实现方案

1. 从movies.txt文件中提取出每条评论的productID和userID
2. 过滤掉不是电影的productID

   若某些product页面不存在导演、演员信息，或者该product的时长过短，则判断其不为电影，直接过滤。最终得到movie.csv文件。
3. 自定义规则，得到同一系列的电影的ID集合

   将上述的movie.csv文件根据电影名进行排序，计算各电影名的编辑距离，若编辑距离小于某一阈值，则判断其为同一系列的电影，得到IP.csv文件。
4. 将所有电影的movieID和userID存入neo4j图数据库，而电影以及用户的其余信息存储在mysql数据库中。在neo4j中存储时将movieID和userID设为节点，将评论设为关系

   （1）将所有电影的信息单独存储为一个moviesId_list.csv文件，并设置csv表头，记录csv第一列为movieId，第二列是节点的LABEL，均为movie

1	movieId:ID(node)	:LABEL
2	B00187MZH0	movie
3	B0006HC06E	movie
4	B000UINP1S	movie
5	B0013B34X0	movie
6	B003QP4CNC	movie
7	B00004CWJ9	movie

   （2）将所有用户的信息单独存储为一个userId_list.csv文件，并设置csv表头，记录csv第一列为userId，第二列是节点的LABEL，均为user

1	userId:ID(node)	:LABEL
2	A1MDCM35WOMTX5	user
3	A2SSV5IV6HIB6M	user
4	A2LZKUARB5Y8C2	user
5	A39WT0YMCZISH4	user
6	AE5T5C6CPVT66	user
7	A3BTDYUGZ7RT00	user

   （3）将user与movie的评论关系存储为一个moviesId-userId.csv文件，设置表头，记录csv第一列是关系的终点，第二列是关系的七点，第三列是关系的类型，均为review

1	:END_ID(node)	:START_ID(node)	:TYPE
2	B003AI2VGA	A141HP4LYPWMSR	review
3	B003AI2VGA	A328S9RN3U5M68	review
4	B003AI2VGA	A1I7QGUDP043DG	review
5	B003AI2VGA	A1M5405JH9THP9	review
6	B003AI2VGA	ATXL536YX71TR	review
7	B003AI2VGA	A3QYDL5CDNYN66	review

