

Student name: Pothumulla Kankanamge Mewan Madhusa

MongoDB NoSQL Assignment

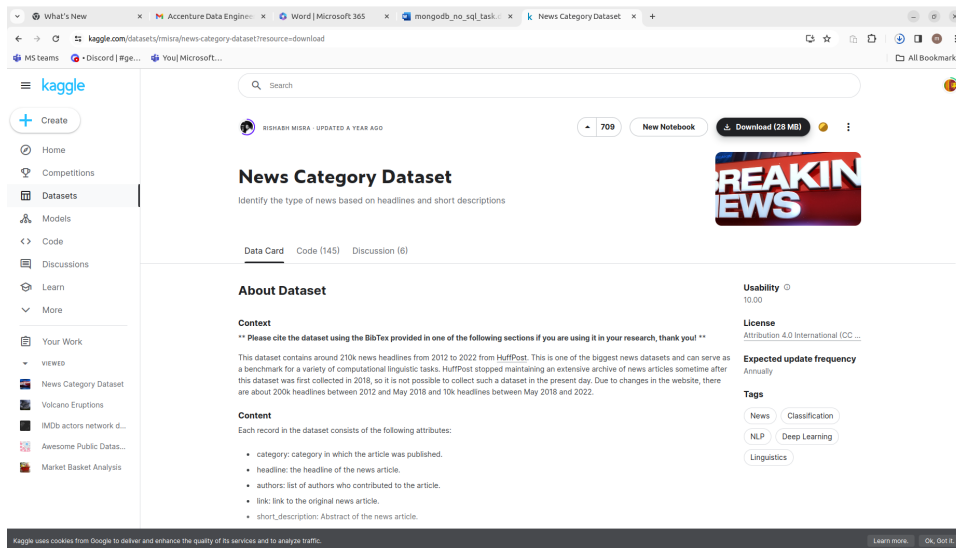
Task 1.

Download dataset from Kaggle*:

News Category Dataset

<https://www.kaggle.com/datasets/rmisra/news-category-dataset>

** for that you will need to create free account on kaggle.com. Go ahead.*

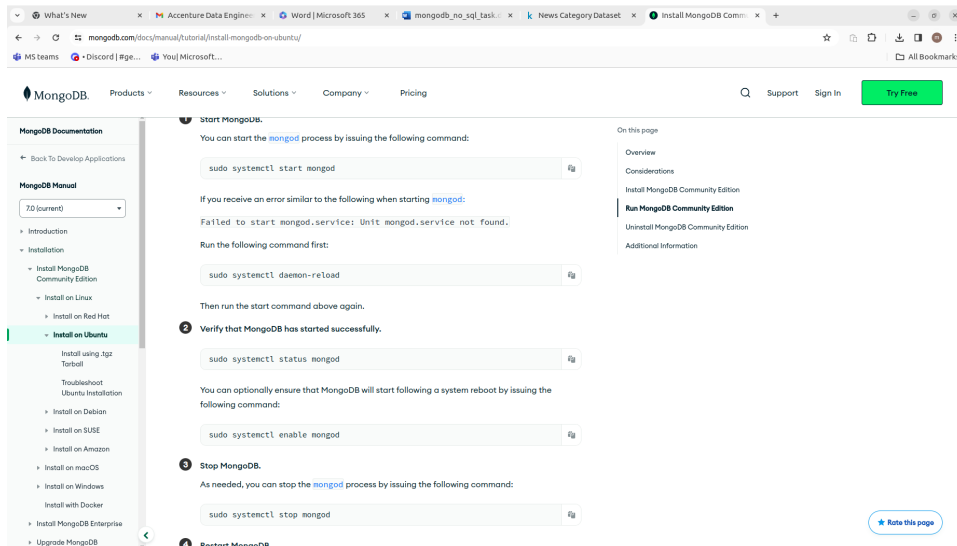


Create new database "news_db" in MongoDB on your local machine.

Provide command of newly created DB.

Follow the below steps to install the MongoDB community server on ubuntu

<https://www.mongodb.com/docs/manual/tutorial/install-mongodb-on-ubuntu/>



Installing DB

```
bootcamp@aldis-HP-EliteBook-840-G5: ~  
  
● mongod.service - MongoDB Database Server  
   Loaded: loaded (/lib/systemd/system/mongod.service; enabled; preset: enabled)  
   Active: active (running) since Tue 2024-01-16 12:01:51 EET; 13min ago  
     Docs: https://docs.mongodb.org/manual  
    Main PID: 1662 (mongod)  
      Memory: 227.4M  
        CPU: 7.587s  
    CGroup: /system.slice/mongod.service  
            └─1662 /usr/bin/mongod --config /etc/mongod.conf  
  
jan 16 12:01:51 aldis-HP-EliteBook-840-G5 systemd[1]: Started MongoDB Database Serv  
er.
```

Create new database mongosh command - use news_db

```
mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
Lines 1-17/17 (END)
[1]: Stopped
      sudo systemctl status mongod
bootcamp@idls-HP-EliteBook-840-G5:~$
bootcamp@idls-HP-EliteBook-840-G5:~$
bootcamp@idls-HP-EliteBook-840-G5:~$ sudo systemctl status mongod
● mongod.service - MongoDB Database Server
   Loaded: loaded (/lib/systemd/system/mongod.service; enabled; preset: enabled)
   Active: active (running) since Tue 2024-01-16 12:01:51 EET; 17min ago
     Docs: https://docs.mongodb.org/manual
    Main PID: 1662 (mongod)
      Memory: 227.4M
         CPU: 9.188s
    CGroup: /system.slice/mongod.service
           └─ssd /usr/bin/mongod ->config /etc/mongod.conf

Jan 16 12:01:51 aIdls-HP-EliteBook-840-G5 systemd[1]: Started MongoDB Database Server.
Jan 16 12:01:51 aIdls-HP-EliteBook-840-G5 mongod[1662]: {"t":{"$date":"2024-01-16T18:01:51.265Z"},"s":"I",  "c":"CONTROL",  "id":7484580, "ctx":"","msg":"Environment variable MONGODB_CONFIG_OVERRIDE_NO"}
Lines 1-17/17 (END)
[2]: Stopped
      sudo systemctl status mongod
bootcamp@idls-HP-EliteBook-840-G5:~$
bootcamp@idls-HP-EliteBook-840-G5:~$
bootcamp@idls-HP-EliteBook-840-G5:~$ mongosh
Current Mongosh Log ID: 65ae583424ae17f1cee0ab
Connecting to:
  mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&appName=mongosh2.1.1
Using MongoDB:
  5.0.23
Using Mongosh:
  2.1.1

For mongosh info see: https://docs.mongodb.com/mongosh-shell/

To help improve our products, anonymous usage data is collected and sent to MongoDB periodically (https://www.mongodb.com/legal/privacy-policy).
You can opt-out by running the disableTelemetry() command.

-----
The server generated these startup warnings when booting
2024-01-16T12:01:51.274Z:80: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine. See http://dochub.mongodb.org/core/prodnotes-filesystem
2024-01-16T12:01:52.586+02:00: Access control is not enabled for the database. Read and write access to data and configuration is unrestricted
-----

test>
test> show dbs;
admin    40.00 KiB
config   60.00 KiB
local    80.00 KiB
test> use news_db
switched to db news_db
news_db>

news_db> show dbs;
admin    40.00 KiB
config   60.00 KiB
local    80.00 KiB
news_db>
```

Task 2.

Load data from news dataset to news_db.

Explain how you solved the task.

