Practical 8

Aim: Implement an application that stores big data in Hbase / MongoDB and manipulate it using R / Python.

Theory:

MongoDB is an open-source and the leading NoSQL database. It is a document-oriented database that offers high performance, easy scalability, and high availability. It uses documents and collections to organize data rather than relations. This makes it an ideal database management system for the storage of unstructured data.

MongoDB uses replica sets to ensure there is a high availability of data. Each replica set is made up of two or more replicas of data. This gives its users the ability to access their data at any time. The replica sets also create fault tolerance. MongoDB scales well to accommodate more data. It uses the sharing technique to scale horizontally and meet the changing storage needs of its users. MongoDB was developed to help developers unleash the power of data and software.

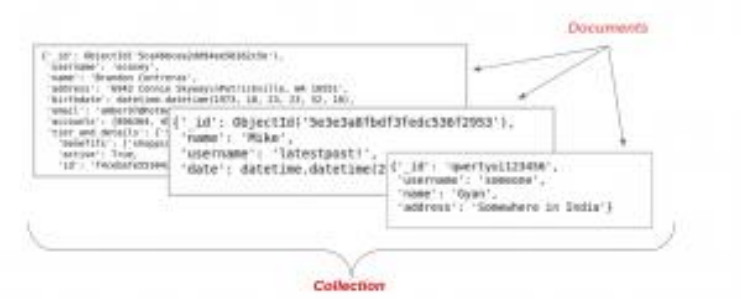
MongoDB is an unstructured database. It stores data in the form of documents. MongoDB is able to handle huge volumes of data very efficiently and is the most widely used NoSQL database as it offers rich query language and flexible and fast access to data.

The Architecture of a MongoDB Database

The information in MongoDB is stored in ***documents***. Here, a document is analogous to ***rows*** in structured databases.

- Each document is a collection of key-value pairs
- Each key-value pair is called a ***field***
- Every document has an ***_id*** field, which uniquely identifies the documents
- A document may also contain nested documents
- Documents may have a varying number of fields (they can be blank as well)

These documents are stored in a ***collection***. A collection is literally a collection of documents in MongoDB. This is analogous to ***tables*** in traditional databases.



Unlike traditional databases, the data is generally stored in a single collection in MongoDB, so there is no concept of joins (except ***$lookup*** operator, which performs ***left-outer-join*** like operation). MongoDB has the nested document instead.

PyMongo is a Python library that enables us to connect with MongoDB. It allows us to perform basic operations on the MongoDB database.

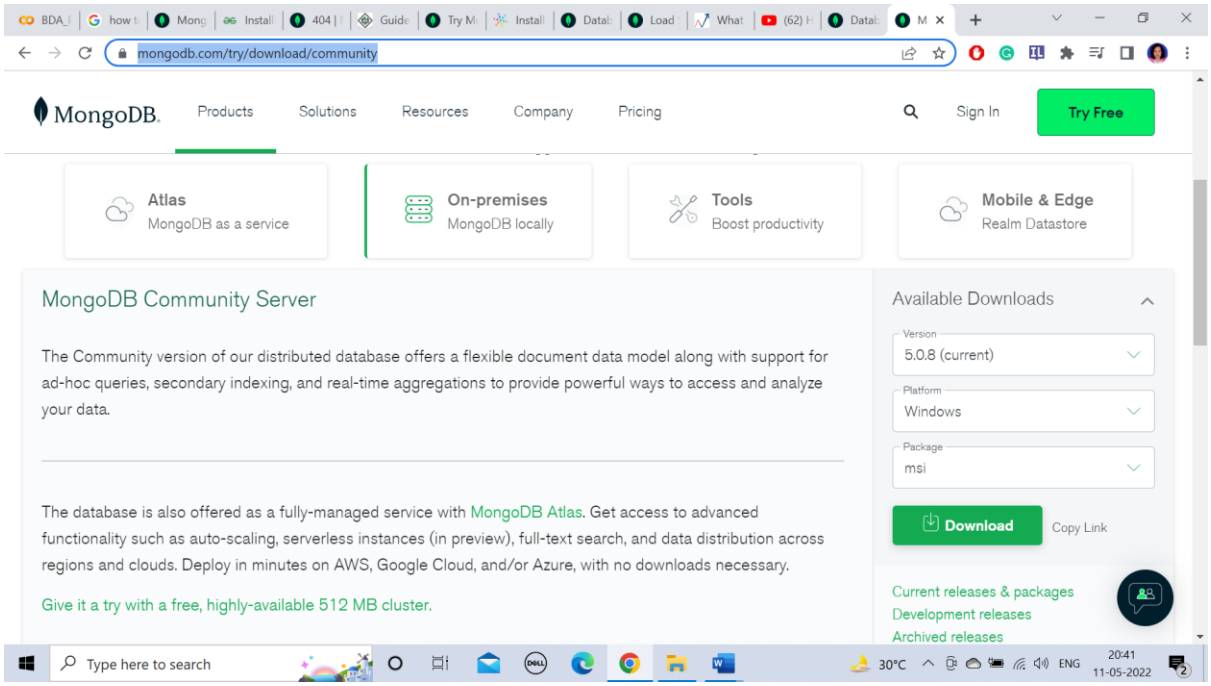
We have chosen Python to interact with MongoDB because it is one of the most commonly used and considerably powerful languages for data science. PyMongo allows us to retrieve the data with dictionary-like syntax.

We can also use the dot notation to access MongoDB data. Its easy syntax makes our job a lot easier. Additionally, PyMongo's rich documentation is always standing there with a helping hand. We will use this library for accessing MongoDB.

Steps of the installation:

Step 1: Download MongoDB

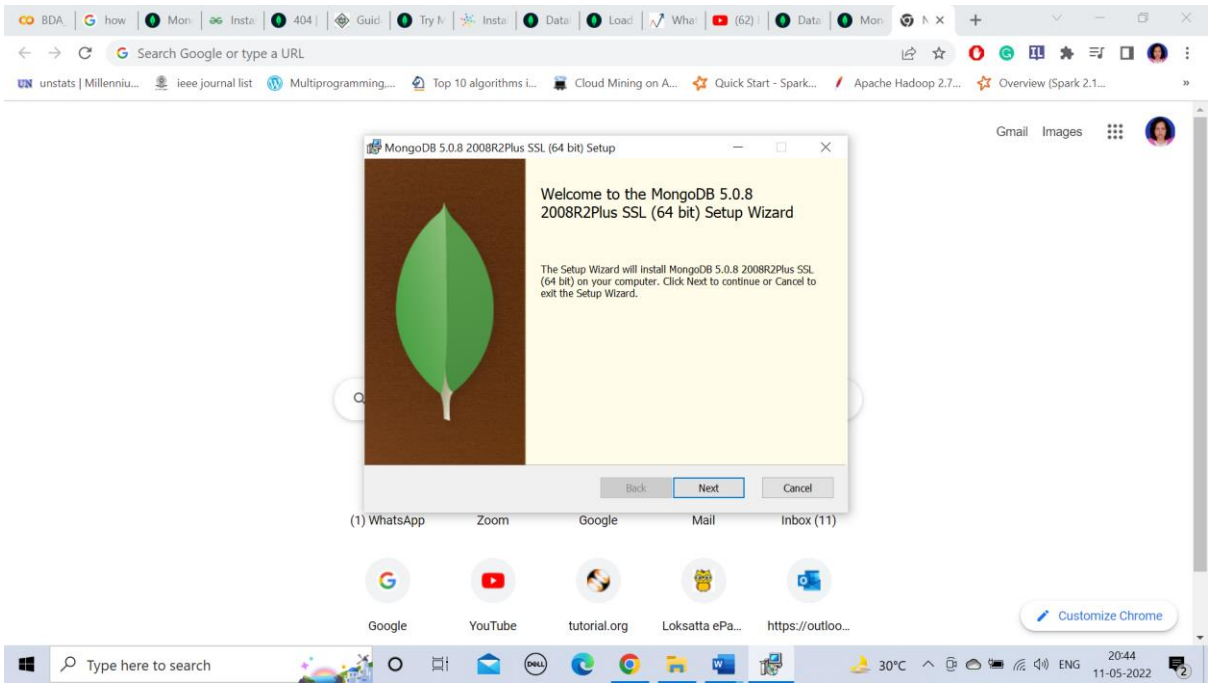
Go to official website: MongoDB Community server
https://www.mongodb.com/try/download/community

