INFSCI 2725 DATA ANALYTICS

ASSIGNEMENT 2

MongoDB

4TH FEBRUARY 2018

Team Members:

Prashant Wankhede (PRW19@pitt.edu)

Abhishek Shankarnarayanan (ABS100@pitt.edu)

ETL Section

Python modules used:

- Codecs
- Pymongo

Strategy:

Open files (movies.dat, ratings.dat, tags.dat) in a read mode using codecs module as file handler. Using lines.strip(), so that then all lines won't be held at the same time. Then splitting each line on :: separated values according to column names. Finally, inserting each line using db.collectionname.insert_one() command to successfully record into the mongo database “db” at location C:\data\db where the server is running at localhost:27017.

Questions

Answers for all questions are as follows including additional personal queries from Question 5 to 7.

1)What genre is the movie CopyCat in?

Answer:

Crime|Drama|Horror|Mystery|Thriller

2)what genre has the most movies?

Answer:

Genre | Count

Action | 1473

Adventure | 1025

Animation | 286

Children's | 0

Comedy | 3703

Crime | 1118

Documentary | 482

Drama | 5339 ← Maximum value

Fantasy | 543

Film-Noir | 148

Horror | 1013

Musical | 436

Mystery | 509

Romance | 1685

Sci-Fi | 754

Thriller | 1706

War | 511

Western | 275

3)what tags did user 146 use to describe the movie "2001: A Space Odyssey”

Answer:

set(Arthur C. Clarke,artificial intelligence,based on a book)

4)What are the top 5 movies with the highest avg rating?

Answer:

Movie ID: 64275 | Avg Rating: 5.0

Movie ID: 42783 | Avg Rating: 5.0

Movie ID: 33264 | Avg Rating: 5.0

Movie ID: 53355 | Avg Rating: 5.0

Movie ID: 51209 | Avg Rating: 5.0

5) How many the number of movies a particular user has rated?

Answer:

userid:14717

44

6) List the users similar to a particular user in terms of rating?

Answer:

userid: 1

139

149

182

215

217

281

326

351

357

426

456

459

494

517

524

556

588

589

590

601

621

634

672

701

719

745

757

775

780

785

…..

7) Which tag appears most?

Answer:

sci-fi