

hw_7

November 22, 2025

1 homework 7: mongodb

1.1 e-library: parts 1 - 3

1.1.1 Insert Data: part 1 & 2

```
[179]: import json
from bson.json_util import dumps
from pymongo import MongoClient

# initialize a client object
client = MongoClient("localhost", 27017)

# select db
db = client["library"] # or client.<name>

# select collection
ebooks = db["ebooks"]
ebooks.delete_many({})

# insert first documents
docs = [
    {
        "title": "The Elements of Statistical Learning: Data Mining, Inference, and Prediction",
        "primary_author": "Trevor Hastie",
        "secondary_author": ["Robert Tibshirani", "Jerome H. Friedman"],
        "published": 2001,
        "pages": 745,
        "publisher": "Springer",
        "topic": "statistics"
    },
    {
        "title": "Deep Learning",
        "primary_author": "Ian Goodfellow",
        "secondary_author": ["Yoshua Bengio", "Aaron Courville"],
        "published": 2016,
        "pages": 800,
    }
]
```

```
        "publisher": "MIT Press",
        "topic": "deep learning"
    }
]

ebooks.insert_many(docs)
print(client.list_database_names())
print(db.list_collection_names())

['admin', 'config', 'congress', 'library', 'local', 'mongoTest']
['users', 'ebooks', 'checkouts']
```

```
[180]: docs = [
    {
        "title": "Zorrie",
        "primary_author": "Laird Hunt",
        "published": 2021,
        "pages": 400,
        "publisher": "Quercus",
        "topic": "fiction"
    },
    {
        "title": "Cicada",
        "primary_author": "Phoebe Giannisi",
        "published": 2022,
        "pages": 128,
        "publisher": "New Directions",
        "translator": "Brian Sneeden",
        "topic": "poetry"
    },
]
ebooks.insert_many(docs)
```

```
[180]: InsertManyResult([ObjectId('69228574944d64b91aacb3eb'),
 ObjectId('69228574944d64b91aacb3ec')], acknowledged=True)
```

```
[181]: users = db["users"]
users.delete_many({})

docs = [
    {
        "id": 1001,
        "name": "Pranav Gundrala",
        "phone": "123-456-7899",
        "address": "69 Brown St, Providence, RI 02912",
        "university": "Brown University"
    },
]
```

```

{
    "id": 1003,
    "name": "John Doe",
    "phone": "246-810-1214",
    "address": "123 Main St, Providence, RI 02906",
    "university": "University of Rhode Island"
},
{
    "id": 1002,
    "name": "Jane Doe",
    "phone": "369-121-5182",
    "address": "456 Benefit St, Providence, RI 02906",
    "university": "Brown University"
}
]

users.insert_many(docs)
print(db.list_collection_names())

```

['users', 'ebooks', 'checkouts']

```
[182]: checkouts = db["checkouts"]
checkouts.delete_many({})

docs = [
    {
        "date": "2025-11-22",
        "book": "The Elements of Statistical Learning: Data Mining, Inference, and Prediction",
        "user": 1003
    },
    {
        "date": "2025-09-16",
        "book": "The Elements of Statistical Learning: Data Mining, Inference, and Prediction",
        "user": 1002
    },
    {
        "date": "2025-09-25",
        "book": "Deep Learning",
        "user": 1002
    },
    {
        "date": "2025-11-21",
        "book": "Cicada",
        "user": 1001
    },
]
```

```

    {
        "date": "2025-10-13",
        "book": "Zorrie",
        "user": 1001
    },
    {
        "date": "2025-10-15",
        "book": "Deep Learning",
        "user": 1001
    },
]
checkouts.insert_many(docs)
print(db.list_collection_names())

```

['users', 'ebooks', 'checkouts']

1.1.2 Queries: part 3

query 1

```
[183]: # 1. Which users have checked out 'Elements of Statistical Learning: Data Mining, Inference, and Prediction'?
query = {
    "book": "The Elements of Statistical Learning: Data Mining, Inference, and Prediction"
}
projection = {"_id" : 0, "user" : 1}

ids = checkouts.find(query, projection)
ids = [x['user'] for x in ids]

query = {
    "id": {"$in": ids}
}

result = users.find(query)
print("Users who have checked out 'Elements of Statistical Learning...'")
for x in result: print(dumps(x, indent=4))
```