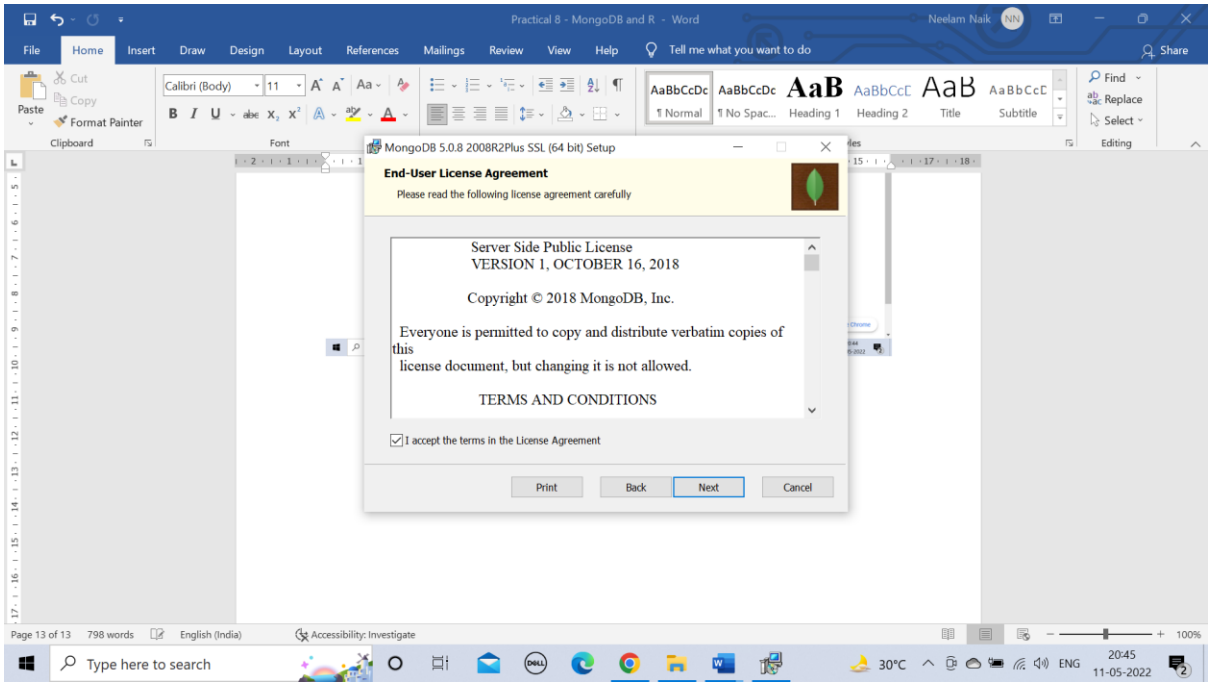


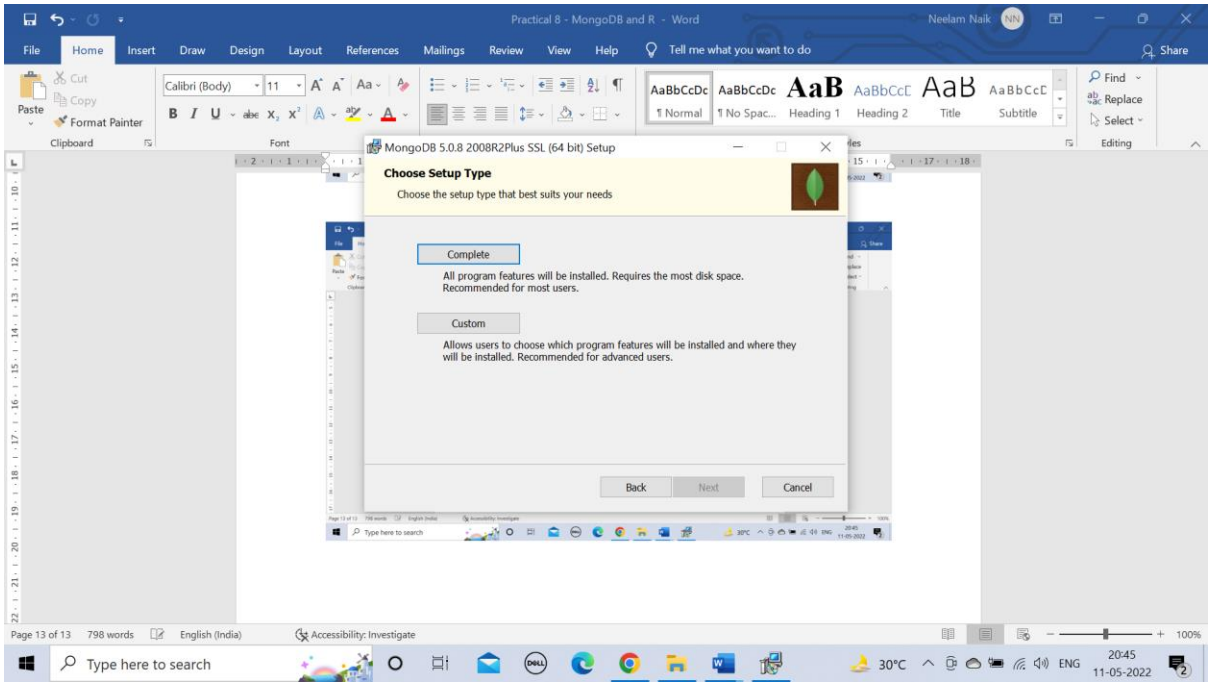
Click on download

Step 2: Install MongoDB

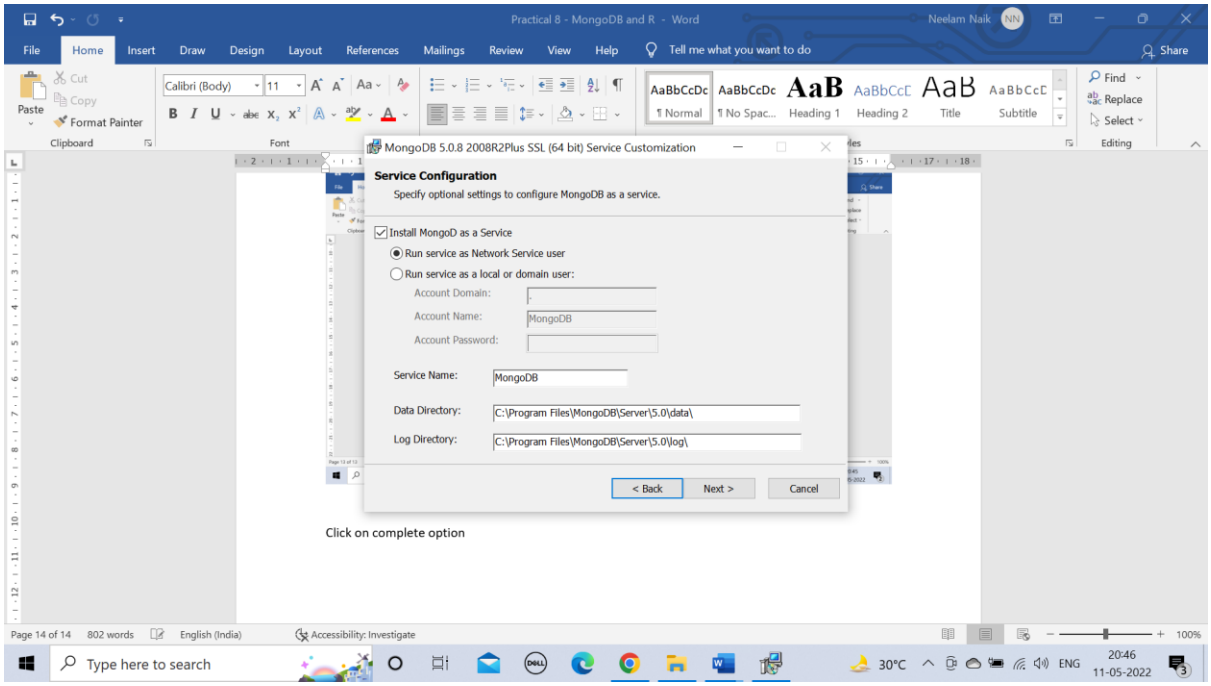
It will download msi file. Click on it and Start the installation.



