CS 657 Massive Mining Datasets Assignment -4

Sai Sahana Bhargavi - G01330358

Introduction: The task is to design and implement a recommendation system to predict the ratings of the movies based on the utility matrix on the Movie Lens 20M dataset.

Approach:

1. Instantiated SparkSession for reading the data files.

```
# spark config
spark = SparkSession \
        .builder \
        .appName("movie recommendation") \
    .config("spark.driver.maxResultSize", "96g") \
        .config("spark.driver.memory", "96g") \
        .config("spark.executor.memory", "8g") \
    .config("spark.master", "local[12]") \
        .getOrCreate()
```

ALS: Alternating Least Squares (ALS) matrix factorization.
ALS attempts to estimate the ratings matrix R as the product of two lower-rank matrices, X and Y, i.e. X * Yt = R. Typically these approximations are called 'factor' matrices. The general approach is iterative. During each iteration, one of the factor matrices is held constant, while the other is solved for using least squares. The newly-solved factor matrix is then held constant while solving for the other factor matrix.

2. Read the ratings.csv file that consists of userId, movieId, ratings and the 80% of the data is used for cross-validation over ALS model for 10 folds using parameters for rank as [10, 20, 500, 100] and regularization parameter as [.01, .05, .1, .20] and used the rest of the dataset (20%) to test the system after tuning and computing RMSE, MSE, and MAE using RegressionEvaluator() as evaluator.
The obtained MSE,RMSE, and MAE values are 0.6177 0.7859 0.6080.

3. Item-Item-CF:
mapped ratings to key / value pairs: user ID;; movie ID, rating
normalize ratings using computed mean for ratings values ,here(value pairs=> (movie ID, rating)
used ratings.partitioned(100) to read the large scale data into selective partitions.
For userID => ((movieID, rating), (movieID, rating))
removed every movie rated together by the same user and further filtered out duplicate pairs
Collected all ratings for each movie pair (movie1, movie2) => (rating1, rating2)and computed cosine-similarity (Pearson correlation)

For the given movie-id, filtered out for movies with the similarities that are as defined by the quality thresholds scoreThreshold = 0.85 and coOccurenceThreshold = 10 and the top 10 results are extracted and predictions are calculated as follows.

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

s_{ij}... similarity of items ***i*** and ***j***
r_{xj}...rating of user ***x*** on item ***j***
N(i;x)... set items rated by ***x*** similar to ***i***

4. ALS + Item-Item CF:

ALS predictions
(Columns: _c2: actual, _c4:predicted)

_c0	_c1	_c2	_c3	_c4
12393	83	4.0	843336010	3.6577127
86771	83	3.0	851184922	3.5276146
37679	83	3.0	875952555	3.723758
20132	83	4.0	1001533223	3.864291
1133	83	4.0	969318619	3.3781295
94296	83	4.0	953575695	3.6423743
20519	83	4.0	1030245697	4.136854
41389	83	5.0	840444257	4.352285
49040	83	4.0	929574190	3.7647567
91643	83	4.0	1112671471	3.5994651
13139	83	4.0	861953728	3.6976643
97583	83	3.0	858955600	4.1892276
16078	83	2.0	858161643	2.675203
28049	83	1.0	864909452	2.106198
97606	83	4.0	850142904	3.605704
91696	83	5.0	833287855	3.779149
123067	83	3.0	860769357	3.1990082
30930	83	3.0	859193965	3.7036836
36508	83	3.0	852623481	3.8676968

only showing top 20 rows

Item-Item CF predictions

userId	movieId	rating	score
12393	83	4.0	3.275127
86771	83	3.0	3.52767
37679	83	3.0	3.423458
20132	83	4.0	3.864291
1133	83	4.0	3.3781295
94296	83	4.0	3.3423741
20519	83	4.0	3.8364854
41389	83	5.0	3.352285
49040	83	4.0	3.1647567
91643	83	4.0	3.5994651
13139	83	4.0	3.6976643
97583	83	3.0	3.892276
16078	83	2.0	1.675212
28049	83	1.0	1.65437
97606	83	4.0	3.901704
91696	83	5.0	3.61523
123067	83	3.0	3.823413
30930	83	3.0	3.5036836
36508	83	3.0	3.7567596

only showing top 20 rows

ALS + Item-Item CF predictions

uid1	mid1	new_pred
67196	83	3.42816
116846	83	3.42816
36508	83	3.52764
89414	83	3.54357
55848	83	3.86429
97707	83	3.37812
17535	83	3.46237
64154	83	3.95663
25771	83	3.75228
79192	83	3.40475
41389	83	3.59946
91696	83	3.69766
132237	83	4.01105
47002	83	2.07520
105485	83	1.83510
92598	83	3.78330
32878	83	3.68079
1133	83	3.57365
11931	83	3.58368
113669	83	3.80113

only showing top 20 rows

Combined predictions are calculated as 0.4* rating_item_item + 0.6 * rating_als

The RMSE, MSE, MAE scores have significantly improved when hybrid model is used
MSE : 0.346, RMSE : 0.451 , MAE: 0.3318.

