Adam Vaccaro
INF 553
Professor Yao-Yi Chiang
USC Spring 2018

Description for Assignment 3

Spark version: 2.2.1
Scala version: 2.11 (2.10 for Task 1: Cosine LSH)

Task 1: Jaccard LSH

To submit using spark-submit:
spark-submit --class JaccardLSH Adam_Vaccaro_hw3.jar <input path> <output path>

Precision: 1.0
Recall: 0.81119525

Task 1: Cosine LSH

Precision: 0.94031835
Recall: 0.77129173

Task 2: Model-based CF

To submit using spark-submit:
spark-submit --class ModelBasedCF Adam_Vaccaro_hw3.jar <rating file path> <testing file path> <output file path>

Accuracy and runtime table:

>=0 and <1	3261
>=1 and <2	3383
>=2 and <3	964
>=3 and <4	92
>=4	0
RMSE	1.3899617662375763
Time (sec)	8

Task 2 Improvements: For each of the CF implementations in Task 2, I shifted the ratings in the training set down by 3 so that they ranged from -2 to 2 (instead of 1 to 5) and then shifted them back up to range from 1 to 5 before calculating RMSE. Also, predictions that were less than 1 were set to 1 and predictions that were greater than 5 were set to 5.

Task 2: User-based CF

To submit using spark-submit:
spark-submit --class UserBasedCF Adam_Vaccaro_hw3.jar <rating file path> <testing file path> <output file path>

Accuracy and runtime table:

>=0 and <1	1788
>=1 and <2	3513
>=2 and <3	1943
>=3 and <4	396
>=4	60
RMSE	1.446067055545343
Time (sec)	107

Task 2: Item-based CF

To submit using spark-submit (with LSH):
spark-submit --class ItemBasedCF Adam_Vaccaro_hw3.jar <rating file path> <similarity results path> <testing file path> <output file path>

*The similarity results path is the path to the .txt file generated from the Jaccard or Cosine similarity LSH in Task 1.

Accuracy and runtime table (with LSH):

>=0 and <1	1662
>=1 and <2	3628
>=2 and <3	2052
>=3 and <4	328
>=4	30
RMSE	1.4081396034687552
Time (sec)	19

To submit using spark-submit (without LSH):
spark-submit --class ItemBasedCFWithoutLSH Adam_Vaccaro_hw3.jar <rating file path> <testing file path> <output file path>

Accuracy and runtime table (without LSH):

>=0 and <1	1975
>=1 and <2	3364
>=2 and <3	1864
>=3 and <4	398
>=4	99
RMSE	1.462023329381675
Time (sec)	7

Effect of LSH: Here, the effect of LSH seems to be decreasing the RMSE and making the predictions more accurate, but at the cost of increasing the runtime of the algorithm. However, the increase in runtime might just be because the LSH similarity data was incorporated inefficiently in this implementation, and might not be true for all implementations of LSH in recommender systems.