

Assignment 3

CMSC 691 — Introduction to Data Science

Faisal Rasheed Khan

VB02734

vb02734@umbc.edu

1., 2., 3., 4., 5. Read the movies csv and the initial 5 records in the csv file:

```
{'_id': ObjectId('65665df671272cca11b6bbda'), 'movieId': 1, 'title': 'Toy Story (1995)', 'genres': 'Adventure|Animation|Children|Comedy|Fantasy'}
{'_id': ObjectId('65665df671272cca11b6bbdb'), 'movieId': 2, 'title': 'Jumanji (1995)', 'genres': 'Adventure|Children|Fantasy'}
{'_id': ObjectId('65665df671272cca11b6bbdc'), 'movieId': 3, 'title': 'Grumpier Old Men (1995)', 'genres': 'Comedy|Romance'}
{'_id': ObjectId('65665df671272cca11b6bbdd'), 'movieId': 4, 'title': 'Waiting to Exhale (1995)', 'genres': 'Comedy|Drama|Romance'}
{'_id': ObjectId('65665df671272cca11b6bbde'), 'movieId': 5, 'title': 'Father of the Bride Part II (1995)', 'genres': 'Comedy'}
```

Records inserted into the movies, ratings, tags and the movieId, userId are unique:

Movies:

```
{'_id': ObjectId('65666aec71272cca11b87a2f'), 'movieId': 200000, 'title': 'Pathan (2023)', 'genres': 'Action|Sci-Fi|Thriller'}
{'_id': ObjectId('65666aec71272cca11b87a30'), 'movieId': 200001, 'title': 'Tiger 3 (2023)', 'genres': 'Action|Sci-Fi|Thriller'}
{'_id': ObjectId('65666aec71272cca11b87a31'), 'movieId': 200002, 'title': 'Tu Jhooti Main Makkar (2023)', 'genres': 'Comedy|Romance'}
{'_id': ObjectId('65666aec71272cca11b87a32'), 'movieId': 200003, 'title': 'Raaz 3 (2012)', 'genres': 'Horror|Thriller'}
{'_id': ObjectId('65666aec71272cca11b87a33'), 'movieId': 200004, 'title': 'Dunki (2023)', 'genres': 'Adventure|Drama'}
```

Ratings:

```
{'_id': ObjectId('656672a771272cca11b87a34'), 'userId': 210000, 'movieId': 200000, 'rating': 5}
{'_id': ObjectId('656672a771272cca11b87a35'), 'userId': 210001, 'movieId': 200001, 'rating': 4}
{'_id': ObjectId('656672a771272cca11b87a36'), 'userId': 210002, 'movieId': 200002, 'rating': 2}
{'_id': ObjectId('656672a771272cca11b87a37'), 'userId': 210003, 'movieId': 200003, 'rating': 3}
{'_id': ObjectId('656672a771272cca11b87a38'), 'userId': 210004, 'movieId': 200004, 'rating': 5}
```

Tags:

```
{'_id': ObjectId('656673f671272cca11b87a39'), 'userId': 210000, 'movieId': 200000, 'tag': 'Agent'}
{'_id': ObjectId('656673f671272cca11b87a3a'), 'userId': 210000, 'movieId': 200000, 'tag': 'RAW'}
{'_id': ObjectId('656673f671272cca11b87a3b'), 'userId': 210001, 'movieId': 200001, 'tag': 'RAW'}
{'_id': ObjectId('656673f671272cca11b87a3c'), 'userId': 210002, 'movieId': 200002, 'tag': 'Hero'}
{'_id': ObjectId('656673f671272cca11b87a3d'), 'userId': 210003, 'movieId': 200003, 'tag': 'Emraan'}
{'_id': ObjectId('656673f671272cca11b87a3e'), 'userId': 210004, 'movieId': 200004, 'tag': 'SRK'}
```

6.a.

