Assignment 2

Versions of frameworks

Scala: 2.11.12

Spark: 2.3.1

Running of the program

The jar file produces 2 output text files based on the models passed in the argument to the file. In order to run the file the following commands should be used:

spark-submit –class CLASSNAME Utkarsh_Gera_hw3.jar <rating file path> <testing file path>

where the "CLASSNAME" can be ModelBasedCF or UserBasedCF.

The program is going to produce the output file in the directory it is called from. So the program should have write access to that directory.

ModelBasedCF requires sparkcontext checkpoint which will produce a temporary folder in the directory for the storage of RDD to overcome the problem of stack overflow. However while exiting the program it gets deleted if it doesn't face any exceptions in doing so.

As mentioned in one of the posts in the discussion board – thread 50 by Niharika Gajam The output in the txt file is of form

"userid1","businessid1","4"
"userid2","businessid2","4"
"userid3","businessid3","4"
(The keys or identifiers have double quotes around them)

If the program finds any issue in the file paths or you don't provide these 2 path arguments/provide more than 2 arguments the program will exit with a proper message like:

Need at two arguments - Input rating file and test file

Performance of the program

1. ModelBasedCF

>=0 and <1: 33621
>=1 and <2: 9549
>=2 and <3: 1120
>=3 and <4: 127
>=4: 43
RMSE: 1.0745433191762725
Time: 191 sec

2. UserBasedCF

>=0 and <1: 37062
>=1 and <2: 7866
>=2 and <3: 289
>=3 and <4: 15
>=4: 4

RMSE: 1.062908777455307
Time: 163 sec

Improvements

There were couple of tweaks that were done in the UserBasedCF

1. It stored the entire RDD as a hashmap for faster access
2. Only the users who have rated more than 5 items were considered for similarity with active user
3. Minimum 2 items need to be co-rated by 2 users before calculating the weights for Pearson's correlation constant.
4. If the active user had rated less than 5 items then the average rating by that user was used as default.
5. In case both user id and business id was encountered for the first time average rating of 2.5 was used, but in case one of them was encountered for first time then average of the latter was used.