Users who have checked out 'Elements of Statistical Learning...'

```
{
    "_id": {
        "$oid": "69228574944d64b91aacb3ee"
    },
    "id": 1003,
    "name": "John Doe",
    "phone": "246-810-1214",
    "address": "123 Main St, Providence, RI 02906",
```

```

        "university": "University of Rhode Island"
    }
{
    "_id": {
        "$oid": "69228574944d64b91aacb3ef"
    },
    "id": 1002,
    "name": "Jane Doe",
    "phone": "369-121-5182",
    "address": "456 Benefit St, Providence, RI 02906",
    "university": "Brown University"
}

```

query 2

```
[184]: # 2. Which users from Brown University have checked out books on Deep Learning?
query = {
    "topic": "deep learning"
}
projection = {"_id": 0, "title": 1}
titles = ebooks.find(query, projection)
titles = [x['title'] for x in titles]

query = {
    "book": {"$in": titles}
}
projection = {"_id": 0, "user": 1}
ids = checkouts.find(query, projection)
ids = [x['user'] for x in ids]

query = {
    "id": {"$in": ids},
    "university": "Brown University"
}
result = users.find(query)
print("Users from 'Brown University' who have checked out books on 'deep learning':")
for x in result: print(dumps(x, indent=4))
```

```
Users from 'Brown University' who have checked out books on 'deep learning':
{
    "_id": {
        "$oid": "69228574944d64b91aacb3ed"
    },
    "id": 1001,
    "name": "Pranav Gundrala",
    "phone": "123-456-7899",
    "address": "69 Brown St, Providence, RI 02912",
    "university": "Brown University"
```

```

}
{
    "_id": {
        "$oid": "69228574944d64b91aacb3ef"
    },
    "id": 1002,
    "name": "Jane Doe",
    "phone": "369-121-5182",
    "address": "456 Benefit St, Providence, RI 02906",
    "university": "Brown University"
}

```

[185]: # 2. Using an \$lookup function ...

```

query = [
    # join book info to checkouts
    {
        "$lookup": {
            "from": "ebooks",
            "localField": "book",
            "foreignField": "title",
            "as": "book_info"
        }
    },
    # filter for 'deep learning'
    { "$match": { "book_info.topic": 'deep learning' } },
    # join user info to checkouts
    {
        "$lookup": {
            "from": "users",
            "localField": "user",
            "foreignField": "id",
            "as": "user_info"
        }
    },
    # filter for 'Brown University'
    { "$match": { "user_info.university": "Brown University" } },
    {
        "$project": {
            "_id": 0,
            "user_info": 1
        }
    }
]

result = checkouts.aggregate(query)
for x in result: print(dumps(x, indent=4))

```

{

```

"user_info": [
    {
        "_id": {
            "$oid": "69228574944d64b91aacb3ef"
        },
        "id": 1002,
        "name": "Jane Doe",
        "phone": "369-121-5182",
        "address": "456 Benefit St, Providence, RI 02906",
        "university": "Brown University"
    }
]
}
{
    "user_info": [
        {
            "_id": {
                "$oid": "69228574944d64b91aacb3ed"
            },
            "id": 1001,
            "name": "Pranav Gundrala",
            "phone": "123-456-7899",
            "address": "69 Brown St, Providence, RI 02912",
            "university": "Brown University"
        }
    ]
}

```

query 3

```
[186]: # 3. How many times is the book 'Deep Learning' been checked out?
query = {
    "book": "Deep Learning"
}
result = checkouts.count_documents(query)
print(f"'Deep Learning' has been checked out {result} times.")
```

'Deep Learning' has been checked out 2 times.

1.2 library of congress dataset: part 4

1.2.1 Insert Data

```
[187]: # initialize a client object
client = MongoClient("localhost", 27017)
# select db
db = client["congress"]
# select collection
books = db["books"]
```

```
books.delete_many({})
```

[187]: DeleteResult({'n': 1000, 'ok': 1.0}, acknowledged=True)

```
[188]: # get documents and add
with open('./data/mongodb_sample.json') as f:
    docs = json.load(f)
books.insert_many(docs)
```