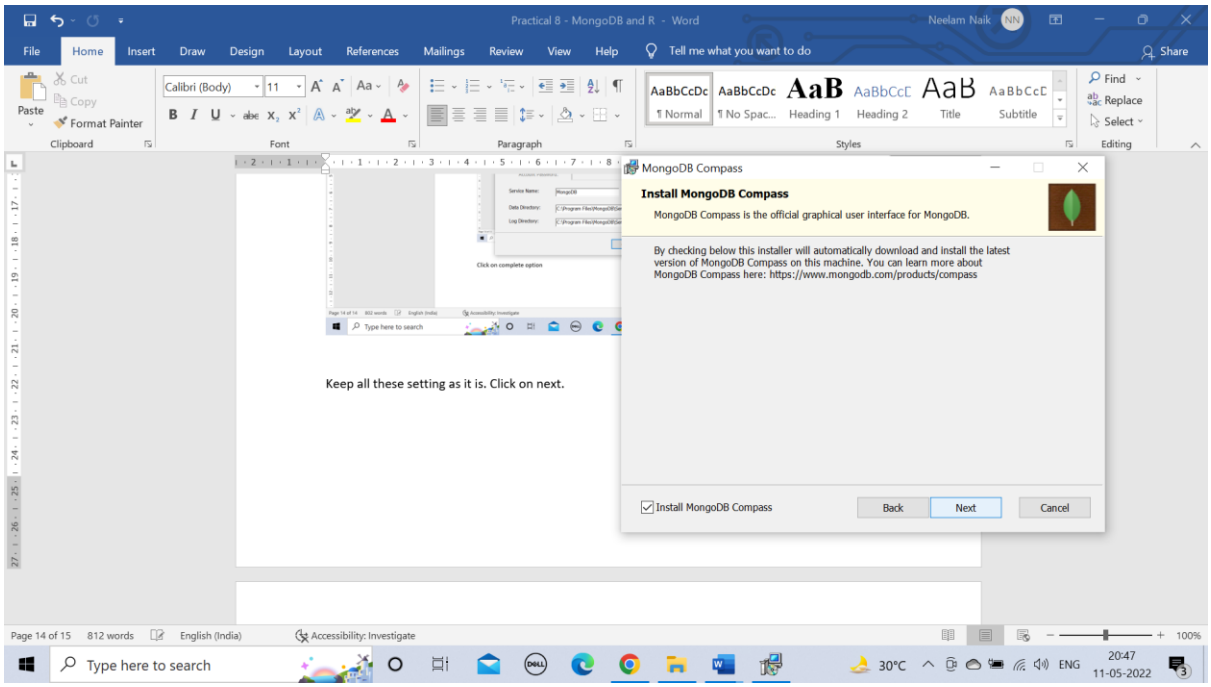




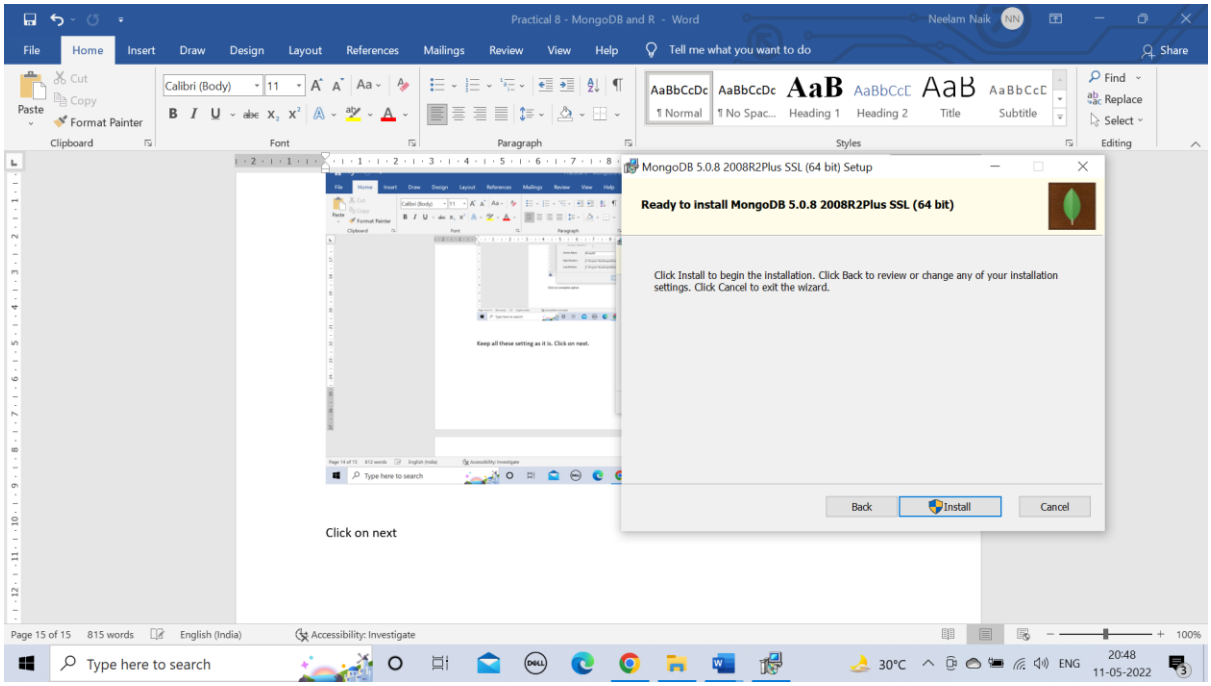
Click on complete option



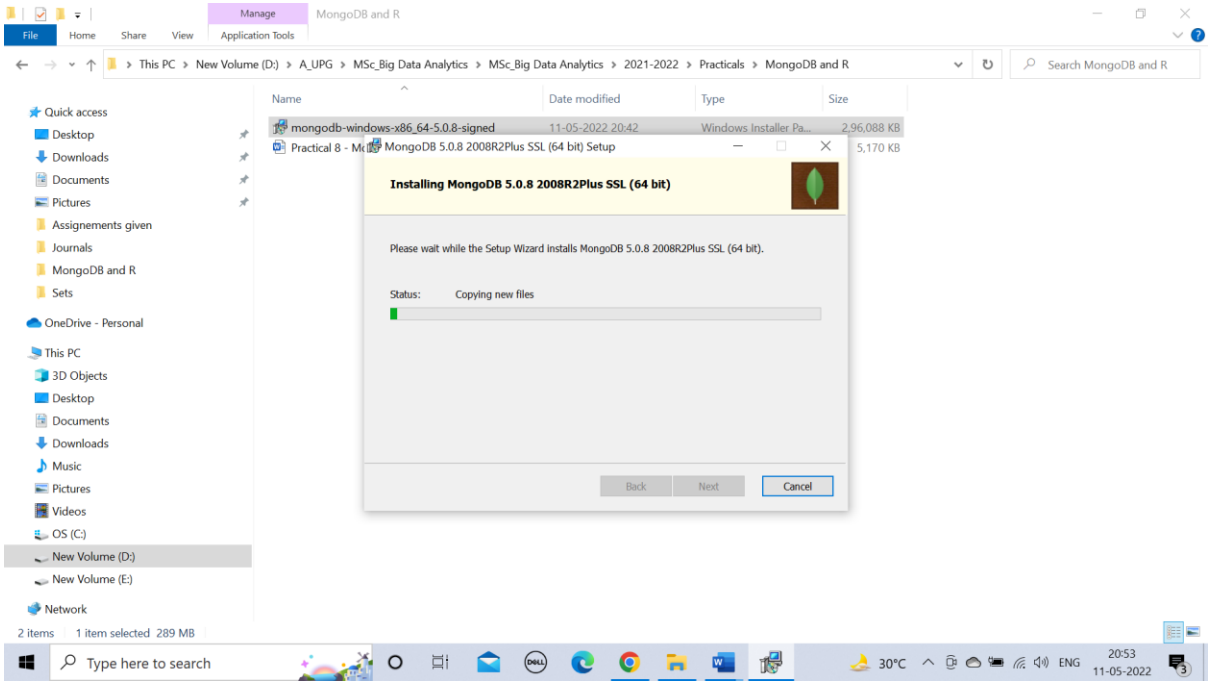
Keep all these setting as it is. Click on next.