First create collection inside the news_db - `db.createCollection("newscategory");`

```
mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
Active: active (running) since Tue 2024-01-16 12:01:51 EET; 17min ago
 Docs: https://docs.mongodb.org/manual
  Main PID: 1662 (mongod)
    Memory: 227.4M
         CPU: 9.188s
    CGroup: /system.slice/mongod.service
           └─ssd /usr/bin/mongod ->config /etc/mongod.conf

Jan 16 12:01:51 aIdls-HP-EliteBook-840-G5 systemd[1]: Started MongoDB Database Server.
Jan 16 12:01:51 aIdls-HP-EliteBook-840-G5 mongod[1662]: {"t":{"$date":"2024-01-16T18:01:51.265Z"},"s":"I",  "c":"CONTROL",  "id":7484580, "ctx":"","msg":"Environment variable MONGODB_CONFIG_OVERRIDE_NO"}
Lines 1-17/17 (END)
[2]: Stopped
      sudo systemctl status mongod
bootcamp@idls-HP-EliteBook-840-G5:~$
bootcamp@idls-HP-EliteBook-840-G5:~$
bootcamp@idls-HP-EliteBook-840-G5:~$ mongosh
Current Mongosh Log ID: 65ae583424ae17f1cee0ab
Connecting to:
  mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&appName=mongosh2.1.1
Using MongoDB:
  5.0.23
Using Mongosh:
  2.1.1

For mongosh info see: https://docs.mongodb.com/mongosh-shell/

To help improve our products, anonymous usage data is collected and sent to MongoDB periodically (https://www.mongodb.com/legal/privacy-policy).
You can opt-out by running the disableTelemetry() command.

-----
The server generated these startup warnings when booting
2024-01-16T12:01:51.274Z:80: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine. See http://dochub.mongodb.org/core/prodnotes-filesystem
2024-01-16T12:01:52.586+02:00: Access control is not enabled for the database. Read and write access to data and configuration is unrestricted
-----

test>
test> show dbs;
admin    40.00 KiB
config   60.00 KiB
local    80.00 KiB
test> use news_db
switched to db news_db
news_db>

news_db> show dbs;
admin    40.00 KiB
config   60.00 KiB
local    80.00 KiB
news_db>

news_db> db.createCollection(newscategory);
{ ok: 1 }
news_db> db.createCollection("newscategory");
{ ok: 1 }
news_db>
```

Use mongoimport functionality to load data into the collection

```
mongoimport --db news_db --collection newscategory Downloads/News_Category_Dataset_v3.json
```

First give the DB name then provide collection name and then give the dataset location in the local system, I have provided downloads file location since my file was in there.

```
bootcamp@aldis-HP-ElliteBook-840-G5: -
bootcamp@aldis-HP-ElliteBook-840-G5: $
bootcamp@aldis-HP-ElliteBook-840-G5: $
bootcamp@aldis-HP-ElliteBook-840-G5: $ ls
bootcamp@aldis-HP-ElliteBook-840-G5: $ mongoimport --db news_db --collection newscategory Downloads/news_Category_Dataset_v3.json
2024-01-10T12:35:08.518+0200   connected to mongodb://localhost/
2024-01-10T12:35:08.519+0200   [#####] news_db.newscategory 42.4MB/83.2MB (51.0%)
2024-01-10T12:35:11.277+0200   [#####] news_db.newscategory 83.2MB/83.2MB (100.0%)
2024-01-10T12:35:11.277+0200   209527 document(s) imported successfully, 0 document(s) failed to import.
bootcamp@aldis-HP-ElliteBook-840-G5: $
```

Let's check data set got imported

```

mongosh mongodb://127.0.0.1:27017/directConnection=true&serverSelectionTimeoutMS=2000
bootcamp@ldis-MP-ElitteBook-R40-G51:~$
bootcamp@ldis-MP-ElitteBook-R40-G51:~$
bootcamp@ldis-MP-ElitteBook-R40-G51:~$ ls
bootcamp@ldis-MP-ElitteBook-R40-G51:~$ mongo --useNewUrlParser --collection newscategory Downloads/News_Category_Dataset_v3.json
2024-01-16T12:35:08.518+0200   connected to: mongodb://localhost/
2024-01-16T12:35:08.519+0200   [#####] news_db.newscategory 42.4MB/83.2MB (51.0K)
2024-01-16T12:35:11.277+0200   [#####] news_db.newscategory 83.2MB/83.2MB (100.0K)
2024-01-16T12:35:11.277+0200   209577 document(s) imported successfully. 0 document(s) failed to import.
bootcamp@ldis-MP-ElitteBook-R40-G51:~$
bootcamp@ldis-MP-ElitteBook-R40-G51:~$
bootcamp@ldis-MP-ElitteBook-R40-G51:~$ mongosh
Current Mongosh Log ID: 65ae5d07ad0554aeef5343
Connecting to:  mongosh://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&appName=mongosh+2.1.1
Using MongoDB:      5.0.23
Using Mongosh:      2.1.1

For mongosh info see: https://docs.mongodb.com/mongodb-shell/

-----
The server generated these startup warnings when booting
2024-01-16T12:28:15.1375+02:00: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine. See http://dochub.mongodb.org/core/prodatodes-filesystem
2024-01-16T12:01:52.586+02:00: Access control is not enabled for the database. Read and write access to data and configuration is unrestricted

test>

test> use news_db
switched to db news_db
news_db.newscategory.find();
{
  _id: ObjectId('65ae5bd07ad0554aeef530a7'),
  link: 'https://www.huffpost.com/entry/covid-booster-punished-flight-attendant-punch-justice-department_u_632e2d3de4be0f47896329fe',
  headline: 'American Airlines Flyer Charged, Banned for Life After Punching Flight Attendant on Video',
  category: 'U.S. NEWS',
  short_description: "he was subdued by passengers and crew when he fled to the back of the aircraft after the confrontation, according to the U.S. attorney's office in Los Angeles.",
  authors: 'Mary Ragenfuss',
  date: '2022-09-23'
},
{
  _id: ObjectId('65ae5bd07ad0554aeef530a9'),
  link: 'https://www.huffpost.com/entry/covid-boosters-upstate-us_u_632d719ee4b0f7faefeaac9',
  headline: 'Over 4 million Americans Roll up Sleeves for Omicron-Targeted COVID Boosters',
  category: 'U.S. NEWS',
  short_description: "Health experts said it is too early to predict whether demand would match up with the 171 million doses of the new boosters the U.S. ordered for the fall.",
  authors: 'Celia A. Johnson, AP',
  date: '2022-09-23'
},
{
  _id: ObjectId('65ae5bd07ad0554aeef530ab'),
  link: 'https://www.huffpost.com/entry/stolen-us-russian-army-offrnt-to-bodys-charter-u_632ade4be0f4dfbf5ff',
  headline: 'Stolen as He To Call Russian War An Offrnt To Body's Charter',
  category: 'WORLD NEWS',
  short_description: "While Mr. Putin officials say the crow of the president's visit to the U.N. this year will be a full-throated condemnation of Russia and its brutal war,"
  authors: 'Ramar Nadhani, AP',

```

Task 3.

Describe the dataset loaded to news_db

The dataset loaded to `news_db` appears to be in JSON format and contains information about a news article. Here's a breakdown of the key-value pairs:

category: Category article belongs to

headline: Headline of the article

authors: Person authored the article

link: Link to the post

short_description: Short description of the article

date: Date the article was published

The data structure uses key-value pairs, which is common in JSON formats. This structure is database-friendly as it allows for easy retrieval and querying of specific information such as the headline, author, or publication date. The use of strings and basic data types makes it suitable for storage and retrieval in various database systems.

Print news_db schema.