   （4）进入数据库的bin文件夹，利用neo4j-admin import 工具将上述csv文件导入一个空的neo4j数据库

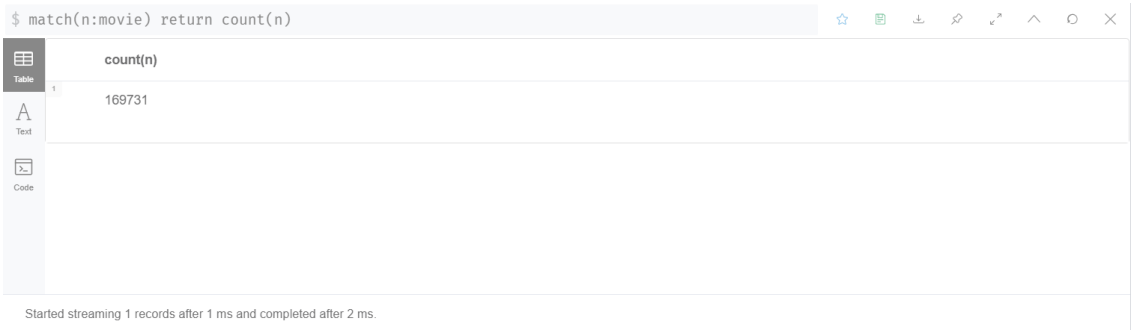
导入时设置上述moviesId_list.csv，userId_list.csv为nodes文件，设置moviesId-userId.csv为relationship文件，设置导入的id-type为STRING，导入结果如下

```
PS C:\Users\cheng fu\AppData\Local\Neo4j\Relate\Data\dbmss\dbms-4fae42e5-dd19-4c26-80a1-f083e2027951\bin> .\neo4j-admin import --nodes "\moviesId_list.csv" --nodes "\userId_list.csv" --relationships "\moviesId-userId.csv" --ignore-missing-nodes --id-type=STRING
Neo4j version: 3.5.23
Importing the contents of these files into C:\Users\cheng fu\AppData\Local\Neo4j\Relate\Data\dbmss\dbms-4fae42e5-dd19-4c26-80a1-f083e2027951\data\databases\graph.db:
Nodes:
  \moviesId_list.csv

  \userId_list.csv
Relationships:
  \moviesId-userId.csv

Available resources:
  Total machine memory: 15.83 GB
  Free machine memory: 7.36 GB
  Max heap memory : 3.52 GB
  Processors: 8
  Configured max memory: 11.08 GB
  High-IO: false

Import starting 2020-11-13 22:34:59.478+0800
  Estimated number of nodes: 863.04 k
  Estimated number of node properties: 863.04 k
  Estimated number of relationships: 6.67 M
  Estimated number of relationship properties: 0.00
  Estimated disk space usage: 263.23 MB
  Estimated required memory usage: 1.01 GB

InteractiveReporterInteractions command list (end with ENTER):
  c: Print more detailed information about current stage
  i: Print more detailed information

(1/4) Node import 2020-11-13 22:34:59.668+0800
  Estimated number of nodes: 863.04 k
  Estimated disk space usage: 46.91 MB
  Estimated required memory usage: 1.01 GB
..........  ..........  ..........  ..........  ..........   5% ?855ms
..........  ..........  ..........  ..........  ......-...  10% ?63ms
..........  ..........  ..........  ..........  ..........  15% ?5ms
..........  ..........  ..........  ..........  ..........  20% ?0ms
..........  ..........  ..........  ..........  ..........  25% ?290ms
..........  ..........  ..........  ..........  ..........  30% ?207ms
..........  ..........  ..........  ..........  ..........  35% ?2ms
..........  ..........  ..........  ..........  ..........  40% ?0ms
..........  ..........  ..........  ..........  ..........  45% ?0ms
..........  ..........  ..........  ..........  ..........  50% ?0ms
..........  ..........  ..........  ..........  ..........  55% ?0ms
..........  ..........  ..........  ..........  ..........  60% ?0ms
..........  ..........  ..........  ..........  ..........  65% ?16ms
..........  ..........  ..........  ..........  ..........  70% ?0ms
..........  ..........  ..........  ..........  ..........  75% ?0ms
```

