:

TASK : 1

LSH

Command :
$SPARK_HOME/bin/spark-submit Surbhi_Batra_task1_Jaccard.py MovieLens.Small.csv

Precision: 1.000000
Recall: 0.909214
Time : 184.732897997

TASK : 2

	Task 2A		**Task2B**
	SMALL	**LARGE**	**SMALL**
>=0 and <1	13808	3223235	15052
>=1 and <2	4146	733558	4160
>=2 and <3	665	81967	890
>=3 and <4	105	7383	152
>=4	9	188	2
RMSE	0.949470593765	0.819786213789	0.962587911664
TIME (seconds)	18.9599509239	1797.52702594	81.8034739494

MODEL - BASED

rank = 10
numIterations = 10
lambdas = 0.1

MovieLens.Small.csv

Command : $SPARK_HOME/bin/spark-submit Surbhi_Batra_task2_ModelBasedCF.py MovieLens.Small.csv testing_small.cs

```
small

 No of rows in ground truth: 20256.000000
 No of rows in Predictions: 18733.000000
(10, 10, 0.1)

 >=0 and <1 : 13808more
 >=1 and <2 : 4146 less
 >=2 and <3: 665 less
 >=3 and <4 : 105 less
 >=4        : 9 less

 RMSE : 0.949470593765 less

Endtime:  18.9599509239
```

MovieLens.Big.csv
Command :

$SPARK_HOME/bin/spark-submit Surbhi_Batra_task2_ModelBasedCF.py MovieLens.Big.csv testing_20m.csv

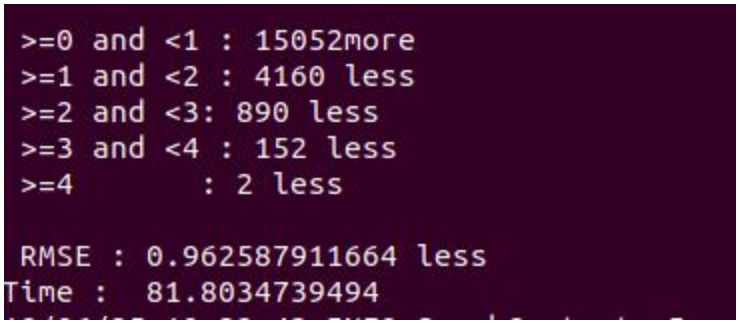
Output file not attached as of now.. Because too big

```
small

 No of rows in ground truth: 4054451.000000
 No of rows in Predictions: 4046331.000000
(10, 10, 0.1)

 >=0 and <1 : 3223235more
 >=1 and <2 : 733558more
 >=2 and <3: 81967more
 >=3 and <4 : 7383more
 >=4        : 188more

 RMSE : 0.819786213789 less

Endtime:  1797.52702594
```

USER - BASED

Used Pearson co-relation in the algorithm.

$SPARK_HOME/bin/spark-submit Surbhi_Batra_task2_UserBasedCF.py MovieLens.Small.csv testing_small.csv

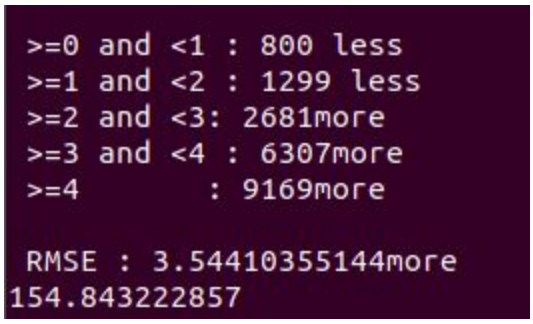


TASK : 3

Command : $SPARK_HOME/bin/spark-submit Surbhi_Batra_task2_ItemBasedCF.py MovieLens.Small.csv testing_small.csv

JACCARD SIMILARITY

With LSH



RMSE : 3.54410355144more

Pearson

RMSE : 0.96

The result is very bad because Jaccard similarity doesn't consider the actual ratings
But just 0 and 1s and hence we do not have an accurate similarity measure.
While Pearson co - relation calculates the actual similarity using corated items.