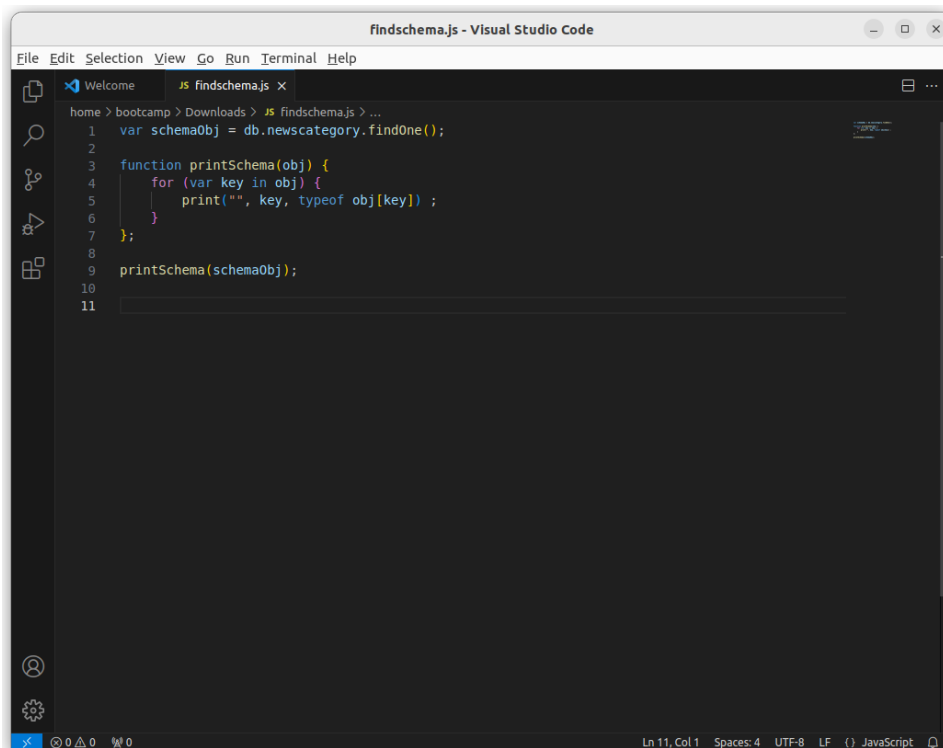
```
var schemaObj = db.newscategory.findOne();
```

```
function printSchema(obj) {
```

```
  for (var key in obj) {
```

```
    print("", key, typeof obj[key]) ;}}
```

```
printSchema(schemaObj);
```



The screenshot shows a Visual Studio Code window titled 'findschema.js - Visual Studio Code'. The editor displays the following JavaScript code:

```
1 var schemaObj = db.newscategory.findOne();
2
3 function printSchema(obj) {
4   for (var key in obj) {
5     print("", key, typeof obj[key]) ;
6   }
7 }
8
9 printSchema(schemaObj);
10
11
```

The status bar at the bottom indicates 'Ln 11, Col 1', 'Spaces: 4', 'UTF-8', 'LF', and 'JavaScript'.

```
load("Downloads/findschema.js");
```

Add new collections link_db, headline_db, category_db, short_description_db, suthors_db, date_db, link db.

Fill in each collection from news_db corresponding value.

Explain how you solved the task.

Code:

```
var originalCollection = db.newscategory;
```

```
// Create and fill 'link_db' collection
```

```
var linkCollection = db.link_db;
```

```
linkCollection.insertMany(originalCollection.find({}, { _id: 0, link: 1 }).toArray());
```

```
// Create and fill 'headline_db' collection
```

```
var headlineCollection = db.headline_db;
```

```
headlineCollection.insertMany(originalCollection.find({}, { _id: 0, headline: 1 }).toArray());
```

```
// Create and fill 'category_db' collection
```

```
var categoryCollection = db.category_db;
```

```
categoryCollection.insertMany(originalCollection.find({}, { _id: 0, category: 1 }).toArray());
```

```
// Create and fill 'short_description_db' collection
```

```
var shortDescriptionCollection = db.short_description_db;
```

```
shortDescriptionCollection.insertMany(originalCollection.find({}, { _id: 0, short_description: 1 }).toArray());
```

```
// Create and fill 'authors_db' collection
```

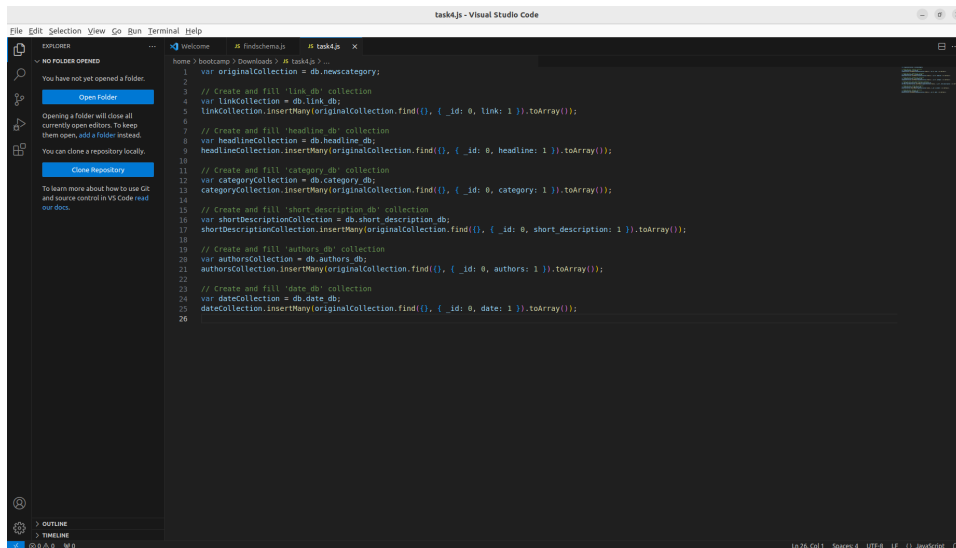
```
var authorsCollection = db.authors_db;
```

```
authorsCollection.insertMany(originalCollection.find({}, { _id: 0, authors: 1 }).toArray());
```

```
// Create and fill 'date_db' collection
```

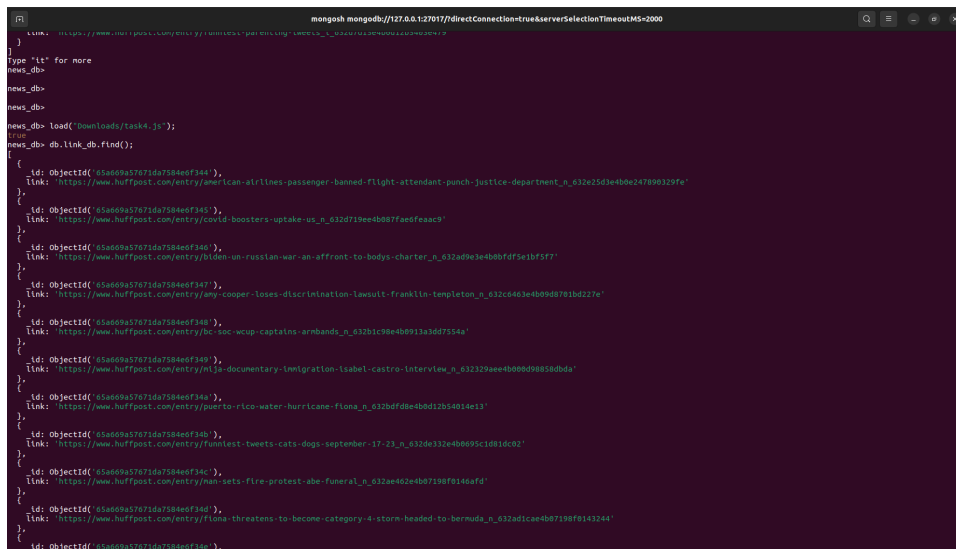
```
var dateCollection = db.date_db;
```

```
dateCollection.insertMany(originalCollection.find({}, { _id: 0, date: 1 }).toArray());
```



```
task4.js - Visual Studio Code
File Edit Selection View Go Run Terminal Help
home > bootcamp > Downloads > # task4.js
1 var originalCollection = db.news.category;
2
3 // Create and fill 'link' db collection
4 var linkCollection = db.link_db;
5 linkCollection.insertMany(originalCollection.find({}, { _id: 0, link: 1 }).toArray());
6
7 // Create and fill 'headline' db collection
8 var headlineCollection = db.headline_db;
9 headlineCollection.insertMany(originalCollection.find({}, { _id: 0, headline: 1 }).toArray());
10
11 // Create and fill 'category' db collection
12 var categoryCollection = db.category_db;
13 categoryCollection.insertMany(originalCollection.find({}, { _id: 0, category: 1 }).toArray());
14
15 // Create and fill 'short_description' db collection
16 var shortDescriptionCollection = db.short_description_db;
17 shortDescriptionCollection.insertMany(originalCollection.find({}, { _id: 0, short_description: 1 }).toArray());
18
19 // Create and fill 'authors' db collection
20 var authorsCollection = db.authors_db;
21 authorsCollection.insertMany(originalCollection.find({}, { _id: 0, authors: 1 }).toArray());
22
23 // Create and fill 'date' db collection
24 var dateCollection = db.date_db;
25 dateCollection.insertMany(originalCollection.find({}, { _id: 0, date: 1 }).toArray());
26
```