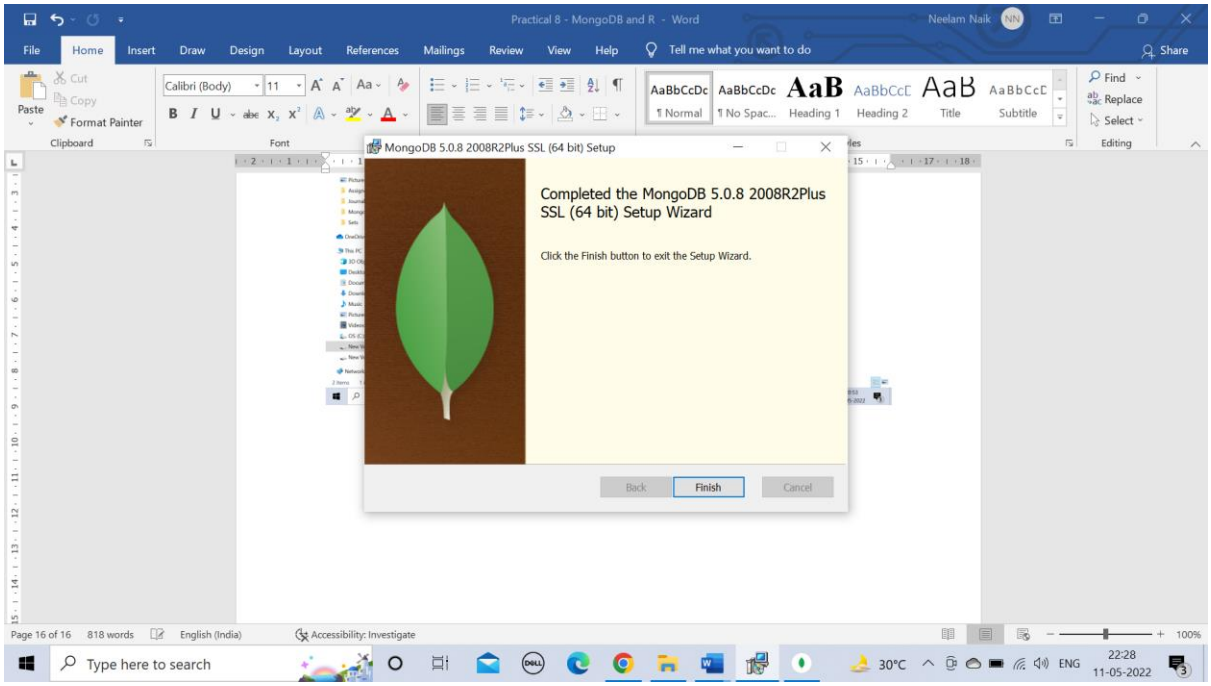


Click on next



Click on install



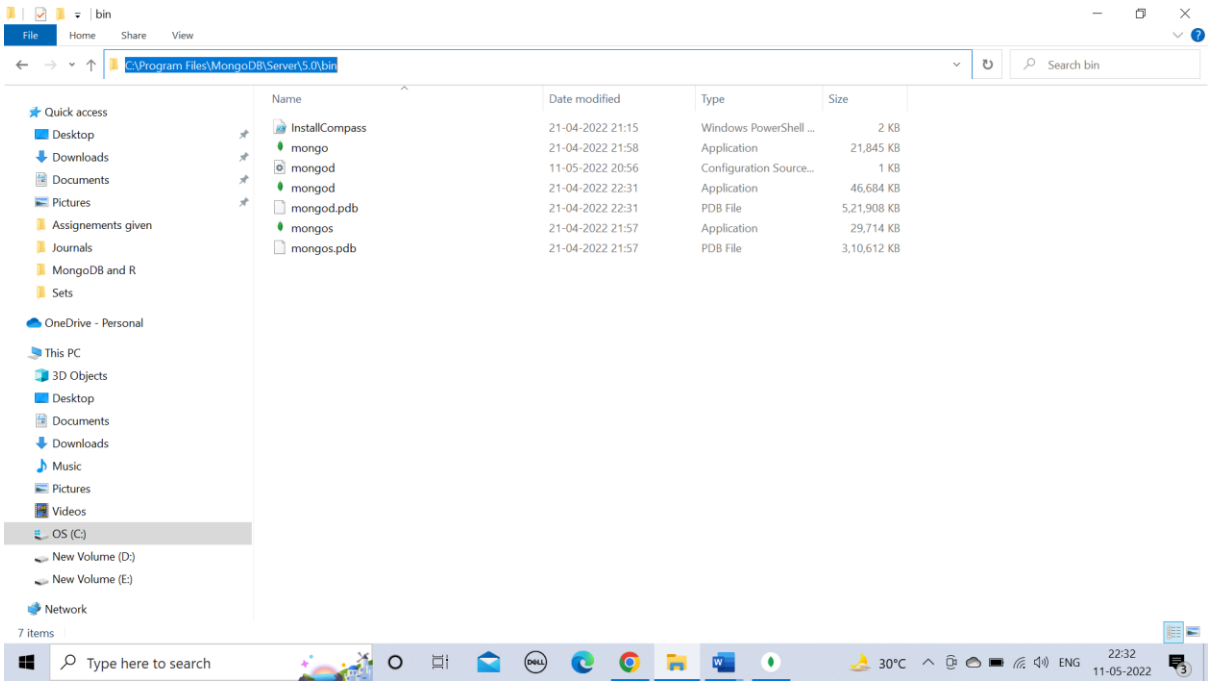


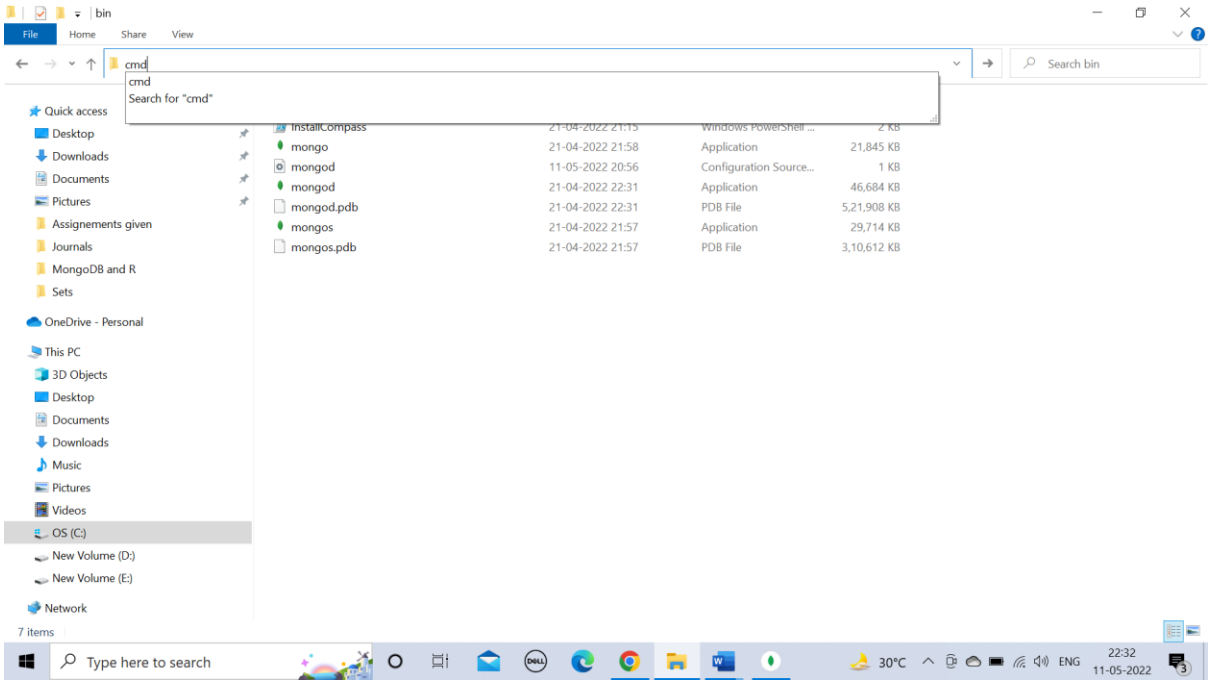
Step 3: Verify MongoDB Installation

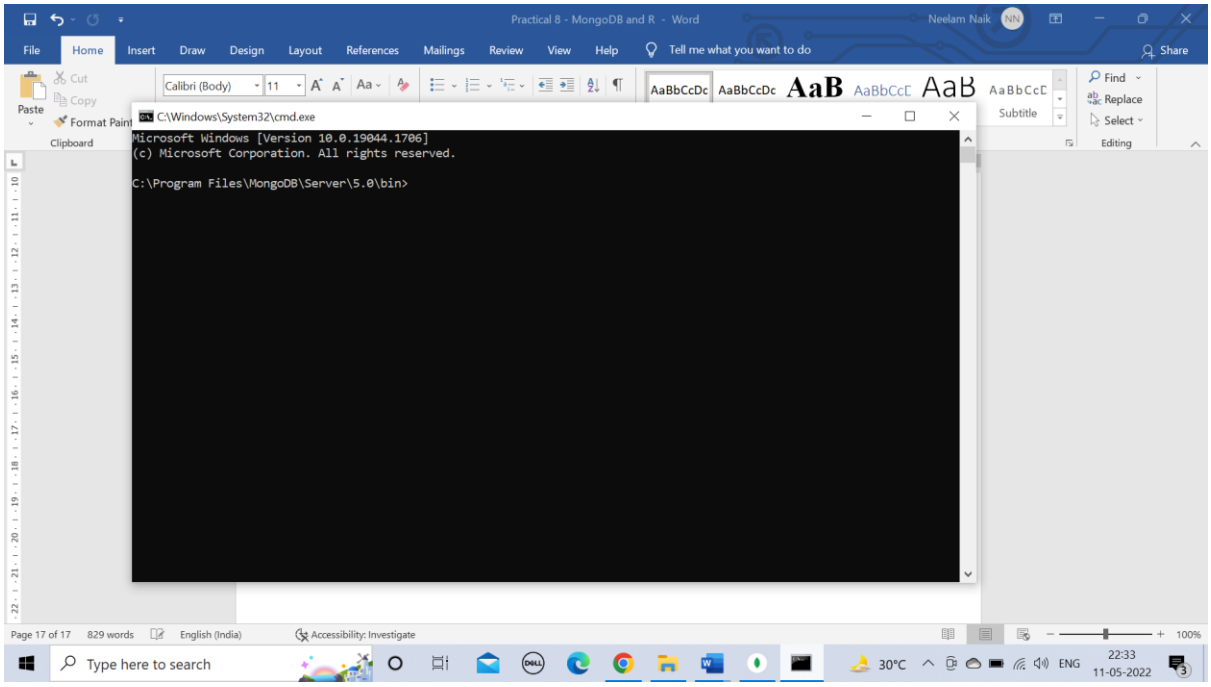
Now go to

C:\Program Files\MongoDB\Server\5.0\bin

Start command prompt from this location

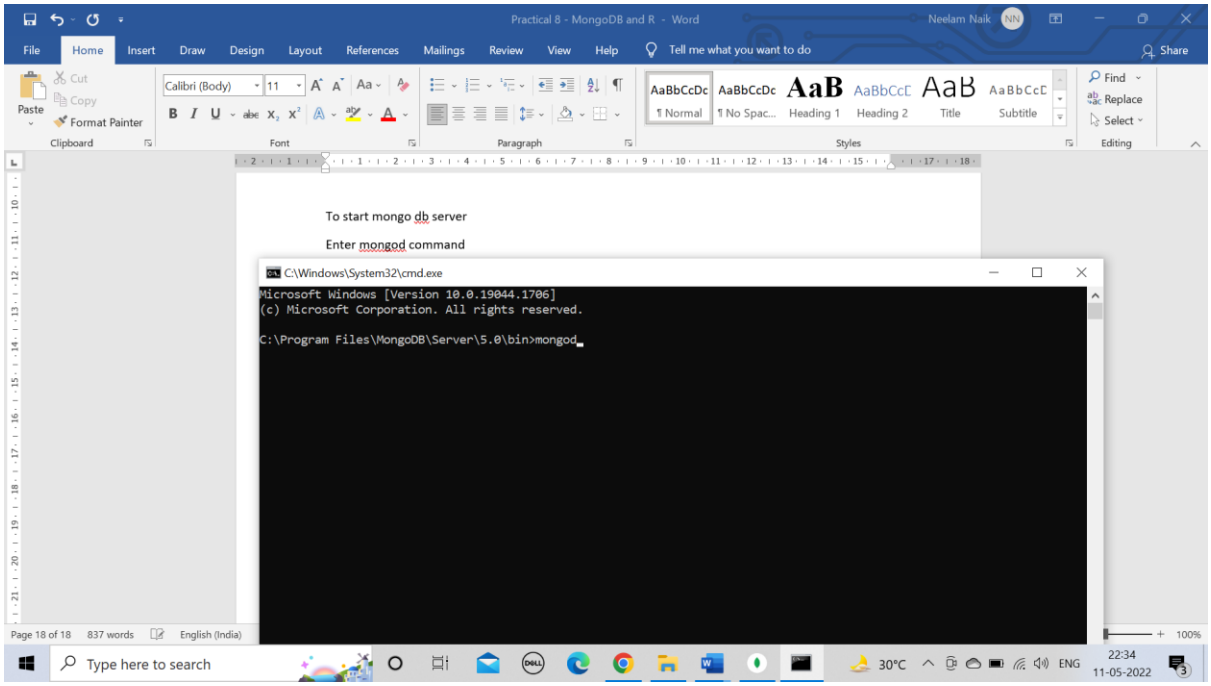


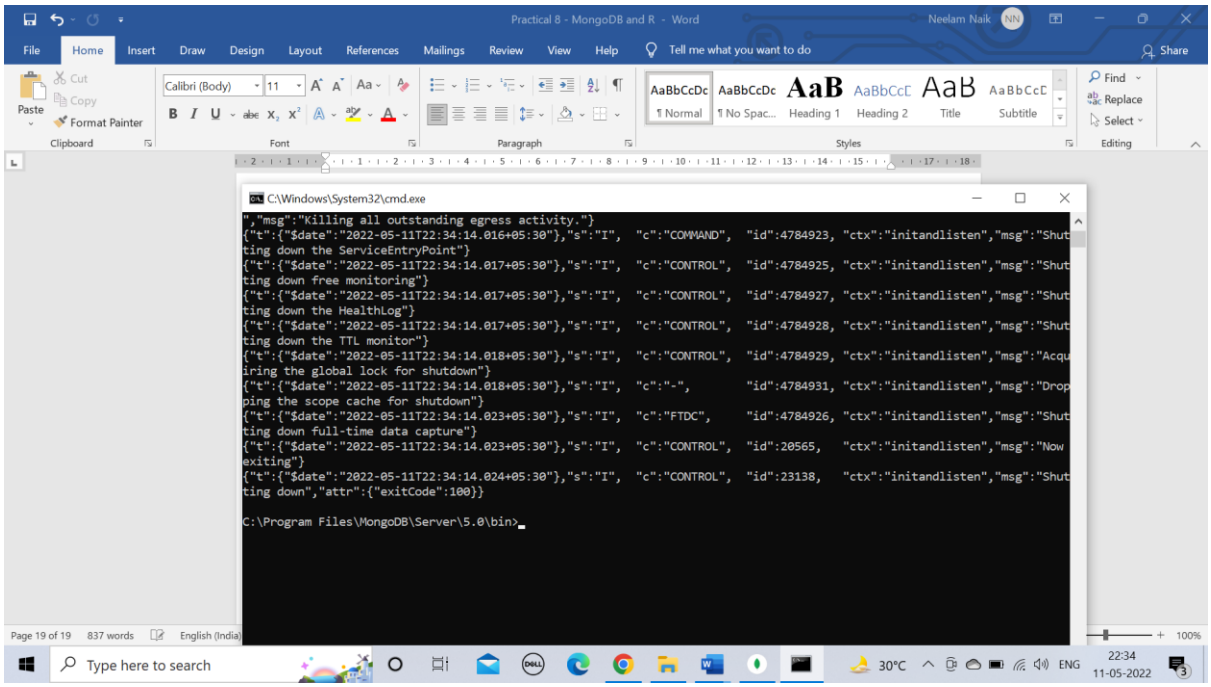




To start mongo db server

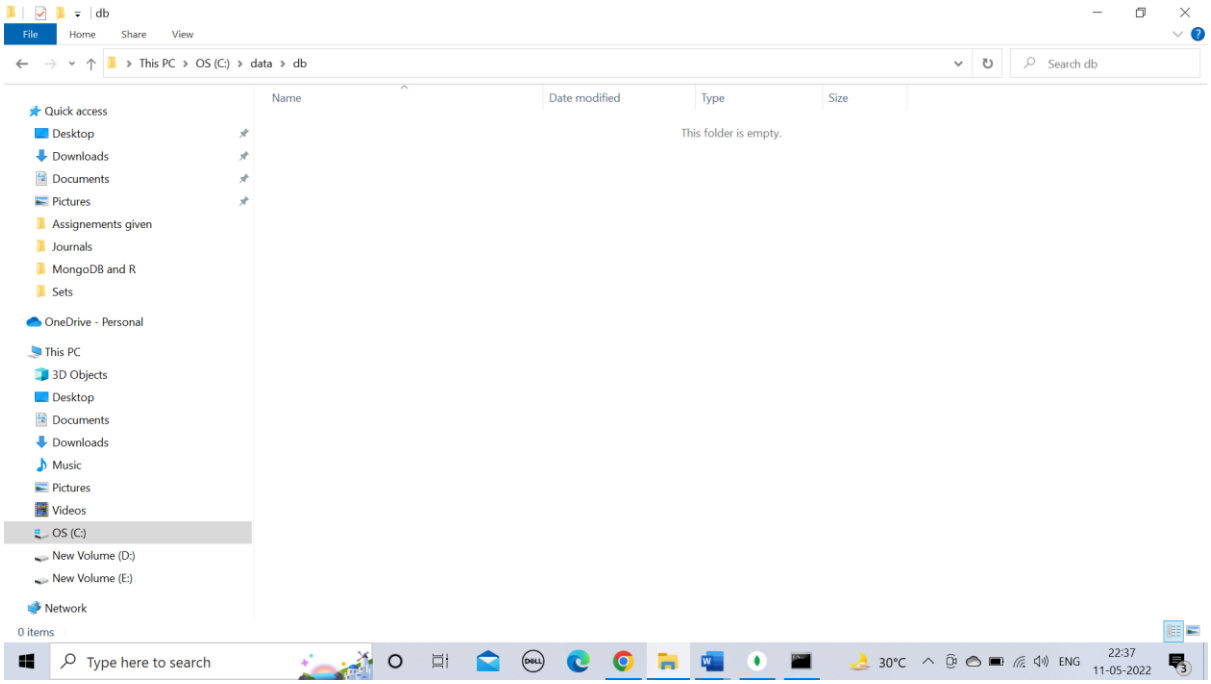
Enter mongod command


Practical 8 - MongoDB and R - Word
Neelam Naik
File Home Insert Draw Design Layout References Mailings Review View Help Tell me what you want to do Share
Paste Cut Copy Format Painter
Calibri (Body) 11
Clipboard Font Paragraph Styles Editing
Find Replace Select
To start mongo db server
Enter mongod command
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19044.1706]
(c) Microsoft Corporation. All rights reserved.
C:\Program Files\MongoDB\Server\5.0\bin>mongod
