Hardik Jain
USC ID – 5623917353

Assignment 3 Description

LSH:

Below is the command to run the program:

bin/spark-submit Hardik_Jain_task1_Jaccard.py <ratings.csv path>

For scala:

bin/spark-submit –class JaccardLSH Hardik_Jain_hw3.jar <ratings file name>

Precision = 1.0
Recall = 0.9999375379001804
Time = 233.6866729259491

ModelBasedCF: using rank = 6, number of iterations = 10

Small Dataset:

Below is the command to run the program:

bin/spark-submit Hardik_Jain_task2_ModelBasedCF.py <ratings_small.csv path> <testing_small.csv path>

For scala:

bin/spark-submit –class ModelBasedCF Hardik_Jain_hw3.jar <ratings file name> <testing_small.csv filename>

Big Dataset:

Note: The zip file uploaded does not contain the output for this dataset due to issues on blackboard.

Below is the command to run the program:

bin/spark-submit Hardik_Jain_task2_ModelBasedCF.py <ratings_big.csv path> <testing_20m.csv path>

For Scala:

bin/spark-submit –class ModelBasedCF Hardik_Jain_hw3.jar <ratings file name> <testing_20m.csv filename>

UserBasedCF:

Below is the command to run the program:

bin/spark-submit Hardik_Jain_task2_UserBasedCF.py <ratings_small.csv path> <testing_small.csv path>

ItemBasedCF:

I am using the same code that I had written for LSH to find out movies with a Jaccard Similarity greater than 0.5. However, here I change the threshold for Jaccard Similarity to 0.1 and used only these similar movies to predict ratings. Also, after computing similarities, I have subtracted testing_small from ratings and then made predictions.

Below is the command to run the program for ItemBased with LSH:

bin/spark-submit Hardik_Jain_task2_ItemBasedCF.py <ratings_small.csv path> <testing_small.csv path>

Below is the command to run the program for ItemBased without LSH:

bin/spark-submit Hardik_Jain_task2_ItemBasedCF_withoutLSH.py <ratings_small.csv path> <testing_small.csv path>

Accuracy of ItemBased CF with LSH is as follows:
>=0 and <1: 15351
>=1 and <2: 4023
>=2 and <3: 751
>=3 and <4: 127
>=4: 4
RMSE: 0.9316358720931318
Time: 309.67094564437866

Accuracy of ItemBased CF without LSH is as follows:

>=0 and <1: 5684
>=1 and <2: 5760
>=2 and <3: 4759
>=3 and <4: 2502
>=4: 1551
RMSE: 2.3443435824686003
Time: 140.07782673835754

To improve RMSE, while predicting a rating, I only chose those combinations that have a correlation greater than 0. This improvement, gave me the following result:

>=0 and <1: 13208
>=1 and <2: 5263
>=2 and <3: 1404
>=3 and <4: 319
>=4: 62
RMSE: 1.0887322491102356
Time: 115.29548192024231

Accuracy of Recommendation System:

	Task1 (ModelBased CF)		Task2 (User Based CF)
	Small	Big	Small
>=0 and <1	12878	3243197	14961
>=1 and <2	4343	699523	4424
>=2 and <3	1126	90622	721
>=3 and <4	292	11594	139
>=4	94	1395	11
RMSE	1.151788512464	0.82862935873	0.9543529761570
Time	12.079508304595	692.8881096839	16.3709540367126