Saved into JavaScript file and then load into mongosh



```
mongosh mongoDB://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
Type 'tt' for more
news_db>
news_db>
news_db> load("Downloads/task4.js");
news_db> db.link_db.find();
[
  {
    _id: ObjectId('65a609a7671da7584eef344'),
    link: 'https://www.huffpost.com/entry/american-airlines-passenger-banned-flight-attendant-punch-justice-department_n_632e25d3e4b06247898329fc',
  },
  {
    _id: ObjectId('65a609a7671da7584eef345'),
    link: 'https://www.huffpost.com/entry/covid-boosters-uptake-us_n_632d73ee4b087faedfeac9',
  },
  {
    _id: ObjectId('65a609a7671da7584eef346'),
    link: 'https://www.huffpost.com/entry/biden-un-russian-war-an-affront-to-bodys-charter_n_632ad9e3e4b06f75e1b5f57f',
  },
  {
    _id: ObjectId('65a609a7671da7584eef347'),
    link: 'https://www.huffpost.com/entry/amy-cooper-loses-discrimination-lawsuit-franklin-templeton_n_632c8403e4b0908781b2227e',
  },
  {
    _id: ObjectId('65a609a7671da7584eef348'),
    link: 'https://www.huffpost.com/entry/bc-soc-wcup-captains-armbands_n_632b1c98e4b0913a3d7f554a',
  },
  {
    _id: ObjectId('65a609a7671da7584eef349'),
    link: 'https://www.huffpost.com/entry/nljs-documentary-immigration-isabel-castro-interview_n_632329ee4b09098858b0da',
  },
  {
    _id: ObjectId('65a609a7671da7584eef34a'),
    link: 'https://www.huffpost.com/entry/puerto-rico-water-hurricane-fiona_n_632b0fd8e4b0d12b54014e33',
  },
  {
    _id: ObjectId('65a609a7671da7584eef34b'),
    link: 'https://www.huffpost.com/entry/funniest-tweets-cats-dogs-september-17-23_n_632de332e4b08091c1d81dc02',
  },
  {
    _id: ObjectId('65a609a7671da7584eef34c'),
    link: 'https://www.huffpost.com/entry/nas-nets-fire-protest-abe-funeral_n_632ae402e4b07198f81d6afd',
  },
  {
    _id: ObjectId('65a609a7671da7584eef34d'),
    link: 'https://www.huffpost.com/entry/fiona-threatens-to-become-category-4-storm-headed-to-bermuda_n_632ad1cae4b07198f8143244',
  },
  {
    _id: ObjectId('65a609a7671da7584eef34e'),
  },
]
```

Below I have checked all the newly added collection in news_db.

```
db.link_db.find();
```

```
db.headline_db.find();
```

```
db.category_db.find();
```

```
db.short_description_db.find();
```

```
db.authors_db.find();
```

```
db.date_db.find();
```


Task 5.

Remove all records from news_db, which have at least one empty or NULL value in object.

How many records are left in news_db?156859

```
mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
{
  "headline": "Reporter Gets Adorable Surprise From Her Boyfriend While Live On Tv",
  "category": "U.S. NEWS",
  "short_description": "Who's that behind you? an anchor for New York's PIX11 asked Journalist Michelle Ross as she finished up an interview.",
  "author": "Ethan Samuel",
  "date": "2022-09-22"
}
{
  "_id": ObjectId("63a65b07ad05540e05f3ac9"),
  "link": "https://www.huffpost.com/entry/hurricane-flora-barrels-toward-turks-and-caicos-islands_n_63298597e0de0991abc9",
  "headline": "Flora Barrels Toward Turks And Caicos Islands As Category 3 Hurricane",
  "category": "WORLD NEWS",
  "short_description": "The Turks and Caicos Islands government imposed a curfew as the intensifying storm kept dropping copious rain over the Dominican Republic and Puerto Rico.",
  "author": "Lorena Coto, AP",
  "date": "2022-09-20"
},
{
  "_id": ObjectId("63a65b07ad05540e05f3ac8"),
  "link": "https://www.huffpost.com/entry/russian-controlled-ukrainian-regions-referendum_n_6329d3ae407104f012f023",
  "headline": "A Russian-Controlled Ukrainian Regions Schedule Votes This Week To Join Russia",
  "category": "WORLD NEWS",
  "short_description": "The concerted and quickening Kremlin-backed efforts to swallow up four regions could set the stage for Moscow to escalate the war.",
  "author": "Jon Gambrell, AP",
  "date": "2022-09-20"
},
{
  "_id": ObjectId("63a65b07ad05540e05f3ac9"),
  "link": "https://www.huffpost.com/entry/golden-globes-return-abc_n_6329f151e4bde0991ab0a7f3",
  "headline": "Golden Globes Returning To ABC In January After Year Off-Air",
  "category": "ENTERTAINMENT",
  "short_description": "For the past 18 months, Hollywood has effectively boycotted the Globes after reports that the NPPA's 87 members of non-American journalists included no Black members.",
  "author": "",
  "date": "2022-09-20"
},
{
  "_id": ObjectId("63a65b07ad05540e05f3ac9"),
  "link": "https://www.huffpost.com/entry/funniest-parenting-tweets_1_n_632d7d1e40d12b540e479",
  "headline": "The Funniest Tweets From Parents This Week (Sept. 17-23)",
  "category": "PARENTING",
  "short_description": "Accidentally put grown-up toothpaste on my toddler's toothbrush and he screamed like I was cleaning his teeth with a Carolina Reaper dipped in Tabasco sauce.",
  "author": "Cassidy Wilson",
  "date": "2022-09-23"
}
}
type: "tt" for more
news_db>
news_db> load('downloads/task5.js');
Number of records remaining in 'news_db': 156859
news_db> db.newscategory.find();
{
  "_id": ObjectId("63a65b07ad05540e05f3ac9"),
  "link": "https://www.huffpost.com/entry/american-airline-passenger-banned-flight-attendant-punch-justice-department_n_632e2d3e40d27090129fe",
  "headline": "American Airlines Flyer Charged, Banned For Life After Punching Flight Attendant On Video",
  "category": "U.S. NEWS",
}
```

Task 6.

Explain how you solved the task and provide screenshots.

```
var newsCollection = db.newscategory;
```

```
// Remove records with at least one empty or null value
```

```
var result = newsCollection.deleteMany({
```

```
$or: [
```

```
{ link: { $eq: null } },
```

```
{ link: { $eq: "" } },
```

```
{ headline: { $eq: null } },
```

```
{ headline: { $eq: "" } },
```

```
{ category: { $eq: null } },
```

```
{ category: { $eq: "" } },
```

```
{ short_description: { $eq: null } },
```

```
{ short_description: { $eq: "" } },
```

```
{ authors: { $eq: null } },
```

```
{ authors: { $eq: "" } },
```

```
{ date: { $eq: null } },
```

```
{ date: { $eq: "" } }
```

```
]
```

```
});
```

```
var remainingCount = newsCollection.countDocuments();
```

```
print("Number of records remaining in 'news_db':", remainingCount);
```

Save the code in JS file and then load into mongosh after run the command

Based on the above code uses the deleteMany method to remove records that have at least one field with an empty or null value. After the deletion, it prints the number of records remaining in the news_db collection.

