Recommendation System

Environment requirements- Scala 2.11 and Spark 2.3.1

1. Configure spark environment.sh to meet the execution time requirements

2. Set $SPARK_HOME to the directory in which spark-2.3.1-bin- hadoop2.7 is stored
   Run source ~/.bashrc

   Where .bashrc file which is in home directory and contains
   export SPARK_HOME=$HOME/Downloads/spark-2.3.1-bin-hadoop2.7
   export PATH=$PATH:~/opt/node/bin

<u>Task 1: Model Based Collaborative Filtering</u>

```
$SPARK_HOME/bin/spark-submit --class NamanaMadanMohanRao_Pawar_ModelBasedCF
NamanaMadanMohanRao_Pawar_hw2.jar <Path of ratings.csv file> <Path of
test.csv file>
```

Parameters: rank = 2, iterations = 21, lambda_ = 0.28

Output:

```
>=0 and <1: 29192
>=1 and <2: 13396
>=2 and <3: 2311
>=3 and <4: 303
>4: 32
Root Mean Squared Error = 1.0762807101689864
Time = 31.032113412
```

Results of model-based CF have been written to output file
“NamanaMadanMohanRao_Pawar_ModelBasedCF.txt”

- Handled missing rating for new users/products by predicting the rating as the average rating of the concerned user or the average rating of the concerned product respectively. If both user and product are new, then an average value of 2.5 is predicted

Task 2: Item Based Collaborative Filtering

```
$SPARK_HOME/bin/spark-submit --class NamanaMadanMohanRao_Pawar_ItemBasedCF
NamanaMadanMohanRao_Pawar_hw2.jar <Path of ratings.csv file> <Path of
test.csv file>
```

Output:

```
>=0 and <1: 28923.0
>=1 and <2: 13159.0
>=2 and <3: 2615.0
>=3 and <4: 493.0
>4: 0
Root Mean Squared Error = 1.0919400169631637
Time = 92.265144454
```

Results of item-based CF have been written to output file "NamanaMadanMohanRao_Pawar_ItemBasedCF.txt"

- Handled cold start and NaN values for Pearson correlation for new users/products by predicting the rating as the average rating of the concerned user or the average rating of the concerned product respectively. If both user and product are new then an average value of 2.5 is predicted
- If predictions were calculated as NaN then the value is substituted by the rating of the item which is most similar to the concerned item
- While calculating Pearson correlation of two items, if the two items have no common users who rated both the items the value is predicted as a low constant to state that the two items are not collectively preferred by users