Page 18 of 18 837 words English (India)
100%
Type here to search
30°C ENG 22:34 11-05-2022

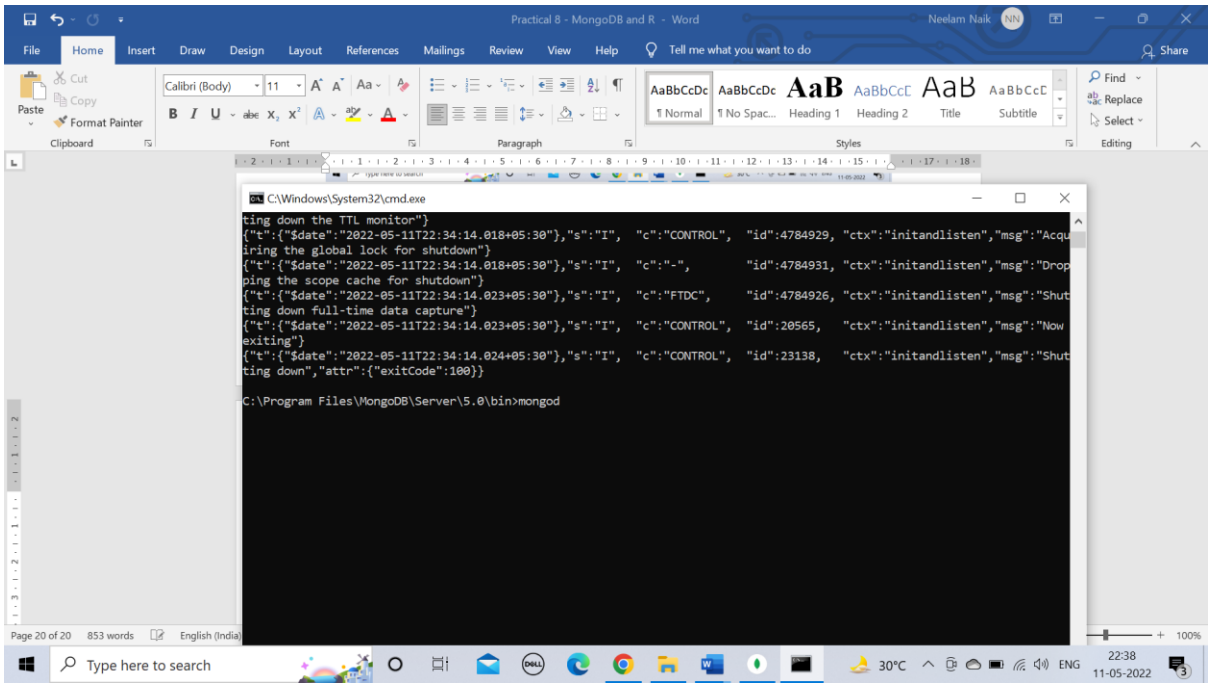

Practical 8 - MongoDB and R - Word
Neelam Naik
File Home Insert Draw Design Layout References Mailings Review View Help Tell me what you want to do Share
Paste Cut Copy Format Painter
Calibri (Body) 11
Clipboard Font Paragraph Styles Editing
Find Replace Select
C:\Windows\System32\cmd.exe
","msg":"Killing all outstanding egress activity."}
{"t":{"$date":"2022-05-11T22:34:14.016+05:30"},"s":"I", "c":"COMMAND", "id":4784923, "ctx":"initandlisten","msg":"Shut
ting down the ServiceEntryPoint"}
{"t":{"$date":"2022-05-11T22:34:14.017+05:30"},"s":"I", "c":"CONTROL", "id":4784925, "ctx":"initandlisten","msg":"Shut
ting down free monitoring"}
{"t":{"$date":"2022-05-11T22:34:14.017+05:30"},"s":"I", "c":"CONTROL", "id":4784927, "ctx":"initandlisten","msg":"Shut
ting down the HealthLog"}