How many categories are in news_db? 42

```
db.newscategory.distinct("category").length;
```

Retrieves the distinct values for the "category" field in the newscategory collection.

```
mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
{
  "_id": ObjectId("65a5b0d7ad05548e05f39c6"),
  "link": "https://www.huffpost.com/entry/reporter-gets-adorable-surprise-from-her-boyfriend-while-working-live-on-tv_n_632ccf43e4b057027010074",
  "headline": "Reporter Gets Adorable Surprise From Her Boyfriend While Live On TV",
  "category": "TV & NEWS",
  "short_description": "Who's that behind you? an anchor for New York's PIX11 asked Journalist Michelle Ross as she finished up an interview.",
  "authors": "Elyse Wanshel",
  "date": "2022-09-22"
},
{
  "_id": ObjectId("65a5b0d7ad05548e05f39c7"),
  "link": "https://www.huffpost.com/entry/hurricane-fiona-barrels-toward-turks-and-caicos-islands_n_63298597e4b0e991abc6f9",
  "headline": "Fiona Barrels Toward Turks And Caicos Islands As Category 3 Hurricane",
  "category": "WORLD NEWS",
  "short_description": "The Turks and Caicos Islands government imposed a curfew as the intensifying storm kept dropping copious rain over the Dominican Republic and Puerto Rico.",
  "authors": "Helen Cote, AP",
  "date": "2022-09-20"
},
{
  "_id": ObjectId("65a5b0d7ad05548e05f39c8"),
  "link": "https://www.huffpost.com/entry/russian-controlled-ukrainian-regions-referendum_n_6329d3ae4b0719f612f023",
  "headline": "4 Russian-Controlled Ukrainian Regions Schedule Votes This Week To Join Russia",
  "category": "WORLD NEWS",
  "short_description": "The concerted and quickening Kremlin-backed efforts to swallow up four regions could set the stage for Moscow to escalate the war.",
  "authors": "Jon Cornwell, AP",
  "date": "2022-09-20"
},
{
  "_id": ObjectId("65a5b0d7ad05548e05f39ca"),
  "link": "https://www.huffpost.com/entry/funniest-parenting-tweets_1_n_632d7015e4b0d1b5403e479",
  "headline": "The Funniest Tweets From Parents This Week (Sept. 17-23)",
  "category": "Parenting",
  "short_description": "Accidentally put grown-up toothpaste on my toddler's toothbrush and he screamed like I was cleaning his teeth with a Carolina Reaper dipped in Tabasco sauce.",
  "authors": "Caroline Belkina",
  "date": "2022-09-23"
},
{
  "_id": ObjectId("65a5b0d7ad05548e05f39cb"),
  "link": "https://www.huffpost.com/entry/ukraine-festival_n_632f6aee4b0e748e05c2f",
  "headline": "Beautiful And Sad At The Same Time: Ukrainian Cultural Festival Takes On A Deeper Meaning This Year",
  "category": "POLITICS",
  "short_description": "An annual celebration took on a different feel as Russia's invasion dragged into Day 186.",
  "authors": "Jonathan Michelson",
  "date": "2022-09-19"
}
]
Type 'it' for more
news_db> db.newscategory().distinct('category').length();
news_db> db.newscategory().distinct('category').length();
news_db> db.newscategory.distinct('category').length();
news_db> db.newscategory.distinct('category').length();
news_db>
```

How many news count is for every category?

```
db.newscategory.aggregate([{$group: { _id: "$category", count: { $sum: 1 } } }]);
```

```
[
  { _id: 'PARENTS', count: 3491 },
  { _id: 'COMEDY', count: 3934 },
  { _id: 'BLACK VOICES', count: 3313 },
  { _id: 'HOME & LIVING', count: 3523 },
  { _id: 'MONEY', count: 1539 },
  { _id: 'STYLE & BEAUTY', count: 7275 },
  { _id: 'IMPACT', count: 2945 },
  { _id: 'ENTERTAINMENT', count: 13463 },
  { _id: 'U.S. NEWS', count: 1093 },
  { _id: 'FIFTY', count: 1042 },
  { _id: 'GREEN', count: 1682 },
  { _id: 'MEDIA', count: 2105 },
  { _id: 'POLITICS', count: 29685 },
  { _id: 'CULTURE & ARTS', count: 693 },
  { _id: 'HEALTHY LIVING', count: 5072 },
  { _id: 'COLLEGE', count: 860 },
  { _id: 'QUEER VOICES', count: 4700 },
  { _id: 'ARTS', count: 863 },
  { _id: 'ENVIRONMENT', count: 778 },
  { _id: 'FOOD & DRINK', count: 4527 }
]
```

```

mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
{
  date: '2022-09-20'
},
{
  _id: ObjectId('63a5b0d7ad05540e05f9ca'),
  link: 'https://www.huffpost.com/entry/funniest-parenting-tweets_1_632d7d1e400d115403e479',
  headline: 'The Funniest Tweets From Parents This Week (Sept. 17-23)',
  category: 'PARENTING',
  short_description: 'Incidentally put grown-up toothpaste on my toddler's toothbrush and he screamed like I was cleaning his teeth with a Carolina Reaper dipped in Tabasco sauce.',
  authors: 'Caroline Bologna',
  date: '2022-09-23'
},
{
  _id: ObjectId('63a5b0d7ad05540e05f9cb'),
  link: 'https://www.huffpost.com/entry/ukraine-festival_n_632f0a4be0002740be032c7',
  headline: 'Sensational And Sad At The Same Time': Ukrainian Cultural Festival Takes On A Deeper Meaning This Year',
  category: 'POLITICS',
  short_description: 'An annual celebration took on a different feel as Russia's invasion dragged into Day 280.',
  authors: 'Samantha Schuchman',
  date: '2022-09-19'
}
}
Type "!!" for more
news_db> db.newscategory().distinct('category').length();
SyntaxError: db.newscategory is not a function
news_db> db.newscategory(distinct('category')).length();
ReferenceError: distinct is not defined
news_db> db.newscategory.distinct('category').length();
0
news_db> db.newscategory.aggregate([{$group: {_id: '$category', count: { $sum: 1 } } }]);
{
  "_id": "PARENTS", "count": 3493 },
  { "_id": "COMEDY", "count": 3234 },
  { "_id": "BLACK VOICES", "count": 3313 },
  { "_id": "HOME & LIVING", "count": 3523 },
  { "_id": "MONEY", "count": 1330 },
  { "_id": "STYLE & BEAUTY", "count": 7275 },
  { "_id": "SPORTS", "count": 2345 },
  { "_id": "ENTERTAINMENT", "count": 13463 },
  { "_id": "U.S. NEWS", "count": 10093 },
  { "_id": "FIFTY", "count": 1842 },
  { "_id": "GREEN", "count": 1802 },
  { "_id": "MUSIC", "count": 2180 },
  { "_id": "POLITICS", "count": 29095 },
  { "_id": "CULTURE & ARTS", "count": 493 },
  { "_id": "HEALTHY LIVING", "count": 5072 },
  { "_id": "OUTLAGE", "count": 800 },
  { "_id": "QUEER VOICES", "count": 4100 },
  { "_id": "ARTS", "count": 863 },
  { "_id": "ENVIRONMENT", "count": 778 },
  { "_id": "FOOD & DRINK", "count": 4327 }
}
Type "!!" for more
news_db>
(To exit, press Ctrl+C again or Ctrl+D or type .exit)
news_db>

```

Explain how you solved the task.

This aggregation pipeline uses the “\$group” stage to group documents by the "category" field and then uses the \$sum accumulator to count the number of documents in each group. The result will display each category along with its corresponding news count.

Task 7.

How many news are created in 2016?

```
db.newscategory.countDocuments({ date: { $gte: "2016-01-01", $lt: "2017-01-01" } })
```

In previous code I have just defined a query that selects documents where the "date" field is greater than or equal to January 1, 2016, and less than January 1, 2017.