5. Supervised Learning Model:

Combined both the features in movies.csv and ratings.csv into a single dataframe that containing title, genre, userId and respective ratings based on common movie-id and using Tokenizer and Word2Vec, the genres are broken into word-tokens and are vectorized and VectorAssembler combines the vectors into single set of features and ratings as labels and used linear regression model to predict ratings.
We joined the ratings dataframe with the movie_genres according to the user ID. It is done so that the we will get recommendations according to each user.For implementing the Linear regression supervised learning model the dataset is dense, which we converted to a sparse one.

```
[mpitale@perseus 4]$ hdfs dfs -mkdir /user/mpitale/ml-20m
[mpitale@perseus 4]$ hdfs dfs -put ml-20m/* /user/mpitale/ml-20m
[mpitale@perseus 4]$ python3 a.py
2022-12-01 03:35:20,942 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2022-12-01 03:35:22,558 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
2022-12-01 03:35:22,558 WARN util.Utils: Service 'SparkUI' could not bind on port 4041. Attempting port 4042.
2022-12-01 03:35:22,559 WARN util.Utils: Service 'SparkUI' could not bind on port 4042. Attempting port 4043.
+-------+--------------------+--------------------+
|movieId|               title|              genres|
+-------+--------------------+--------------------+
|      1|    Toy Story (1995)|Adventure|Animati...|
|      2|      Jumanji (1995)|Adventure|Childre...|
|      3|Grumpier Old Men ...|      Comedy|Romance|
|      4|Waiting to Exhale...|Comedy|Drama|Romance|
|      5|Father of the Bri...|              Comedy|
|      6|         Heat (1995)|Action|Crime|Thri...|
|      7|      Sabrina (1995)|      Comedy|Romance|
|      8| Tom and Huck (1995)|  Adventure|Children|
|      9| Sudden Death (1995)|              Action|
|     10|    GoldenEye (1995)|Action|Adventure|...|
|     11|American Presiden...|Comedy|Drama|Romance|
|     12|Dracula: Dead and...|       Comedy|Horror|
|     13|        Balto (1995)|Adventure|Animati...|
|     14|        Nixon (1995)|               Drama|
|     15|Cutthroat Island ...|Action|Adventure|...|
|     16|       Casino (1995)|         Crime|Drama|
|     17|Sense and Sensibi...|       Drama|Romance|
|     18|   Four Rooms (1995)|              Comedy|
|     19|Ace Ventura: When...|              Comedy|
|     20|  Money Train (1995)|Action|Comedy|Cri...|
+-------+--------------------+--------------------+
only showing top 20 rows
```

```
+-------+------+-----------+----------------+--------------------+
| userId|rating|genre_array|genre_array_indexed|              vector|
+-------+------+-----------+----------------+--------------------+
|  [490]|   3.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[126388]|   3.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|  [603]|   2.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[127245]|   1.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|  [741]|   4.5|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[127911]|   1.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|  [903]|   3.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[128653]|   2.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
| [1259]|   5.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[129522]|   5.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
| [1716]|   2.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[130122]|   3.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
| [1931]|   2.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[130456]|   1.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
| [2671]|   3.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[130531]|   1.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
| [3335]|   5.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[130986]|   3.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
| [3433]|   2.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
|[132093]|   2.0|    [Drama]|           [0.0]|[1, 0, 0, 0, 0, 0...|
+-------+------+-----------+----------------+--------------------+
only showing top 20 rows

+------+--------------------+--------------------+
|rating|              vector|      features_array|
+------+--------------------+--------------------+
|   3.0|[1, 0, 0, 0, 0, 0...|[490, 1, 0, 0, 0,...|
|   3.0|[1, 0, 0, 0, 0, 0...|[126388, 1, 0, 0,...|
|   2.0|[1, 0, 0, 0, 0, 0...|[603, 1, 0, 0, 0,...|
|   1.0|[1, 0, 0, 0, 0, 0...|[127245, 1, 0, 0,...|
|   4.5|[1, 0, 0, 0, 0, 0...|[741, 1, 0, 0, 0,...|
|   1.0|[1, 0, 0, 0, 0, 0...|[127911, 1, 0, 0,...|
|   3.0|[1, 0, 0, 0, 0, 0...|[903, 1, 0, 0, 0,...|
|   2.0|[1, 0, 0, 0, 0, 0...|[128653, 1, 0, 0,...|
|   5.0|[1, 0, 0, 0, 0, 0...|[1259, 1, 0, 0, 0...|
|   5.0|[1, 0, 0, 0, 0, 0...|[129522, 1, 0, 0,...|
|   2.0|[1, 0, 0, 0, 0, 0...|[1716, 1, 0, 0, 0...|
|   3.0|[1, 0, 0, 0, 0, 0...|[130122, 1, 0, 0,...|
|   2.0|[1, 0, 0, 0, 0, 0...|[1931, 1, 0, 0, 0...|
|   1.0|[1, 0, 0, 0, 0, 0...|[130456, 1, 0, 0,...|
|   3.0|[1, 0, 0, 0, 0, 0...|[2671, 1, 0, 0, 0...|
|   1.0|[1, 0, 0, 0, 0, 0...|[130531, 1, 0, 0,...|
|   5.0|[1, 0, 0, 0, 0, 0...|[3335, 1, 0, 0, 0...|
|   3.0|[1, 0, 0, 0, 0, 0...|[130986, 1, 0, 0,...|
|   2.0|[1, 0, 0, 0, 0, 0...|[3433, 1, 0, 0, 0...|
|   2.0|[1, 0, 0, 0, 0, 0...|[132093, 1, 0, 0,...|
+------+--------------------+--------------------+
```