{"t":{"$date":"2022-05-11T22:34:14.017+05:30"},"s":"I", "c":"CONTROL", "id":4784928, "ctx":"initandlisten","msg":"Shut
ting down the TTL monitor"}
{"t":{"$date":"2022-05-11T22:34:14.018+05:30"},"s":"I", "c":"CONTROL", "id":4784929, "ctx":"initandlisten","msg":"Acqu
iring the global lock for shutdown"}
{"t":{"$date":"2022-05-11T22:34:14.018+05:30"},"s":"I", "c":"-", "id":4784931, "ctx":"initandlisten","msg":"Drop
ping the scope cache for shutdown"}
{"t":{"$date":"2022-05-11T22:34:14.023+05:30"},"s":"I", "c":"FTDC", "id":4784926, "ctx":"initandlisten","msg":"Shut
ting down full-time data capture"}
{"t":{"$date":"2022-05-11T22:34:14.023+05:30"},"s":"I", "c":"CONTROL", "id":20565, "ctx":"initandlisten","msg":"Now
exiting"}
{"t":{"$date":"2022-05-11T22:34:14.024+05:30"},"s":"I", "c":"CONTROL", "id":23138, "ctx":"initandlisten","msg":"Shut
ting down","attr":{"exitCode":100}}
C:\Program Files\MongoDB\Server\5.0\bin>
Page 19 of 19 837 words English (India)
100%
Type here to search
30°C ENG 22:34 11-05-2022

