OVERVIEW OF DB:

```
Dataset Overview:
Total records: 7824482

Ratings distribution:
+------+-------+
|rating|  count|
+------+-------+
|   1.0| 901765|
|   2.0| 456322|
|   3.0| 633073|
|   4.0|1485781|
|   5.0|4347541|
+------+-------+


Sample of the data:
+--------------+----------+------+----------+
|       user_id|product_id|rating| timestamp|
+--------------+----------+------+----------+
| AKM1MP6P0OYPR|0132793040|   5.0|1365811200|
|A2CX7LUOHB2NDG|0321732944|   5.0|1341100800|
|A2NWSAGRHCP8N5|0439886341|   1.0|1367193600|
|A2WNBOD3WNDNKT|0439886341|   3.0|1374451200|
|A1GI0U4ZRJA8WN|0439886341|   1.0|1334707200|
+--------------+----------+------+----------+
only showing top 5 rows

Preprocessing data...
Finding similar products for product ID: 0321732944

Similar Products:
+--------------+----------+------+----------+
|       user_id|product_id|rating| timestamp|
+--------------+----------+------+----------+
|A3169ZUL2I57NO|B00AQDG9BA|   5.0|1392163200|
|A2YPTBUWEAMXIB|B000H9J3ZC|   5.0|1393459200|
|A3GOUKUM7PQ6EC|B007WTAJTO|   5.0|1363737600|
| AHO3ONWH3S63H|B000H9J3ZC|   5.0|1314144000|
|A25YCUIR1YC8HM|B0049P6OTI|   5.0|1342396800|
| A2KL86JGIDM7S|B000H9J3ZC|   5.0|1255046400|
|A2Q8YOYXC8892R|B007WTAJTO|   5.0|1405036800|
| AZDJN0CVUU0YN|B000H9J3ZC|   5.0|1318464000|
|A2LKQ20NKZM69I|B0030AZ43A|   5.0|1359331200|
| AMFD3HZR44KJZ|B000H9J3ZC|   5.0|1366588800|
|A1TXAE9VA3POAX|B007WTAJTO|   5.0|1361923200|
| AE48NNP81HR76|B000H9J3ZC|   5.0|1371340800|
|A35J8CE8NE942R|B0049P6OTI|   5.0|1404950400|
|A3TBDSHPW81PZN|B000H9J3ZC|   5.0|1266883200|
|A2FH9PIX34CS7I|B007WTAJTO|   5.0|1381363200|
| AMGUWDM3XPEEB|B000H9J3ZC|   5.0|1267401600|
|A3LP7BQGOZSZPZ|B005H4CDF4|   5.0|1393286400|
| AMLUFPB784PJB|B000H9J3ZC|   5.0|1330041600|
| AJYEAYSX7QSN1|B007WTAJTO|   5.0|1383523200|
|A3BMC6FKX00U4H|B000H9J3ZC|   5.0|1316044800|
+--------------+----------+------+----------+
only showing top 20 rows
```

PRE-PROCESSING:

```
/usr/local/bin/python3 "/Users/santosh/Desktop/CL TASK1/advsimsearch.py"
(base) santosh@Santoshs-MacBook-Air CL TASK1 % /usr/local/bin/python3 "/Users/santosh/Desktop/CL TASK1/advsimsearch.py"
24/12/07 19:49:56 WARN Utils: Your hostname, Santoshs-MacBook-Air.local resolves to a loopback address: 127.0.0.1; using 192.0.0.2 instead (on inte
rface en0)
24/12/07 19:49:56 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/12/07 19:49:56 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
TF-IDF Output:
24/12/07 19:50:03 WARN DAGScheduler: Broadcasting large task binary with size 4.0 MiB
24/12/07 19:50:03 WARN DAGScheduler: Broadcasting large task binary with size 4.0 MiB
+----------+--------------------+--------------------------------------------------------------------------------------------------------+
|product_id|review              |tfidf_features                                                                                          |
+----------+--------------------+--------------------------------------------------------------------------------------------------------+
|0132793040|Great product       |(262144,[52879,261870],[1.157452788691043,2.302585092994046])                                           |
|0321732944|Good quality        |(262144,[43890,113432],[1.6835458845878224,1.4152818979931427])                                         |
|0439886341|Not satisfied       |(262144,[180097],[3.1498829533812494])                                                                  |
|0439886341|Average product     |(262144,[52879,98221],[1.157452788691043,3.5553480614894135])                                           |
|0439886341|Terrible product    |(262144,[52879,239452],[1.157452788691043,2.8622008809294686])                                          |
|0511189877|Excellent product   |(262144,[52879,78745],[1.157452788691043,2.302585092994046])                                            |
|0511189877|Not worth it        |(262144,[51247],[3.1498829533812494])                                                                   |
|0511189877|Good value for money|(262144,[113432,123499,134711],[1.4152818979931427,3.1498829533812494,2.8622008809294686])              |
|0511189877|Really good quality |(262144,[43890,113432,229264],[1.6835458845878224,1.4152818979931427,2.8622008809294686])               |
|0511189877|Fantastic purchase  |(262144,[182344,199176],[3.1498829533812494,2.8622008809294686])                                        |
|0511189877|Very satisfied      |(262144,[180097],[3.1498829533812494])                                                                  |
|0528881469|Highly recommended  |(262144,[13790,19633],[3.5553480614894135,3.5553480614894135])                                          |
|0528881469|Poor quality        |(262144,[43890,85735],[1.6835458845878224,2.6390573296152584])                                          |
|0528881469|Very good product   |(262144,[52879,113432],[1.157452788691043,1.4152818979931427])                                          |
|0528881469|Not good at all     |(262144,[113432],[1.4152818979931427])                                                                  |
|0528881469|Great value for money|(262144,[123499,134711,261870],[3.1498829533812494,2.8622008809294686,2.302585092994046])              |
|0528881469|Okay product        |(262144,[52879,261610],[1.157452788691043,2.8622008809294686])                                          |
|0528881469|Not up to the mark  |(262144,[159292],[3.5553480614894135])                                                                  |
|0528881469|Could be better     |(262144,[235375],[2.8622008809294686])                                                                  |
|0528881469|Very good           |(262144,[113432],[1.4152818979931427])                                                                  |
+----------+--------------------+--------------------------------------------------------------------------------------------------------+
only showing top 20 rows
```