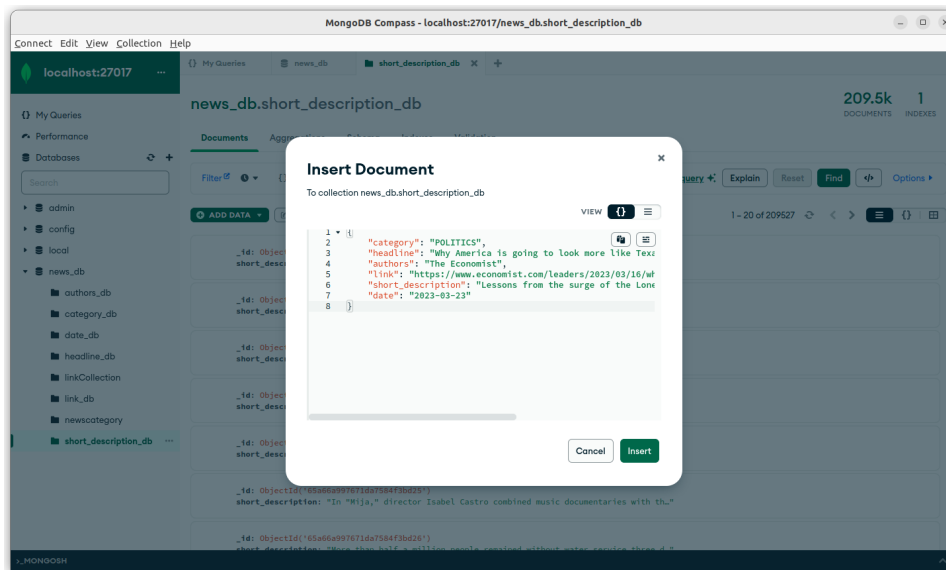
Explain how you solved the task and add the following records to the DB:

```
{ "category": "POLITICS", "headline": "Why America is going to look more like Texas", "authors": "The Economist", "link": "https://www.economist.com/leaders/2023/03/16/why-america-is-going-to-look-more-like-texas", "short_description": "Lessons from the surge of the Lone Star State", "date": "2023-03-23" }
```

```
{ "category": "POLITICS", "headline": "The Federal Reserve must choose between inflation and market chaos", "authors": "The Economist", "link": "https://www.economist.com/finance-and-economics/2023/03/19/the-federal-reserve-must-choose-between-inflation-and-market-chaos", "short_description": "Will policymakers raise interest rates as planned?", "date": "2023-03-23" }
```

Explain how you solved the task.

Since most of the things i have don through the mongosh , here I have decided to use MongoDB compass tool to connect with my local server, after that I have uploaded each file into newscategory collection.



Task 8.

Can you categorize news articles based on their headlines and short descriptions?

Let's try to categories political news based on heading and short description and also using most popular words in political news such as politics, election, government

In the code below I'm searching for keywords related to politics in both the "headline" and "short_description" fields. If a document contains any of these keywords, we set its "category" to "Politics."

```
db.newscategory.aggregate([
{
  $match: {
    $or: [
      { headline: { $regex: /politics|election|government/i } },
      { short_description: { $regex: /politics|election|government/i } }
    ]
  }
}]
```

},

{

\$set: {

category: "POLITICS"

}

}

))

```
mongosh mongod> /127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000

Type 'it' for more
news_db> db.newscategory.aggregate([{$group: { _id: "$category", count: { $sum: 1 } } }]);
[
  { _id: 'TASTE', count: 1891 },
  { _id: 'DIVORCE', count: 597 },
  { _id: 'HEALTHY LIVING', count: 5972 },
  { _id: 'CULTURE & ARTS', count: 593 },
  { _id: 'POLITICS', count: 34603 },
  { _id: 'QUEEN VOICES', count: 4788 },
  { _id: 'ARTS', count: 869 },
  { _id: 'COLLEGE', count: 868 },
  { _id: 'STYLE & BEAUTY', count: 7775 },
  { _id: 'BLACK VOICES', count: 3113 },
  { _id: 'COMEDY', count: 3334 },
  { _id: 'HOME & LIVING', count: 3523 },
  { _id: 'MONEY', count: 1539 },
  { _id: 'PARENTS', count: 1892 },
  { _id: 'MEDIA', count: 2185 },
  { _id: 'FIFTY', count: 1842 },
  { _id: 'GREEN', count: 1842 },
  { _id: 'U.S. NEWS', count: 1893 },
  { _id: 'IMPACT', count: 2845 },
  { _id: 'ENTERTAINMENT', count: 13463 }
]

Type 'it' for more
news_db> db.newscategory.aggregate([
... {
...   $set: {
...     { headline: { $regex: /politics(election|government)/ },
...       short_description: { $regex: /politics(election|government)/ } }
...   }
... }
... {
...   $set: {
...     category: "POLITICS"
...   }
... }
... ])
[
  {
    _id: ObjectId('65a5b097ad85548eb3f9b0f'),
    link: 'https://www.nytimes.com/2025/04/04/politics/asia-fire-protest-abe-funeral_n_632ae42e4b07190f81d6afd',
    headline: 'Man Sets Himself On Fire In Apparent Protest Of Funeral For Japan's Abe',
    category: 'POLITICS',
    short_description: 'The incident underscores a growing wave of protests against the funeral for Shinzo Abe, who was one of the most divisive leaders in postwar Japanese politics.',
    authors: 'Nari Yoneguchi, AP',
    date: '2025-04-04'
  },
  {
    _id: ObjectId('65a5b0d7ad85548eb3f9ac7'),
    link: 'https://www.nytimes.com/2025/04/04/world/hurricane-flora-barrel-toward-turki-and-calcos-islands_n_63298937e4b0e091d6c6cf9',
    headline: 'Flora Barrels Toward Turks And Calcos Islands As Category 3 Hurricane',
    category: 'POLITICS',
    date: '2025-04-04'
  }
]
```

Do news articles from different categories have different writing styles?

Seems yes because categories before and after values were same in the POLITICS category

Explain your answer.

Justification

Before

```
news_db> db.newscategory.aggregate([{$group: { _id: "$category", count: { $sum: 1 } } }]);
[
  { _id: 'TASTE', count: 1891 },
  { _id: 'DIVORCE', count: 1695 },
  { _id: 'HEALTHY LIVING', count: 5072 },
  { _id: 'CULTURE & ARTS', count: 693 },
  { _id: 'POLITICS', count: 29685 },
  { _id: 'QUEER VOICES', count: 4700 },
  { _id: 'ARTS', count: 863 },
  { _id: 'COLLEGE', count: 860 },
  { _id: 'STYLE & BEAUTY', count: 7275 },
  { _id: 'BLACK VOICES', count: 3313 },
  { _id: 'COMEDY', count: 3934 },
  { _id: 'HOME & LIVING', count: 3523 },
  { _id: 'MONEY', count: 1539 },
  { _id: 'PARENTS', count: 3491 },
  { _id: 'MEDIA', count: 2105 },
  { _id: 'FIFTY', count: 1042 },
  { _id: 'GREEN', count: 1682 },
  { _id: 'U.S. NEWS', count: 1093 },
  { _id: 'IMPACT', count: 2945 },
  { _id: 'ENTERTAINMENT', count: 13463 }
]
Type "it" for more
```

After

```
category: 'POLITICS',
short_description: 'Five voters from Greene's district sought to have her removed from the ballot, saying t
authors: 'Kate Brumback, AP',
date: '2022-07-26'
},
{
  _id: ObjectId('65a65bd97ad05548e05f5b1d'),
  link: 'https://www.huffpost.com/entry/trial-expected-to-begin-for-ex-trump-adviser-steve-bannon_n_62d51ff6e
headline: 'Jury Selection Begins In Trial Of Ex-Trump Adviser Steve Bannon',
category: 'POLITICS',
short_description: 'Bannon is charged in Washington's federal court with defying a subpoena from the Jan. 6
authors: 'Gary Fields, Ashraf Khalil, AP',
date: '2022-07-18'
}
]
Type "it" for more
news_db> db.newscategory.aggregate([{$group: { _id: "$category", count: { $sum: 1 } } }]);
[
  { _id: 'ENTERTAINMENT', count: 13463 },
  { _id: 'IMPACT', count: 2945 },
  { _id: 'U.S. NEWS', count: 1093 },
  { _id: 'GREEN', count: 1682 },
  { _id: 'MEDIA', count: 2105 },
  { _id: 'PARENTS', count: 3491 },
  { _id: 'MONEY', count: 1539 },
  { _id: 'COMEDY', count: 3934 },
  { _id: 'HOME & LIVING', count: 3523 },
  { _id: 'BLACK VOICES', count: 3313 },
  { _id: 'STYLE & BEAUTY', count: 7275 },
  { _id: 'FIFTY', count: 1042 },
  { _id: 'COLLEGE', count: 860 },
  { _id: 'QUEER VOICES', count: 4700 },
  { _id: 'ARTS', count: 863 },
  { _id: 'POLITICS', count: 29685 },
  { _id: 'CULTURE & ARTS', count: 693 },
  { _id: 'HEALTHY LIVING', count: 5072 },
  { _id: 'DIVORCE', count: 1695 },
  { _id: 'EDUCATION', count: 893 }
]
Type "it" for more
```

Task 9.

