Spark Version: 2.2.1 Python Version: 2.7.15

Model-Based CF

1. Go to command prompt on Windows OS.
2. In command line change directory to where spark is installed and then run command-
"**bin\spark-submit Snehal_Shirgure_task2_ModelBasedCF.py <ratings file> <test file>**"
And hit enter
3. Output saved to –
Snehal_Shirgure_ModelBasedCF.txt
For small file -
>=0 and <1: 13928
>=1 and <2: 3993
>=2 and <3: 680
>=3 and <4: 125
>=4: 7
RMSE: 0.948179056742
Time: 213.08100009 sec
For big file -
>=0 and <1: 3191803
>=1 and <2: 765525
>=2 and <3: 82308
>=3 and <4: 6573
>=4: 129
RMSE: 0.8385254360124645
Time: 5940. 02302031 sec

User-Based CF

1. Go to command prompt on Windows OS.
2. In command line change directory to where spark is installed and then run command-

```
“bin\spark-submit Snehal_Shirgure_task2_UserBasedCF.py <ratings file> <test file>”
```

And hit enter

3. Output saved to –

Snehal_Shirgure_UserBasedCF.txt

For small file:

```
>=0 and <1: 15433
>=1 and <2: 3987
>=2 and <3: 707
>=3 and <4: 124
>=4: 5
```

RMSE: 0.924241691853

Time: 88.6360001564 sec

Item-Based CF

1. Go to command prompt on Windows OS.
2. In command line change directory to where spark is installed and then run command-

```
“bin\spark-submit Snehal_Shirgure_task2_ ItemBasedCF.py <ratings file> <test file>”
```

And hit enter

3. Output saved to –

Snehal_Shirgure_ItemBasedCF.txt

For small file **with LSH:**

```
>=0 and <1: 13640
>=1 and <2: 5106
>=2 and <3: 1247
>=3 and <4: 234
>=4: 29
```

RMSE: 1.05309732877

Time: 300.947999954 sec

without LSH:

```
>=0 and <1: 13824
>=1 and <2: 5167
>=2 and <3: 1046
>=3 and <4: 203
>=4: 16
```

RMSE: 1.02597760288

Time: 224.019000053 sec

As RMSE value for Pearson correlation(without LSH) is greater than using Jaccard based LSH similarity values, it is a better algorithm.