rating	vector	features_array	features	prediction
1.0	[1, 0, 0, 0, 0, 0...	[131037, 1, 0, 0,...	[131037.0,1.0,0.0...	3.0828377397350346
1.0	[1, 0, 0, 0, 0, 0...	[131095, 1, 0, 0,...	[131095.0,1.0,0.0...	3.08267545126656
1.5	[1, 0, 0, 0, 0, 0...	[131698, 1, 0, 0,...	[131698.0,1.0,0.0...	3.0809882108098363
2.0	[1, 0, 0, 0, 0, 0...	[160, 1, 0, 0, 0,...	[160.0,1.0,0.0,0....	3.4490416688472862
2.0	[1, 0, 0, 0, 0, 0...	[417, 1, 0, 0, 0,...	[417.0,1.0,0.0,0....	3.448322563047322
2.0	[1, 0, 0, 0, 0, 0...	[749, 1, 0, 0, 0,...	[749.0,1.0,0.0,0....	3.4473936014691593
2.0	[1, 0, 0, 0, 0, 0...	[750, 1, 0, 0, 0,...	[750.0,1.0,0.0,0....	3.4473908033921163
2.0	[1, 0, 0, 0, 0, 0...	[774, 1, 0, 0, 0,...	[774.0,1.0,0.0,0....	3.4473236495430926
2.0	[1, 0, 0, 0, 0, 0...	[131078, 1, 0, 0,...	[131078.0,1.0,0.0...	3.0827230185762855
2.0	[1, 0, 0, 0, 0, 0...	[131629, 1, 0, 0,...	[131629.0,1.0,0.0...	3.0811812781257797
2.5	[1, 0, 0, 0, 0, 0...	[131, 1, 0, 0, 0,...	[131.0,1.0,0.0,0....	3.4491228130815235
2.5	[1, 0, 0, 0, 0, 0...	[637, 1, 0, 0, 0,...	[637.0,1.0,0.0,0....	3.4477069860979372
2.5	[1, 0, 0, 0, 0, 0...	[131086, 1, 0, 0,...	[131086.0,1.0,0.0...	3.082700633959944
2.5	[1, 0, 0, 0, 0, 0...	[131100, 1, 0, 0,...	[131100.0,1.0,0.0...	3.082661460881347
2.5	[1, 0, 0, 0, 0, 0...	[131516, 1, 0, 0,...	[131516.0,1.0,0.0...	3.0814974608316
3.0	[1, 0, 0, 0, 0, 0...	[25, 1, 0, 0, 0, ...	[25.0,1.0,0.0,0.0...	3.4494194092480455
3.0	[1, 0, 0, 0, 0, 0...	[95, 1, 0, 0, 0, ...	[95.0,1.0,0.0,0.0...	3.449223543855059
3.0	[1, 0, 0, 0, 0, 0...	[158, 1, 0, 0, 0,...	[158.0,1.0,0.0,0....	3.4490472650013713
3.0	[1, 0, 0, 0, 0, 0...	[164, 1, 0, 0, 0,...	[164.0,1.0,0.0,0....	3.4490304765391153
3.0	[1, 0, 0, 0, 0, 0...	[286, 1, 0, 0, 0,...	[286.0,1.0,0.0,0....	3.448689111139911

only showing top 20 rows

6. ALS + Item-Item CF+ Supervised Learning Model:

SVL	ALS+Item_CF	hybrid_pred
3.0828377397350346	3.42816	3.290031095894014
3.08267545126656	3.42816	3.2899661805066245
3.0809882108098363	3.52764	3.3489792843239345
3.4490416688472862	3.54357	3.505758667538914
3.448322563047322	3.86429	3.697903025218929
3.4473936014691593	3.37812	3.405829440587664
3.4473908033921163	3.46237	3.4563783213568464
3.4473236495430926	3.95663	3.752907459817237
3.0827230185762855	3.75228	3.484457207430514
3.0811812781257797	3.40475	3.275322511250312
3.4491228130815235	3.59946	3.53932512523261
3.4477069860979372	3.69766	3.597678794439175
3.082700633959944	4.01105	3.639710253583978
3.082661460881347	2.0752	2.478184584352539
3.0814974608316	1.8351	2.33365898433264
3.4494194092480455	3.7833	3.649747763699218
3.449223543855059	3.68079	3.588163417542024
3.4490472650013713	3.57365	3.5238089060005486
3.4490304765391153	3.58368	3.5298201906156463
3.448689111139911	3.80113	3.6601536444559644

The RMSE, MSE, MAE scores have significantly improved when hybrid model is used MSE : 0.296, RMSE : 0.381 , MAE: 0.2398.