SIMILARITY CALCULATIONS:

```
24/12/07 19:50:03 WARN DAGScheduler: Broadcasting large task binary with size 4.1 MiB
Cosine Similarity Matrix:
 [[6.64159507 0.         0.         ... 0.         0.         1.33969696]
 [0.         4.8373496  0.         ... 4.8373496  0.         0.        ]
 [0.         0.         9.92176262 ... 0.         0.         0.        ]
 ...
 [0.         4.8373496  0.         ... 4.8373496  0.         0.        ]
 [0.         0.         0.         ... 0.         8.19219388 0.        ]
 [1.33969696 0.         0.         ... 0.         0.         8.30432055]]
24/12/07 19:50:05 WARN DAGScheduler: Broadcasting large task binary with size 4.1 MiB
Jaccard Similarity Matrix:
 [[1.         0.         0.         ... 0.         0.         0.18977451]
 [0.         1.         0.         ... 1.         0.         0.        ]
 [0.         0.         1.         ... 0.         0.         0.        ]
 ...
 [0.         1.         0.         ... 1.         0.         0.        ]
 [0.         0.         0.         ... 0.         1.         0.        ]
 [0.18977451 0.         0.         ... 0.         0.         1.        ]]
24/12/07 19:50:10 WARN DAGScheduler: Broadcasting large task binary with size 4.1 MiB
Laplacian Similarity Matrix:
 [[1.00000000e+00 1.03456474e-05 6.40656378e-08 ... 1.03456474e-05
  3.61216278e-07 4.70670863e-06]
 [1.03456474e-05 1.00000000e+00 3.89223458e-07 ... 1.00000000e+00
  2.19452820e-06 1.96175683e-06]
 [6.40656378e-08 3.89223458e-07 1.00000000e+00 ... 3.89223458e-07
  1.35896617e-08 1.21482202e-08]
 ...
 [1.03456474e-05 1.00000000e+00 3.89223458e-07 ... 1.00000000e+00
  2.19452820e-06 1.96175683e-06]
 [3.61216278e-07 2.19452820e-06 1.35896617e-08 ... 2.19452820e-06
  1.00000000e+00 6.84943601e-08]
 [4.70670863e-06 1.96175683e-06 1.21482202e-08 ... 1.96175683e-06
  6.84943601e-08 1.00000000e+00]]
24/12/07 19:50:13 WARN DAGScheduler: Broadcasting large task binary with size 4.1 MiB
24/12/07 19:50:15 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC), users should configure it(them
) to spark.eventLog.gcMetrics.youngGenerationGarbageCollectors or spark.eventLog.gcMetrics.oldGenerationGarbageCollectors
Euclidean Similarity Matrix:
 [[0.         3.38805913 4.06981052 ... 3.38805913 3.85146582 3.50235945]
 [3.38805913 0.         3.841759   ... 0.         3.60964589 3.62514415]
 [4.06981052 3.841759   0.         ... 3.841759   4.2560494  4.2692017 ]
 ...
 [3.38805913 0.         3.841759   ... 0.         3.60964589 3.62514415]
 [3.85146582 3.60964589 4.2560494  ... 3.60964589 0.         4.06159014]
 [3.50235945 3.62514415 4.2692017  ... 3.62514415 4.06159014 0.        ]]
24/12/07 19:50:24 WARN DAGScheduler: Broadcasting large task binary with size 4.1 MiB
```

```
Manhattan Similarity Matrix:
 [[0.         6.55886566 6.60992084 ... 6.55886566 6.32223876 4.94164242]
 [6.55886566 0.         6.24871074 ... 0.         5.96102866 6.8953379 ]
 [6.60992084 6.24871074 0.         ... 6.24871074 6.01208383 6.94639307]
 ...
 [6.55886566 0.         6.24871074 ... 0.         5.96102866 6.8953379 ]
 [6.32223876 5.96102866 6.01208383 ... 5.96102866 0.         6.658711  ]
 [4.94164242 6.8953379  6.94639307 ... 6.8953379  6.658711   0.        ]]
24/12/07 19:50:36 WARN DAGScheduler: Broadcasting large task binary with size 4.1 MiB
Hamming Similarity Matrix:
 [[0 4 3 ... 4 3 2]
 [4 0 3 ... 0 3 4]
 [3 3 0 ... 3 2 3]
 ...
 [4 0 3 ... 0 3 4]
 [3 3 2 ... 3 0 3]
 [2 4 3 ... 4 3 0]]
```

GRAPH - COMPARISSION