Homework - 3

Contents of the folder -
1) Input (ratings.csv, testing_small.csv)
2) Output (SaiSree_Kamineni_result_task1.txt, SaiSree_Kamineni_result_task2.txt)
3) Solution – Scala code (SaiSree_Kamineni _task1.scala, SaiSree_Kamineni _task2.scala), Jar files (SaiSree_Kamineni _task1.jar, SaiSree_Kamineni _task2.jar)
4) Description File (SaiSree_Kamineni_description.pdf)

Note - I haven't set up SPARK_HOME variable. So I run the scripts and commands from inside "*spark-1.6.3-bin-hadoop2.4*" directory.
Place this UnZipped folder in *spark-1.6.3-bin-hadoop2.4* directory

Steps to run the jar files -
Task - 1
./bin/spark-submit --class "HW3Task1" --master local[4]
./SaiSree_Kamineni_hw3/Solution/SaiSree_Kamineni_task1.jar
./SaiSree_Kamineni_hw3/Input/ratings.csv ./SaiSree_Kamineni_hw3/Input/testing_small.csv

Output will be stored in current directory that is spark-1.6.3-bin-hadoop2.4 with name SaiSree_Kamineni_result_task1.txt/part-00000

Task -2
./bin/spark-submit --class "HW3Task2" --master local[4]
./SaiSree_Kamineni_hw3/Solution/SaiSree_Kamineni_task2.jar
./SaiSree_Kamineni_hw3/Input/ratings.csv ./SaiSree_Kamineni_hw3/Input/testing_small.csv

Output will be stored in current directory that is spark-1.6.3-bin-hadoop2.4 with name SaiSree_Kamineni_result_task2.txt/part-00000

Output format
As mentioned in the problem statement.
Task-1

```
UserId,MovieId,Pred_rating
1,1172,1.5635267889731306
1,1405,2.5034098137841347
1,2193,3.0211803063635245
1,2968,2.3165438621827574
2,52,3.5679453086434565
2,144,2.766309925610061
2,248,2.358145702961475
2,314,2.9316203942882915
2,319,3.6920442782835687
2,370,2.9620403364405155
2,371,2.461985645533563
2,372,2.7224500003022896
2,382,3.481713344843987
2,405,2.291103371446261
```

Task-2

```
UserId,MovieId,Pred_rating
1,1172,2.9550965045753217
1,1405,2.566373045195705
1,2193,2.0963619756235503
1,2968,1.9021411224810363
2,52,3.8200881961443454
2,144,3.2601656150707994
2,248,3.0310648814694763
2,314,4.75178274485157
2,319,4.80848706931415
2,370,2.890601028395516
2,371,3.30668425285272
2,372,3.6515291509314673
2,382,3.825558793935305
2,405,2.82276838629143
```

Accuracy –

Task - 1

>=0 and <1: 17038
>=1 and <2: 2494
>=2 and <3: 546
>=3 and <4: 142
>=4: 36
RMSE = 1.0722270192897556
The total execution time taken is 12.125 sec

Task - 2

>=0 and <1: 17679

>=1 and <2: 2168
>=2 and <3: 326
>=3 and <4: 57
>=4: 26
RMSE = 1.0277773956242549
The total execution time taken is 88.29 sec

Algorithm Used – User based collaborative filtering. Calculated weights predictions using the following formulae. Used the neighborhood approach where the size is 4.

Weights between users -

$$w_{u,v} = \frac{\sum_{i \in I} (r_{u,i} - \overline{r}_u)(r_{v,i} - \overline{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \overline{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \overline{r}_v)^2}}$$

Prediction -

$$P_{a,i} = \overline{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \overline{r}_u) \cdot w_{a,u}}{\sum_{u \in U} | w_{a,u} |}$$

Approach –
Handling outliers – If ratings are greater than 5, I replaced them with 5 and if they are less than 0, I replaced them with 0.

In case of missing ratings, I filled it with the average rating of that user over all the movies in training data.