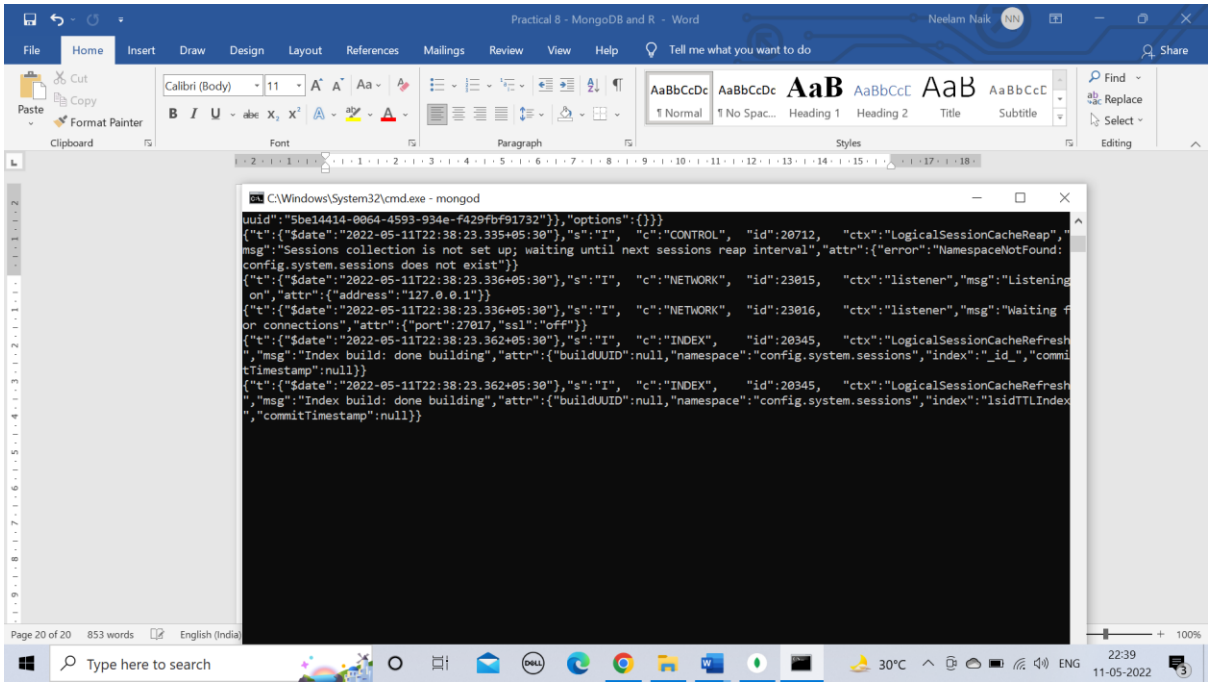
It says c:\data\db directory not found

So create c:\data\db



Run the above command mongod once again





Mongo daemon is started now

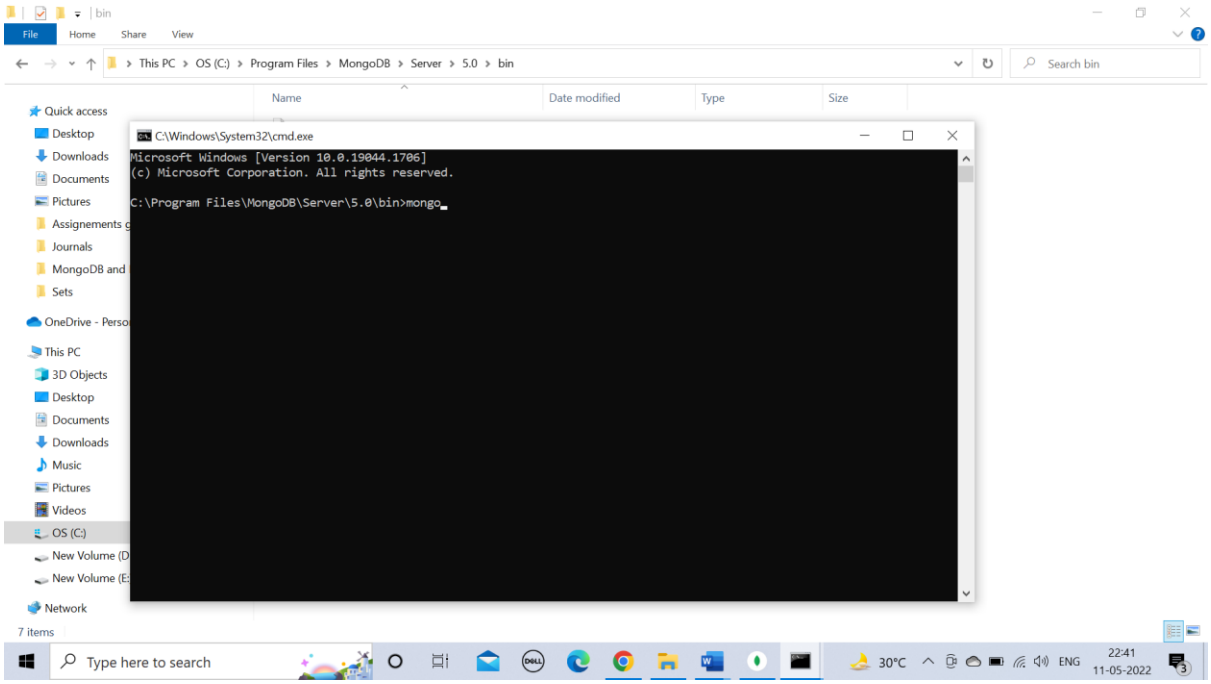
To open the mongo shell

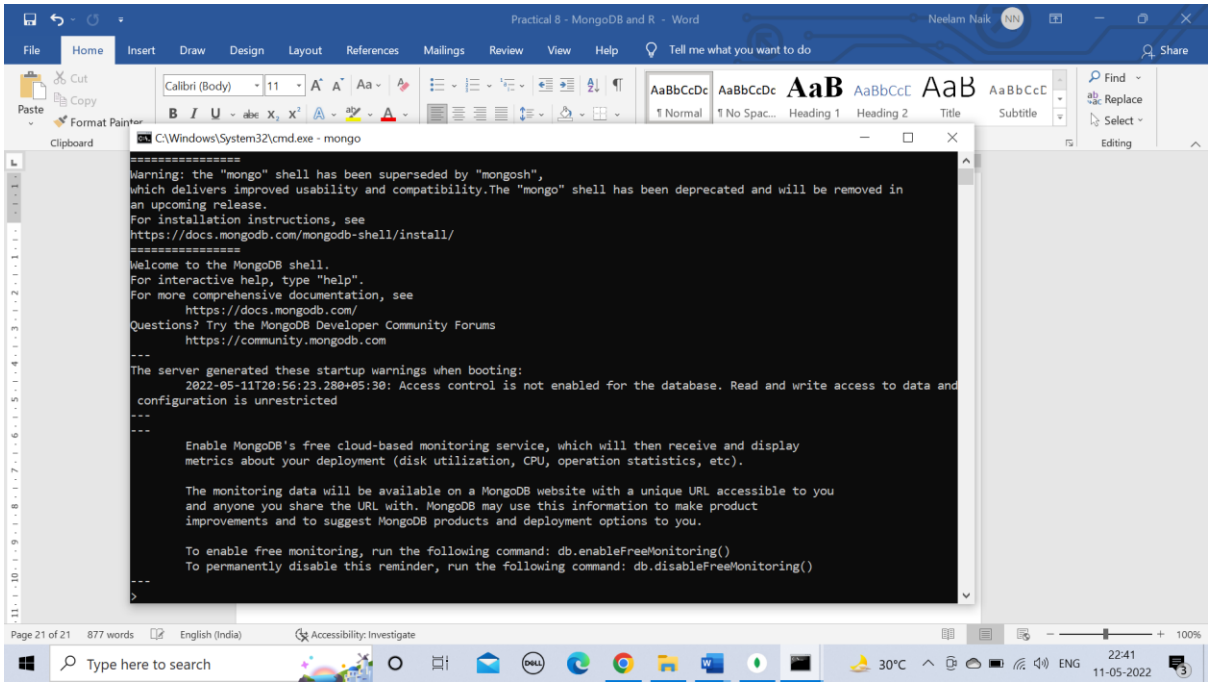
Go to

C:\Program Files\MongoDB\Server\5.0\bin