[188]: InsertManyResult([ObjectId('69228575944d64b91aacb3f7'),
 ObjectId('69228575944d64b91aacb3f8'), ObjectId('69228575944d64b91aacb3f9'),
 ObjectId('69228575944d64b91aacb3fa'), ObjectId('69228575944d64b91aacb3fb'),
 ObjectId('69228575944d64b91aacb3fc'), ObjectId('69228575944d64b91aacb3fd'),
 ObjectId('69228575944d64b91aacb3fe'), ObjectId('69228575944d64b91aacb3ff'),
 ObjectId('69228575944d64b91aacb400'), ObjectId('69228575944d64b91aacb401'),
 ObjectId('69228575944d64b91aacb402'), ObjectId('69228575944d64b91aacb403'),
 ObjectId('69228575944d64b91aacb404'), ObjectId('69228575944d64b91aacb405'),
 ObjectId('69228575944d64b91aacb406'), ObjectId('69228575944d64b91aacb407'),
 ObjectId('69228575944d64b91aacb408'), ObjectId('69228575944d64b91aacb409'),
 ObjectId('69228575944d64b91aacb40a'), ObjectId('69228575944d64b91aacb40b'),
 ObjectId('69228575944d64b91aacb40c'), ObjectId('69228575944d64b91aacb40d'),
 ObjectId('69228575944d64b91aacb40e'), ObjectId('69228575944d64b91aacb40f'),
 ObjectId('69228575944d64b91aacb410'), ObjectId('69228575944d64b91aacb411'),
 ObjectId('69228575944d64b91aacb412'), ObjectId('69228575944d64b91aacb413'),
 ObjectId('69228575944d64b91aacb414'), ObjectId('69228575944d64b91aacb415'),
 ObjectId('69228575944d64b91aacb416'), ObjectId('69228575944d64b91aacb417'),
 ObjectId('69228575944d64b91aacb418'), ObjectId('69228575944d64b91aacb419'),
 ObjectId('69228575944d64b91aacb41a'), ObjectId('69228575944d64b91aacb41b'),
 ObjectId('69228575944d64b91aacb41c'), ObjectId('69228575944d64b91aacb41d'),
 ObjectId('69228575944d64b91aacb41e'), ObjectId('69228575944d64b91aacb41f'),
 ObjectId('69228575944d64b91aacb420'), ObjectId('69228575944d64b91aacb421'),
 ObjectId('69228575944d64b91aacb422'), ObjectId('69228575944d64b91aacb423'),
 ObjectId('69228575944d64b91aacb424'), ObjectId('69228575944d64b91aacb425'),
 ObjectId('69228575944d64b91aacb426'), ObjectId('69228575944d64b91aacb427'),
 ObjectId('69228575944d64b91aacb428'), ObjectId('69228575944d64b91aacb429'),
 ObjectId('69228575944d64b91aacb42a'), ObjectId('69228575944d64b91aacb42b'),
 ObjectId('69228575944d64b91aacb42c'), ObjectId('69228575944d64b91aacb42d'),
 ObjectId('69228575944d64b91aacb42e'), ObjectId('69228575944d64b91aacb42f'),
 ObjectId('69228575944d64b91aacb430'), ObjectId('69228575944d64b91aacb431'),
 ObjectId('69228575944d64b91aacb432'), ObjectId('69228575944d64b91aacb433'),
 ObjectId('69228575944d64b91aacb434'), ObjectId('69228575944d64b91aacb435'),
 ObjectId('69228575944d64b91aacb436'), ObjectId('69228575944d64b91aacb437'),
 ObjectId('69228575944d64b91aacb438'), ObjectId('69228575944d64b91aacb439'),
 ObjectId('69228575944d64b91aacb43a'), ObjectId('69228575944d64b91aacb43b'),
 ObjectId('69228575944d64b91aacb43c'), ObjectId('69228575944d64b91aacb43d'),
 ObjectId('69228575944d64b91aacb43e'), ObjectId('69228575944d64b91aacb43f'),
 ObjectId('69228575944d64b91aacb440'), ObjectId('69228575944d64b91aacb441'),


```

ObjectId('69228575944d64b91aacb790'), ObjectId('69228575944d64b91aacb791'),