How many news are about Turkey?

```
db.newscategory.countDocuments({
```

```
$or: [
```

```
{ headline: { $regex: /turkey/i } },
{ short_description: { $regex: /turkey/i } }
]
});
```

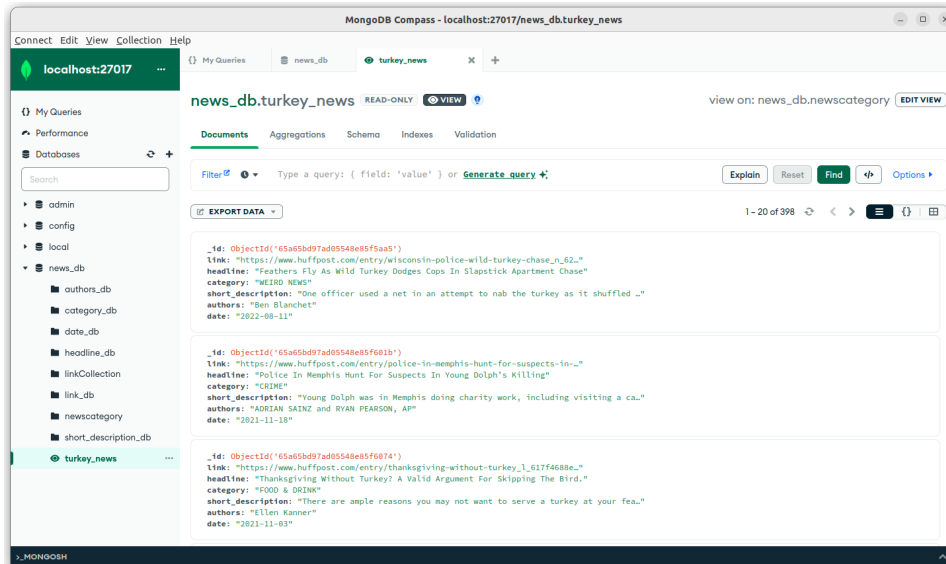
```
news_db> ;
news_db> db.newscategory.countDocuments({
...   $or: [
...     { headline: { $regex: /turkey/i } },
...     { short_description: { $regex: /turkey/i } }
...   ]
... });
398
news_db> □
```

Create a view turkey_news containing news only about Turkey.

Explain your answer.

```
news_db> db.createView("turkey_news", "newscategory", [
...   {
...     $match: {
...       $or: [
...         { headline: { $regex: /turkey/i } },
...         { short_description: { $regex: /turkey/i } }
...       ]
...     }
...   }
... ], {
...   ok: 1
... });
news_db> □
```

```
db.createView("turkey_news", "newscategory", [
{
$match: {
$or: [
{ headline: { $regex: /turkey/i } },
{ short_description: { $regex: /turkey/i } }
]
}
}
]);
```

The regular expression `"/turkey/i "` is used for case-insensitive matching. This ensures that variations like "Turkey," "turkey," or "TURKEY" are all considered when regex matching.

The `$or` operator is used to match documents where the keyword "Turkey" is present in either the headline or short description.

Creating a View: The `createView` method is used to create a view named `turkey_news`. The view is based on the original `newscategory` collection, and the documents in the view are filtered based on the specified criteria.

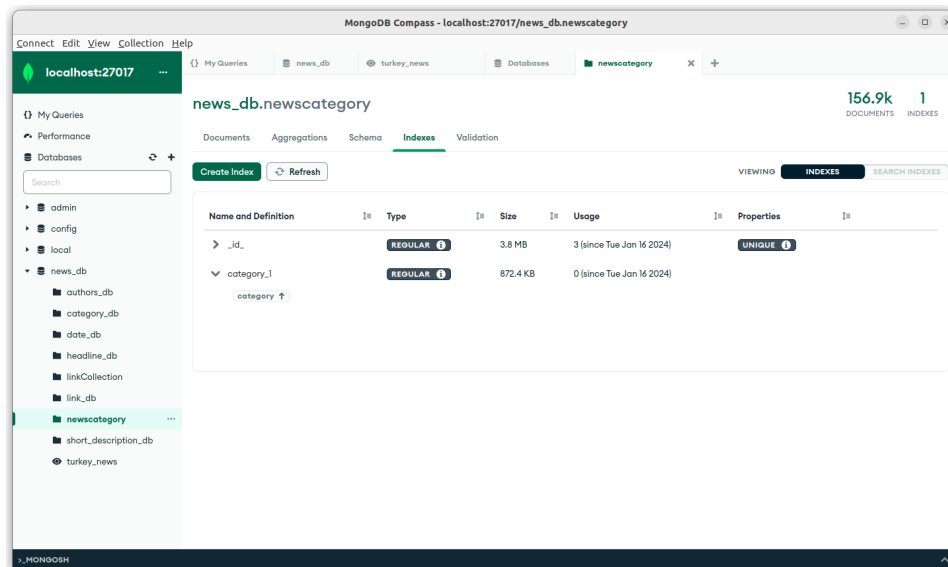
Task 10.

Add indexes to `news_db` based on news categories.

Explain how you solved the task.

```
db.newscategory.createIndex({ category: 1 });
```

To enhance query performance, I added an index to the "category" field in the `newscategory` collection using MongoDB's `createIndex` method. This index enables efficient retrieval of documents based on their categories, optimizing queries involving category-based filtering, sorting, or aggregation. Indexing is a crucial optimization strategy, and it's essential to consider query patterns and strike a balance between improved read performance and potential write operation overhead.



Task 11.

Add new collection that contains number of symbols in short_description.

Explain how you solved the task.

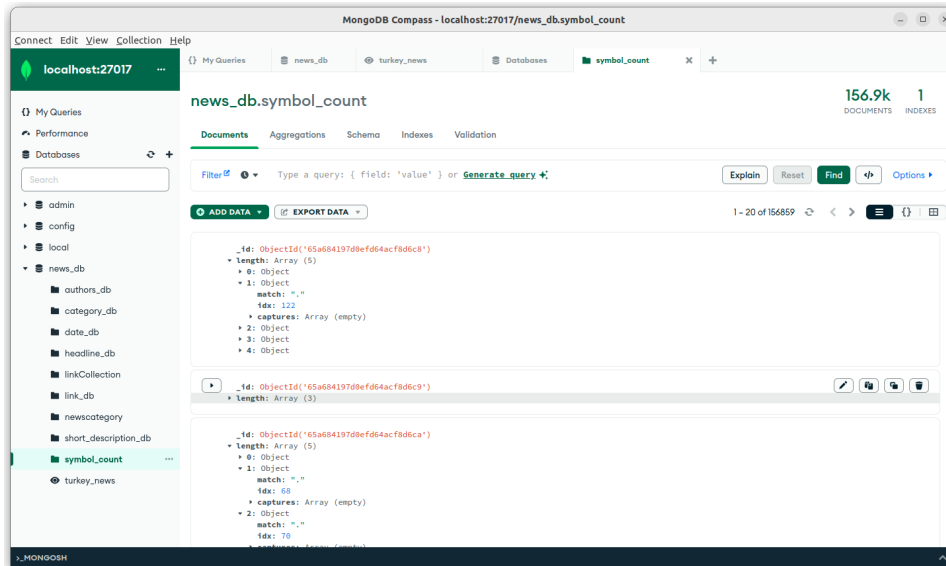
```
db.createCollection("symbol_count");

db.symbol_count.insertMany(
  db.newscategory.aggregate([
    {
      $project: {
        _id: 0,
        length: { $regexFindAll: {
          input: "$short_description",
          regex: /^[^a-zA-Z0-9\s]/g
        }}
      }
    }
  ]).toArray()
);
```

```

9 db.symbol_count.insertMany(
10   db.newscategory.aggregate([
11     {
12       $project: {
13         _id: 0,
14         length: { $regexFindAll: {
15           input: "$short_description",
16           regex: /^[^a-zA-Z0-9\s]/g
17         } }
18       }
19     }
20   ]).toArray()
21 );

```



The createCollection method establishes a new 'symbol_count' collection. Using the aggregation framework, the \$project stage calculates the length of each short description using \$regexFindAll, ensuring accurate Unicode code point counting based on given regex filtering. The results are inserted into 'symbol_count'.