number of movies released per year:

```
{'_id': '1944', 'movies': 16}
{'_id': '1776', 'movies': 1}
{'_id': '1951', 'movies': 22}
{'_id': '2009', 'movies': 281}
{'_id': '1917', 'movies': 1}
{'_id': '1984', 'movies': 99}
{'_id': '1989', 'movies': 142}
{'_id': '1973', 'movies': 59}
{'_id': '1920', 'movies': 2}
{'_id': '1955', 'movies': 36}
{'_id': '1986', 'movies': 139}
{'_id': '1921', 'movies': 1}
{'_id': '1997', 'movies': 259}
{'_id': '1942', 'movies': 23}
{'_id': '1935', 'movies': 13}
{'_id': '1970', 'movies': 33}
{'_id': '1959', 'movies': 37}
{'_id': '2007', 'movies': 283}
{'_id': '2046', 'movies': 1}
{'_id': '1964', 'movies': 43}
{'_id': '1993', 'movies': 198}
{'_id': '1958', 'movies': 31}
{'_id': '1138', 'movies': 1}
{'_id': '2017', 'movies': 144}
{'_id': '1949', 'movies': 25}
{'_id': '2048', 'movies': 1}
{'_id': '1990', 'movies': 147}
{'_id': '1967', 'movies': 42}
{'_id': '1900', 'movies': 2}
{'_id': '1968', 'movies': 41}
{'_id': '2049', 'movies': 1}
{'_id': '1976', 'movies': 42}
{'_id': '1952', 'movies': 16}
{'_id': '1961', 'movies': 34}
{'_id': '1928', 'movies': 4}
{'_id': '2004', 'movies': 277}
{'_id': '2002', 'movies': 310}
{'_id': '1981', 'movies': 92}
{'_id': '1934', 'movies': 11}
{'_id': '1943', 'movies': 10}
{'_id': '1963', 'movies': 39}
{'_id': '2003', 'movies': 278}
{'_id': None, 'movies': 12}
{'_id': '2012', 'movies': 236}
{'_id': '1977', 'movies': 63}
{'_id': '1969', 'movies': 36}
{'_id': '1939', 'movies': 23}
{'_id': '2016', 'movies': 218}
{'_id': '1972', 'movies': 39}
{'_id': '1979', 'movies': 68}
{'_id': '2010', 'movies': 248}
{'_id': '1933', 'movies': 12}
{'_id': '1995', 'movies': 259}
{'_id': '1950', 'movies': 21}
{'_id': '1992', 'movies': 165}
{'_id': '1978', 'movies': 59}
{'_id': '1965', 'movies': 47}
```

{'_id': '1929', 'movies': 4}
{'_id': '1600', 'movies': 1}
{'_id': '1930', 'movies': 5}
{'_id': '1941', 'movies': 19}
{'_id': '1931', 'movies': 14}
{'_id': '2008', 'movies': 269}
{'_id': '1971', 'movies': 46}
{'_id': '1974', 'movies': 44}
{'_id': '1991', 'movies': 147}
{'_id': '1999', 'movies': 261}
{'_id': '1987', 'movies': 152}
{'_id': '1946', 'movies': 23}
{'_id': '1922', 'movies': 1}
{'_id': '2013', 'movies': 238}
{'_id': '1985', 'movies': 127}
{'_id': '2001', 'movies': 295}
{'_id': '1902', 'movies': 1}
{'_id': '1903', 'movies': 1}
{'_id': '1960', 'movies': 37}
{'_id': '1954', 'movies': 23}
{'_id': '1926', 'movies': 5}
{'_id': '1932', 'movies': 9}
{'_id': '1000', 'movies': 1}
{'_id': '1994', 'movies': 237}
{'_id': '1982', 'movies': 87}
{'_id': '1980', 'movies': 89}
{'_id': '1962', 'movies': 40}
{'_id': '2018', 'movies': 42}
{'_id': '1927', 'movies': 7}
{'_id': '1956', 'movies': 30}
{'_id': '2015', 'movies': 274}
{'_id': '2011', 'movies': 254}
{'_id': '1919', 'movies': 1}
{'_id': '1945', 'movies': 17}
{'_id': '1923', 'movies': 4}
{'_id': '1966', 'movies': 42}
{'_id': '2014', 'movies': 278}
{'_id': '1937', 'movies': 16}
{'_id': '2023', 'movies': 4}
{'_id': '1908', 'movies': 1}
{'_id': '1975', 'movies': 42}
{'_id': '1983', 'movies': 83}
{'_id': '1998', 'movies': 256}
{'_id': '2006', 'movies': 296}
{'_id': '1988', 'movies': 164}
{'_id': '1947', 'movies': 20}
{'_id': '1940', 'movies': 25}
{'_id': '1957', 'movies': 33}
{'_id': '1938', 'movies': 15}
{'_id': '1924', 'movies': 5}
{'_id': '1948', 'movies': 20}
{'_id': '1936', 'movies': 18}
{'_id': '1953', 'movies': 30}
{'_id': '2000', 'movies': 290}
{'_id': '1915', 'movies': 1}
{'_id': '1408', 'movies': 1}
{'_id': '3000', 'movies': 3}
{'_id': '1925', 'movies': 4}
{'_id': '2005', 'movies': 272}

```
{'_id': '1996', 'movies': 275}
{'_id': '1492', 'movies': 1}
{'_id': '1916', 'movies': 4}
```