ObjectId('69228575944d64b91aacb792'), ObjectId('69228575944d64b91aacb793'),
ObjectId('69228575944d64b91aacb794'), ObjectId('69228575944d64b91aacb795'),
ObjectId('69228575944d64b91aacb796'), ObjectId('69228575944d64b91aacb797'),
ObjectId('69228575944d64b91aacb798'), ObjectId('69228575944d64b91aacb799'),
ObjectId('69228575944d64b91aacb79a'), ObjectId('69228575944d64b91aacb79b'),
ObjectId('69228575944d64b91aacb79c'), ObjectId('69228575944d64b91aacb79d'),
ObjectId('69228575944d64b91aacb79e'), ObjectId('69228575944d64b91aacb79f'),
ObjectId('69228575944d64b91aacb7a0'), ObjectId('69228575944d64b91aacb7a1'),
ObjectId('69228575944d64b91aacb7a2'), ObjectId('69228575944d64b91aacb7a3'),
ObjectId('69228575944d64b91aacb7a4'), ObjectId('69228575944d64b91aacb7a5'),
ObjectId('69228575944d64b91aacb7a6'), ObjectId('69228575944d64b91aacb7a7'),
ObjectId('69228575944d64b91aacb7a8'), ObjectId('69228575944d64b91aacb7a9'),
ObjectId('69228575944d64b91aacb7aa'), ObjectId('69228575944d64b91aacb7ab'),
ObjectId('69228575944d64b91aacb7ac'), ObjectId('69228575944d64b91aacb7ad'),
ObjectId('69228575944d64b91aacb7ae'), ObjectId('69228575944d64b91aacb7af'),
ObjectId('69228575944d64b91aacb7b0'), ObjectId('69228575944d64b91aacb7b1'),
ObjectId('69228575944d64b91aacb7b2'), ObjectId('69228575944d64b91aacb7b3'),
ObjectId('69228575944d64b91aacb7b4'), ObjectId('69228575944d64b91aacb7b5'),
ObjectId('69228575944d64b91aacb7b6'), ObjectId('69228575944d64b91aacb7b7'),
ObjectId('69228575944d64b91aacb7b8'), ObjectId('69228575944d64b91aacb7b9'),
ObjectId('69228575944d64b91aacb7ba'), ObjectId('69228575944d64b91aacb7bb'),
ObjectId('69228575944d64b91aacb7bc'), ObjectId('69228575944d64b91aacb7bd'),
ObjectId('69228575944d64b91aacb7be'), ObjectId('69228575944d64b91aacb7bf'),
ObjectId('69228575944d64b91aacb7c0'), ObjectId('69228575944d64b91aacb7c1'),
ObjectId('69228575944d64b91aacb7c2'), ObjectId('69228575944d64b91aacb7c3'),
ObjectId('69228575944d64b91aacb7c4'), ObjectId('69228575944d64b91aacb7c5'),
ObjectId('69228575944d64b91aacb7c6'), ObjectId('69228575944d64b91aacb7c7'),
ObjectId('69228575944d64b91aacb7c8'), ObjectId('69228575944d64b91aacb7c9'),
ObjectId('69228575944d64b91aacb7ca'), ObjectId('69228575944d64b91aacb7cb'),
ObjectId('69228575944d64b91aacb7cc'), ObjectId('69228575944d64b91aacb7cd'),
ObjectId('69228575944d64b91aacb7ce'), ObjectId('69228575944d64b91aacb7cf'),
ObjectId('69228575944d64b91aacb7d0'), ObjectId('69228575944d64b91aacb7d1'),
ObjectId('69228575944d64b91aacb7d2'), ObjectId('69228575944d64b91aacb7d3'),
ObjectId('69228575944d64b91aacb7d4'), ObjectId('69228575944d64b91aacb7d5'),
ObjectId('69228575944d64b91aacb7d6'), ObjectId('69228575944d64b91aacb7d7'),
ObjectId('69228575944d64b91aacb7d8'), ObjectId('69228575944d64b91aacb7d9'),
ObjectId('69228575944d64b91aacb7da'), ObjectId('69228575944d64b91aacb7db'),
ObjectId('69228575944d64b91aacb7dc'), ObjectId('69228575944d64b91aacb7dd'),
ObjectId('69228575944d64b91aacb7de')], acknowledged=True)