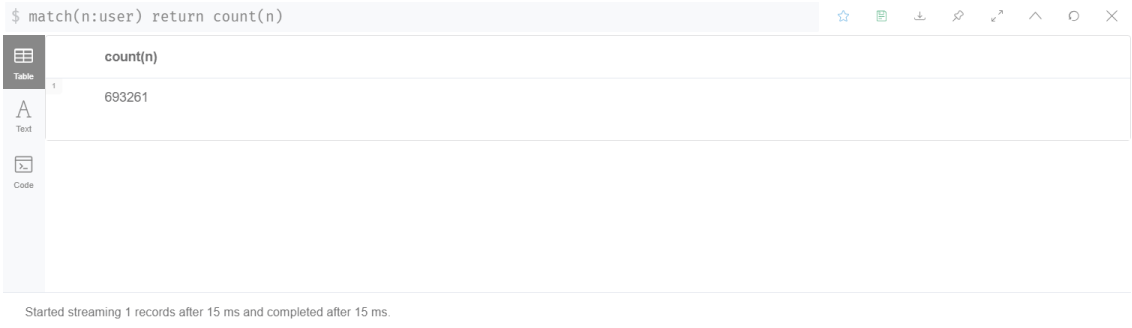
```
..........  ..........  ..........  ..........  ..........  80% ?0ms
..........  ..........  ..........  ..........  ..........  85% ?0ms
..........  ..........  ..........  ..........  ..........  90% ?0ms
..........  ..........  ..........  ..........  ..........  95% ?10ms
..........  ..........  ..........  ..........  .......... 100% ?0ms

(2/4) Relationship import 2020-11-13 22:35:01.268+0800
  Estimated number of relationships: 6.67 M
  Estimated disk space usage: 216.31 MB
  Estimated required memory usage: 1.01 GB
..........  ..........  ..........  ..........  ..........   5% ?670ms
..........  ..........  ..........  ..........  ..........  10% ?435ms
..........  ..........  ..........  ..........  ..........  15% ?208ms
..........  ..........  ..........  ..........  ..........  20% ?2ms
..........  ..........  ..........  ..........  ..........  25% ?4ms
..........  ..........  ..........  ..........  ..........  30% ?485ms
..........  ..........  ..........  ..........  ..........  35% ?8ms
..........  ..........  ..........  ..........  ..........  40% ?268ms
..........  ..........  ..........  ..........  ..........  45% ?248ms
..........  ..........  ..........  ..........  ..........  50% ?0ms
..........  ..........  ..........  ..........  ..........  55% ?2ms
..........  ..........  ..........  ..........  ..........  60% ?465ms
..........  ..........  ..........  ..........  ..........  65% ?5ms
..........  ..........  ..........  ..........  ..........  70% ?5ms
..........  ..........  ..........  ..........  ..........  75% ?549ms
..........  ..........  ..........  ..........  ..........  80% ?6ms
..........  ..........  ..........  ..........  ..........  85% ?208ms
..........  ..........  ..........  ..........  ..........  90% ?230ms
..........  ..........  ..........  ..........  ..........  95% ?3ms
..........  ..........  ..........  ..........  .......... 100% ?1ms

(3/4) Relationship linking 2020-11-13 22:35:05.071+0800
  Estimated required memory usage: 1.00 GB
..........  ..........  ..........  ..........  ..........   5% ?203ms
..........  ..........  ..........  ..........  ..........  10% ?204ms
..........  ..........  ..........  ..........  ....-.....  15% ?260ms
..........  ..........  ..........  ..........  ..........  20% ?240ms
..........  ..........  ..........  ..........  ..........  25% ?0ms
..........  ..........  ..........  ..........  ..........  30% ?0ms
..........  ..........  ..........  ..........  ..........  35% ?201ms
..........  ..........  ..........  ..........  ..........  40% ?1ms
..........  ..........  ..........  ..........  ..........  45% ?221ms
..........  ..........  ..........  ..........  ..........  50% ?7ms
..........  ....-.....  ..........  ..........  ..........  55% ?200ms
..........  ..........  ..........  ..........  ..........  60% ?200ms
..........  ..........  ..........  ..........  ..........  65% ?200ms
..........  ..........  ..........  ..........  ..........  70% ?200ms
..........  ..........  ..........  ..........  ..........  75% ?200ms
..........  ..........  ..........  ..........  ..........  80% ?0ms
..........  ..........  ..........  ..........  ..........  85% ?200ms
..........  ..........  ..........  ..........  ..........  90% ?0ms
..........  ..........  ..........  ..........  ..........  95% ?450ms
..........  ..........  ..........  ..........  .......... 100% ?0ms

(4/4) Post processing 2020-11-13 22:35:08.738+0800
  Estimated required memory usage: 1020.01 MB
-.........  ..........  ..........  ..........  ..........   5% ?207ms
..........  ..........  ..........  ..........  ..........  10% ?1ms
```

```
..........  ..........  ..........  ..........  ..........  15% ?2ms
..........  ..........  ..........  ..........  ..........  20% ?1ms
..........  ..........  ..........  ..........  ..........  25% ?1ms
..........  ..........  ..........  ..........  ..........  30% ?201ms
..........  ..........  ..........  ..........  ..........  35% ?1ms
..........  ..........  ..........  ..........  ..........  40% ?1ms
..........  ..........  ..........  ..........  ..........  45% ?1ms
..........  ..........  ..........  ..........  ..........  50% ?1ms
..........  ..........  ..........  ..........  ..........  55% ?1ms
..........  ..........  ..........  ..........  ..........  60% ?1ms
..........  ..........  ..........  ..........  ..........  65% ?1ms
..........  ..........  ..........  ..........  ..........  70% ?1ms
..........  ..........  ..........  ..........  ..........  75% ?201ms
..........  ..........  ..........  ..........  ..........  80% ?1ms
..........  ..........  ..........  ..........  ..........  85% ?0ms
..........  ..........  ..........  ..........  ..........  90% ?1ms
..........  ..........  ..........  ..........  ..........  95% ?1ms
..........  ..........  ..........  ..........  .......... 100% ?1ms


IMPORT DONE in 11s 181ms.
Imported:
  862992 nodes
  6670561 relationships
  862992 properties
Peak memory usage: 1.04 GB
```