6.b.

number of movies per genre:

```
{'_id': {'genre': 'IMAX'}, 'movies_per_genre': 158}
{'_id': {'genre': 'Film-Noir'}, 'movies_per_genre': 87}
{'_id': {'genre': 'Documentary'}, 'movies_per_genre': 440}
{'_id': {'genre': 'War'}, 'movies_per_genre': 382}
{'_id': {'genre': 'Western'}, 'movies_per_genre': 167}
{'_id': {'genre': 'Action'}, 'movies_per_genre': 1830}
{'_id': {'genre': 'Thriller'}, 'movies_per_genre': 1897}
{'_id': {'genre': 'Adventure'}, 'movies_per_genre': 1264}
{'_id': {'genre': 'Romance'}, 'movies_per_genre': 1597}
{'_id': {'genre': 'Fantasy'}, 'movies_per_genre': 779}
{'_id': {'genre': 'Crime'}, 'movies_per_genre': 1199}
{'_id': {'genre': '(no genres listed)'}, 'movies_per_genre': 34}
{'_id': {'genre': 'Comedy'}, 'movies_per_genre': 3757}
{'_id': {'genre': 'Children'}, 'movies_per_genre': 664}
{'_id': {'genre': 'Sci-Fi'}, 'movies_per_genre': 982}
{'_id': {'genre': 'Animation'}, 'movies_per_genre': 611}
{'_id': {'genre': 'Drama'}, 'movies_per_genre': 4362}
{'_id': {'genre': 'Mystery'}, 'movies_per_genre': 573}
{'_id': {'genre': 'Horror'}, 'movies_per_genre': 979}
{'_id': {'genre': 'Musical'}, 'movies_per_genre': 334}
```

6.c.

number of movies per rating:

```
{'_id': {'rating': 0.5}, 'ratings': 1370}
{'_id': {'rating': 3.0}, 'ratings': 20048}
{'_id': {'rating': 3.5}, 'ratings': 13136}
{'_id': {'rating': 5.0}, 'ratings': 13213}
{'_id': {'rating': 4.0}, 'ratings': 26819}
{'_id': {'rating': 1.0}, 'ratings': 2811}
{'_id': {'rating': 1.5}, 'ratings': 1791}
{'_id': {'rating': 2.0}, 'ratings': 7552}
{'_id': {'rating': 2.5}, 'ratings': 5550}
{'_id': {'rating': 4.5}, 'ratings': 8551}
```

6.d.

number of movies tagged:

```
{'total tags': 1577}
```

6.e.

the most popular tag:

```
{'_id': 'In Netflix queue', 'count': 131}
```

7. a. IDS_Assign3.ipynb

7.b.

Summarization of all the data added:

First, I have read the csv files movies, ratings, tags using python and inserted into the mongo database. Next, I have added customized data of 5 rows for movies, ratings and tags maintaining the unique ids. Next, I have used aggregate pipeline for the tasks to find: number of movies released per year, number of movies per genre, number of movies per rating, number of movies per tagged, the most popular tag.

7.c.

Summarization of my learnings:

Successfully installed MongoDB and its setup with python, Installed MongoDB and used by creating custom Environment in Anaconda and adding path to the OS of my computer.

Done Connection to MongoDB with the help of Python's pymongo library.

Inserted data in MongoDB with the help of python's pymongo.

Read data from csv and inserted the csv data to MongoDB collections with Pandas library using python.

Good experience and learnings of working with MongoDB collections and also python interaction with the MongoDB.

Updated the MongoDB collections which already had rows of the data from csv with additional 5 rows of data into the MongoDB collections.

Learned to add and manipulate the data in the MongoDB collections.

Learnt to apply Aggregation pipeline for the data which is there in the MongoDB and done complex operations of grouping and counting to get the described result as per the requirements. Used many more like sum, sort.

Learned to manipulate data with MongoDB collections and perform analytical tasks.

Overall learnings, installed, connected, add, manipulate data, used pipelines and operations like count, sum, sort, group for the MongoDB.

References:

https://blackboard.umbc.edu/bbcswebdav/pid-6404242-dt-content-rid-72305900_1/xid-72305900_1

https://blackboard.umbc.edu/bbcswebdav/pid-6488710-dt-content-rid-73120009_1/xid-73120009_1

https://blackboard.umbc.edu/bbcswebdav/pid-6488711-dt-content-rid-73120010_1/xid-73120010_1

[MongoDB Documentation](#)

[Our Documentation | Python.org](#)