Task 12.

Remove obsolete records that are older than 1 Jan 2016 from news db. How many records left in the database?

Explain how you solved the task.

Initial data count - `db.newscategory.countDocuments();` - 156859

```
news_db> db.newscategory.deleteMany({ date: { $lt: Date("2016-01-01T00:00:00Z") } });
{ acknowledged: true, deletedCount: 156859 }
news_db> db.newscategory.deleteMany({ date: { $lt: "2016-01-01T00:00:00Z" } });
{ acknowledged: true, deletedCount: 130357 }
news_db> db.newscategory.countDocuments();
79170
news_db> 
```

Deleted document count - `db.newscategory.deleteMany({ date: { $lt: "2016-01-01T00:00:00Z" } });` - **130357**

Remaining count - 79170

Task 13.

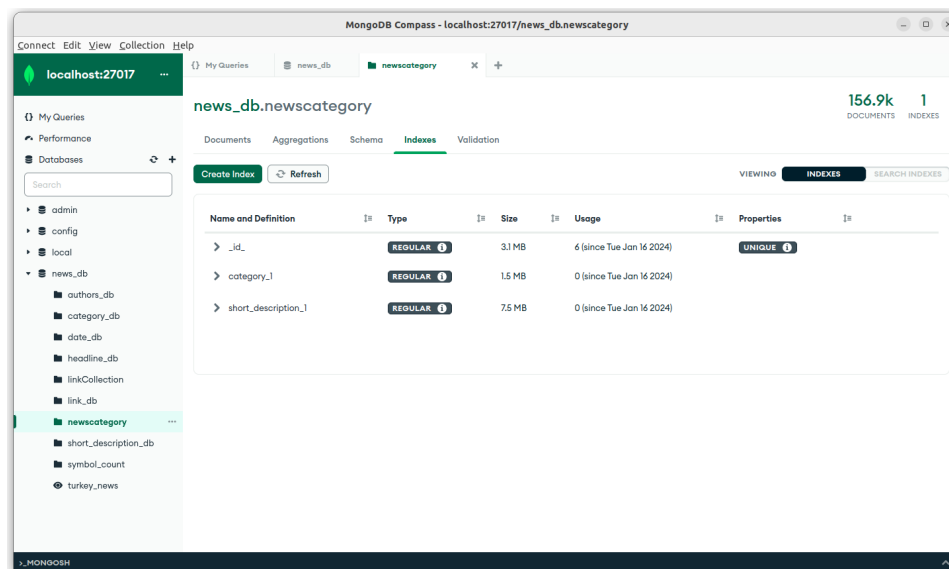
Apply aggregated function to news_db that will sort dataset ascending based on indexed categories and length of short_description. (You worked on them in previous tasks).

Explain how you solved the task.

First create index in both categories in ascending order and length of short_description

`db.newscategory.createIndex({ category: 1 });`

`db.newscategory.createIndex({ short_description: 1 });`



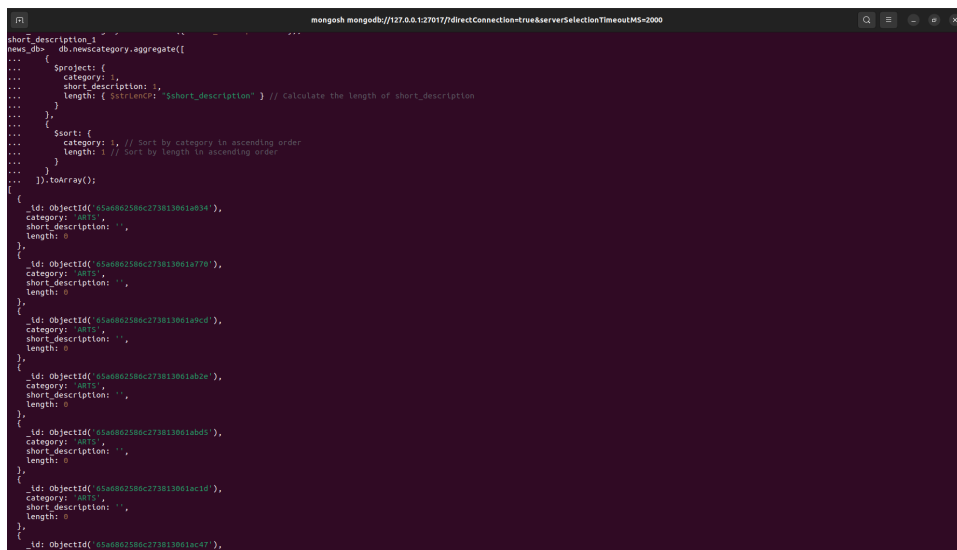
Use the \$sort stage in the aggregation pipeline to sort the dataset based on the "category" field in ascending order. Additionally, sort based on the assumed "length" field, which should contain the length of the short_description.

`db.newscategory.aggregate([`

```

{
  $project: {
    category: 1,
    short_description: 1,
    length: { $strLenCP: "$short_description" } // Calculate the length of short_description
  }
},
{
  $sort: {
    category: 1, // Sort by category in ascending order
    length: 1 // Sort by length in ascending order
  }
}
]).toArray();

```



```

mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000
short_description_1
news_db> db.news.category.aggregate([
... {
...   $project: {
...     category: 1,
...     short_description: 1,
...     length: { $strLenCP: "$short_description" } // Calculate the length of short_description
...   }
... },
... {
...   $sort: {
...     category: 1, // Sort by category in ascending order
...     length: 1 // Sort by length in ascending order
...   }
... },
... ]).toArray();
[
  {
    _id: ObjectId('65a68d586c2738138d1a934'),
    category: 'news',
    short_description: '',
    length: 0
  },
  {
    _id: ObjectId('65a68d586c2738138d1a770'),
    category: 'news',
    short_description: '',
    length: 0
  },
  {
    _id: ObjectId('65a68d586c2738138d1a9cd'),
    category: 'news',
    short_description: '',
    length: 0
  },
  {
    _id: ObjectId('65a68d586c2738138d1a2e'),
    category: 'news',
    short_description: '',
    length: 0
  },
  {
    _id: ObjectId('65a68d586c2738138d1ad1'),
    category: 'news',
    short_description: '',
    length: 0
  },
  {
    _id: ObjectId('65a68d586c2738138d1ad'),
    category: 'news',
    short_description: '',
    length: 0
  },
  {
    _id: ObjectId('65a68d586c2738138d1ad7'),
    category: 'news',
    short_description: '',
    length: 0
  }
]

```

In the above MongoDB aggregation pipeline, the `createIndex` method is initially used to ensure indexes on the "category" and "short_description" fields for optimized sorting. The aggregation pipeline then

employs the \$project stage to calculate the length of the "short_description" using the \$strLenCP operator, creating a new field called "length." Subsequently, the \$sort stage arranges the dataset in ascending order based on both the "category" and "length" fields. The resulting sorted dataset is converted to an array using toArray(). This comprehensive process facilitates efficient sorting of the news_db collection, incorporating both category and short_description length considerations.