```

1.2.2 Queries

query 1

```
[189]: # debug #
# look at first book
# head = books.find_one({})
```

```
# print(dumps(head, indent=4))
```

```
[200]: # debug #
# import numpy as np
# test = books.find({}, {"_id": 0, "date": 1})
# test = np.array([int(x['date']) for x in test])
# print(test[test < 1800])
# expected answer is one book from 1780
```

```
[191]: # 1. What books available at the LoC were written before 1800?
```

```
query = {
    "date": {"$lt": "1800"}
}
projection = {"_id": 0, "item": {"title": 1}, "date": 1}
result = books.find(query, projection)
for x in result: print(dumps(x, indent=4))

{
    "date": "1780",
    "item": {
        "title": "Historie der waereld,"
    }
}
```

query 2

```
[192]: # 2. How many books are written in English?
```

```
query = {
    "language": "english"
}
result = books.count_documents(query)
print(f"{result} books are written in english")
```

918 books are written in english

query 3

```
[193]: # debug #
# for x in (books.find({}, {"_id":0, "item.contributors":1})): ↴
#     print(dumps(x, indent=4))
# there are books with "item" = {} and "contributors" missing
```

```
[194]: # 3. What books have more than 1 contributor?
```

```
query = [
    # check that the item and contributors fields exist
    {
        "$match": {
            "item": {"$exists": True},
            "item.contributors": {"$exists": True, "$ne": []}
        }
}
```

```

},
# calculate the length
{
    "$addFields": {
        "len": {"$size": "$item.contributors" }
    }
},
# check if the len > 1
{
    "$match": {"len": {"$gt": 1}}
},
{
    "$project": {"_id":0, "title":1, "item.contributors":1}
}
]

# print first 3 results
result = list(books.aggregate(query))
for x in result[:3]: print(dumps(x, indent=4))

{
    "item": {
        "contributors": [
            "Hyslop, James H. (James Hervey), 1854-",
            "Woodrow Wilson Collection (Library of Congress)"
        ]
    },
    "title": "Democracy; a study of government,"
}
{
    "item": {
        "contributors": [
            "Strong, Josiah, 1847-1916. [from old catalog]",
            "Congregational home missionary society. [from old catalog]"
        ]
    },
    "title": "Our country: its possible future and its present crisis."
}
{
    "item": {
        "contributors": [
            "Butler, James Davie, 1815-1905.",
            "Joseph Meredith Toner Collection (Library of Congress)"
        ]
    },
    "title": "Portraits of Columbus : A monograph"
}

```

query 4

```
[197]: # debug #
# books do exist with multiple languages
# query = [
#     {"$match": {"language": {"$exists": True, "$ne": []}}},
#     {"$addFields": {"len": {"$size": "$language"}}},
#     {"$match": {"len": {"$gt": 1}}},
#     {"$project": {"_id": 0, "language": 1, "title": 1}}
# ]
# result = list(books.aggregate(query))
# for x in result[:1]: print(dumps(x, indent=4))
```

[196]: # 4. How many books per language does the data have? (Hint: use a pipeline)

```
pipeline = [
    # unwind for books with multiple languages
    { "$unwind": "$language" },
    # group by language and count docs
    { "$group": {"_id": "$language",
                 "count": {"$count": {}}}
    },
    # sort descending
    { "$sort": {"count": -1}}
]
# print
for x in books.aggregate(pipeline): print(dumps(x, indent=4))
```

```
{
    "_id": "english",
    "count": 918
}
{
    "_id": "german",
    "count": 17
}
{
    "_id": "french",
    "count": 14
}
{
    "_id": "spanish",
    "count": 13
}
{
    "_id": "russian",
    "count": 9
}
{
    "_id": "latin",
```

```

        "count": 7
    }
    {
        "_id": "italian",
        "count": 6
    }
    {
        "_id": "flemish",
        "count": 4
    }
    {
        "_id": "dutch",
        "count": 4
    }
    {
        "_id": "czech",
        "count": 2
    }
    {
        "_id": "portuguese",
        "count": 2
    }
    {
        "_id": "polish",
        "count": 2
    }
    {
        "_id": "danish",
        "count": 2
    }
    {
        "_id": "englat",
        "count": 2
    }
    {
        "_id": "engspa",
        "count": 1
    }
    {
        "_id": "latgrc",
        "count": 1
    }
    {
        "_id": "latgre",
        "count": 1
    }
    {
        "_id": "multiple languages",

```

```
        "count": 1
    }
    {
        "_id": "swedish",
        "count": 1
    }
    {
        "_id": "engund",
        "count": 1
    }
    {
        "_id": "lithuanian",
        "count": 1
    }
    {
        "_id": "sanskrit",
        "count": 1
    }
    {
        "_id": "engger",
        "count": 1
    }
    {
        "_id": "lateng",
        "count": 1
    }
}
```