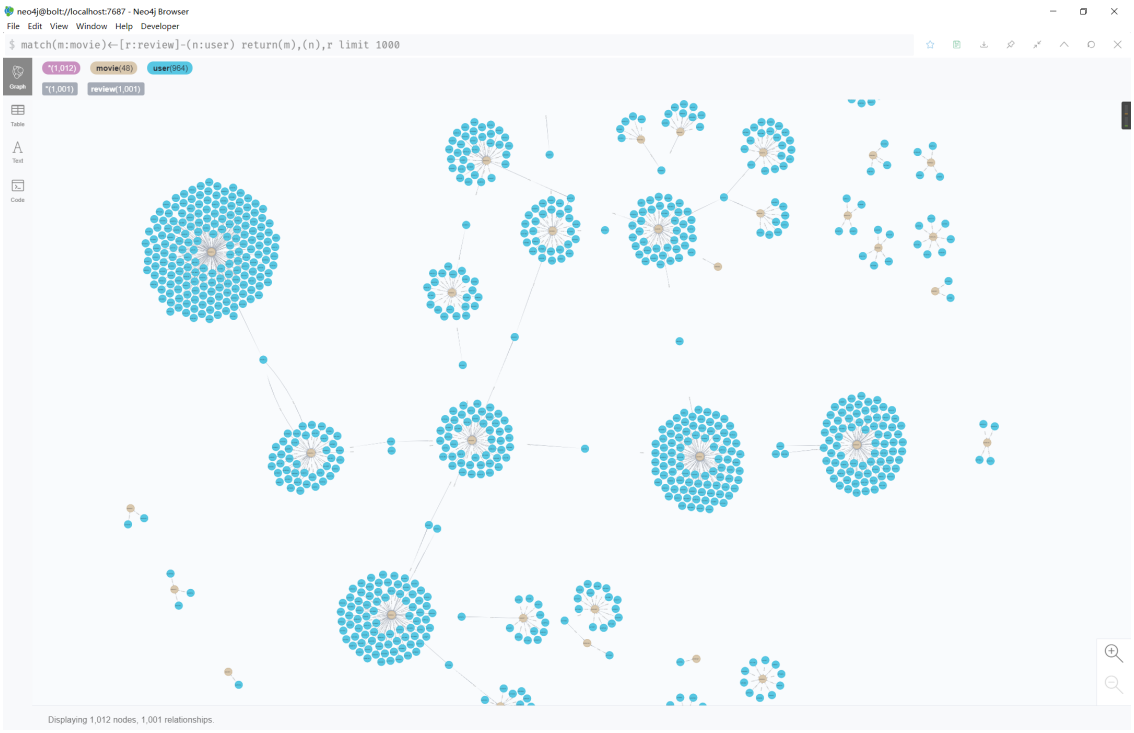
共导入169731部电影：



693261个用户



6670561个评论关系



5. 查询所有系列电影的评论人集合，得到最多的用户集合

   运行neo4j.py文件，查询上述步骤提取出的各个系列的电影的评论用户集合，得到最多的用户集合数量为5737，将用户userId写出为result.txt中

遇到的问题

match、merge、create语句速度过慢

在导入数据时，最初使用match、merge、create等语句进行数据插入，性能过慢，将数据全部导入完毕需要几百个小时……最终改为了使用neo4j-admin import直接进行导入

neo4j-admin import进行导入时jdk版本不兼容

```
警告: ERROR! Neo4j cannot be started using java version 1.8.0_261
警告: * Please use Oracle(R) Java(TM) 11, OpenJDK(TM) 11 to run Neo4j Server.
* Please see https://neo4j.com/docs/ for Neo4j installation instructions.
Invoke-Neo4jAdmin : This instance of Java is not supported
所在位置 C:\Users\cheng fu\AppData\Local\Neo4j\Relate\Data\dbmss\dbms-02c10639-b64e-4ad2-aba3-3c416e7fcecb\bin
\neo4j-admin.ps1:13 字符: 7
+ Exit (Invoke-Neo4jAdmin -Verbose:$Arguments.Verbose -CommandArgs $Arg ...
+       ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
    + CategoryInfo          : NotSpecified: (:) [Write-Error], WriteErrorException
    + FullyQualifiedErrorId : Microsoft.PowerShell.Commands.WriteErrorException,Invoke-Neo4jAdmin
```

最初使用的是neo4j 4.1，jdk8，提示版本不兼容，将neo4j版本降级为3.5后解决