Start command prompt from this location

Fire the command: mongo

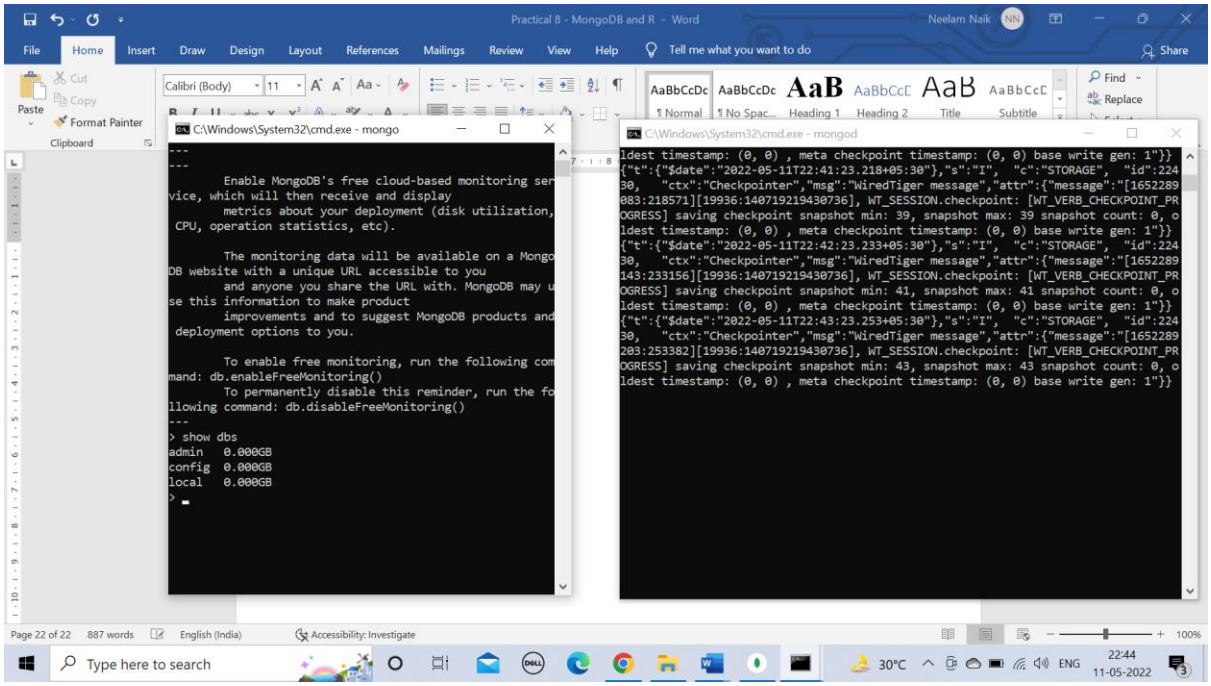




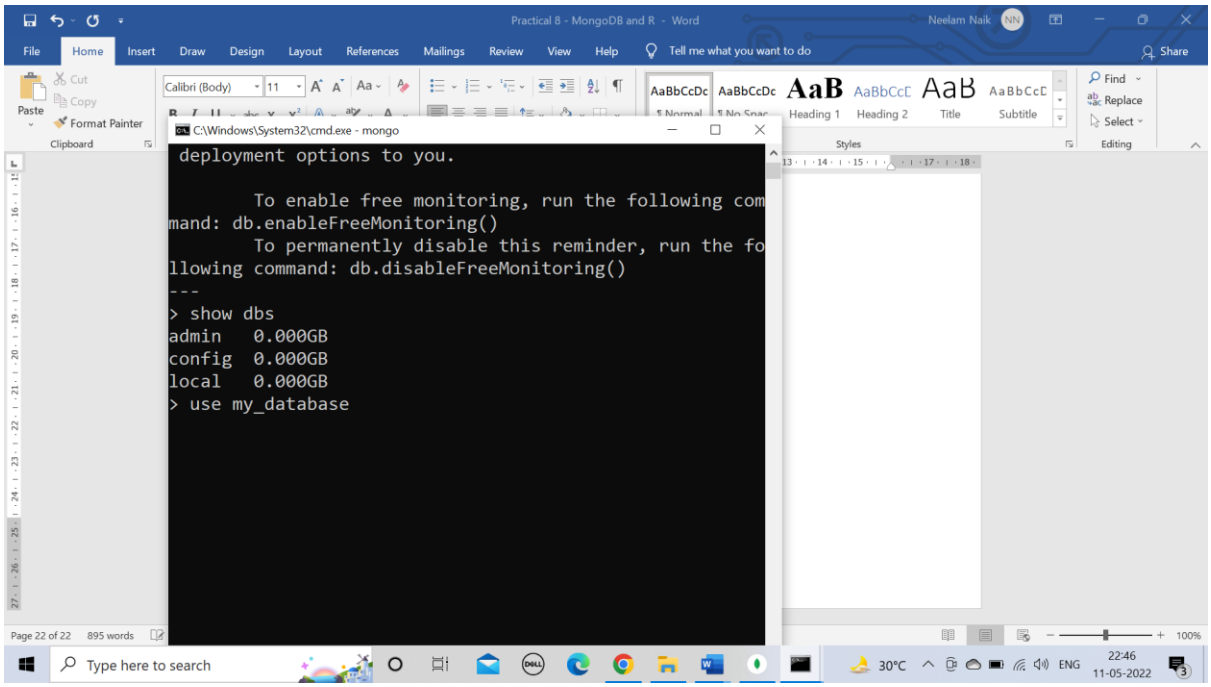
Mongo shell is started

To see all the default databases:

>show dbs

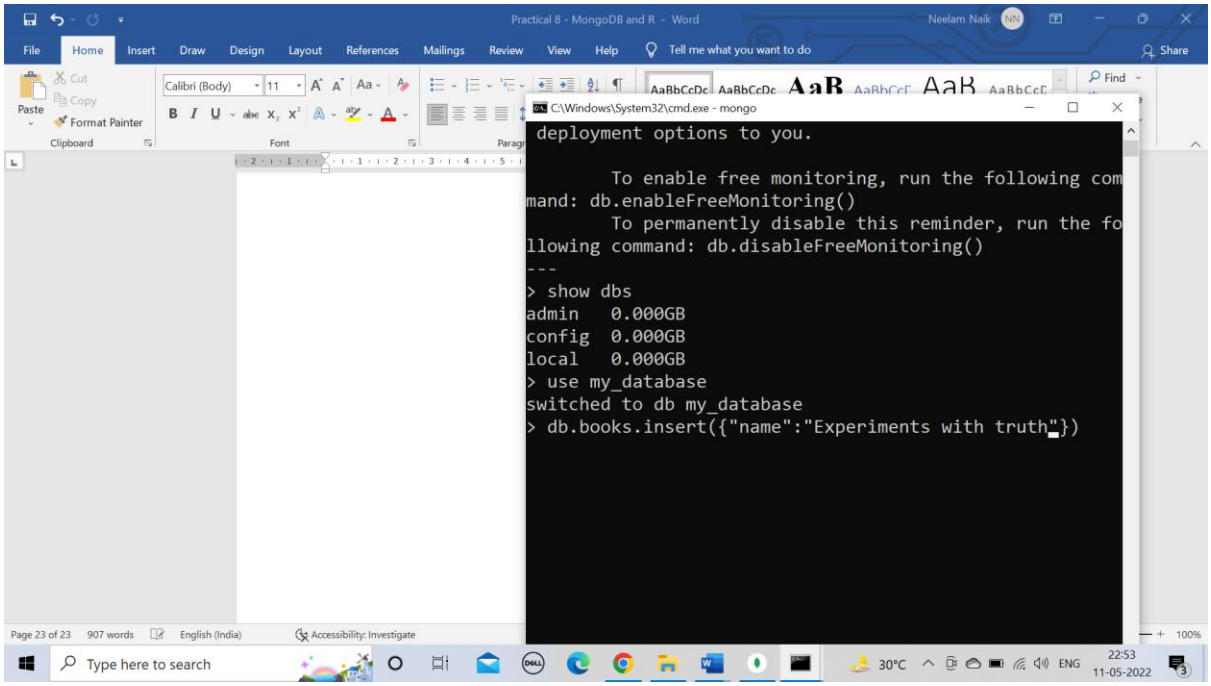


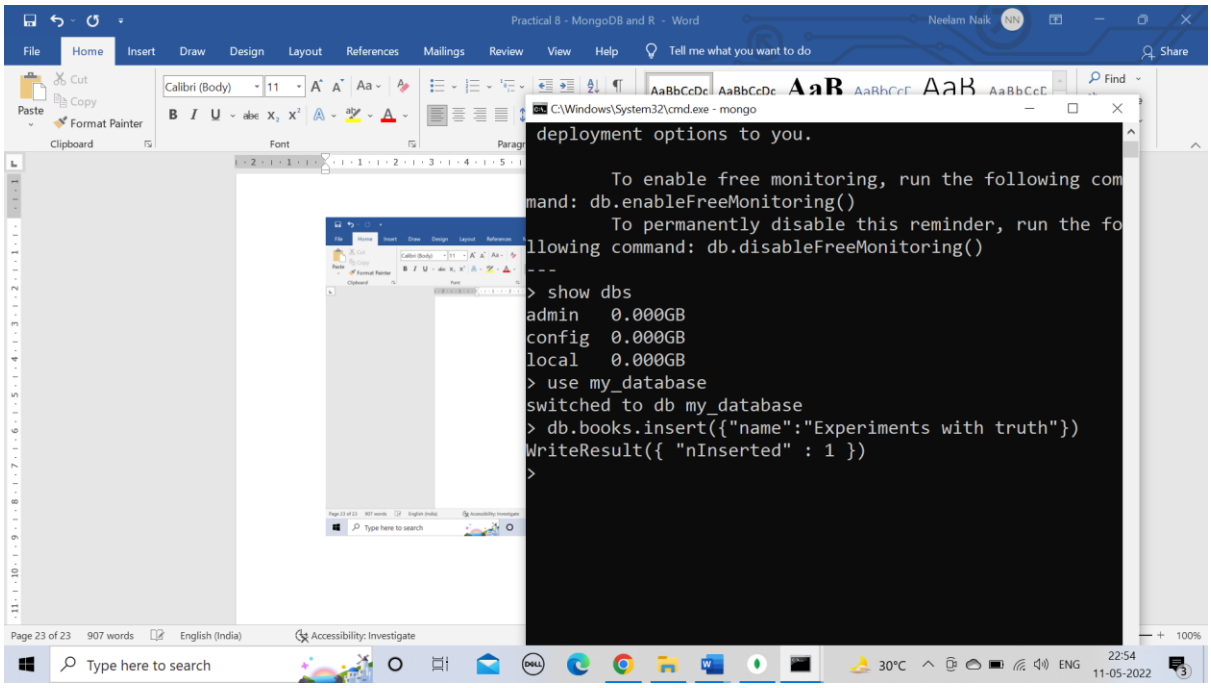
To create new database named my_database:



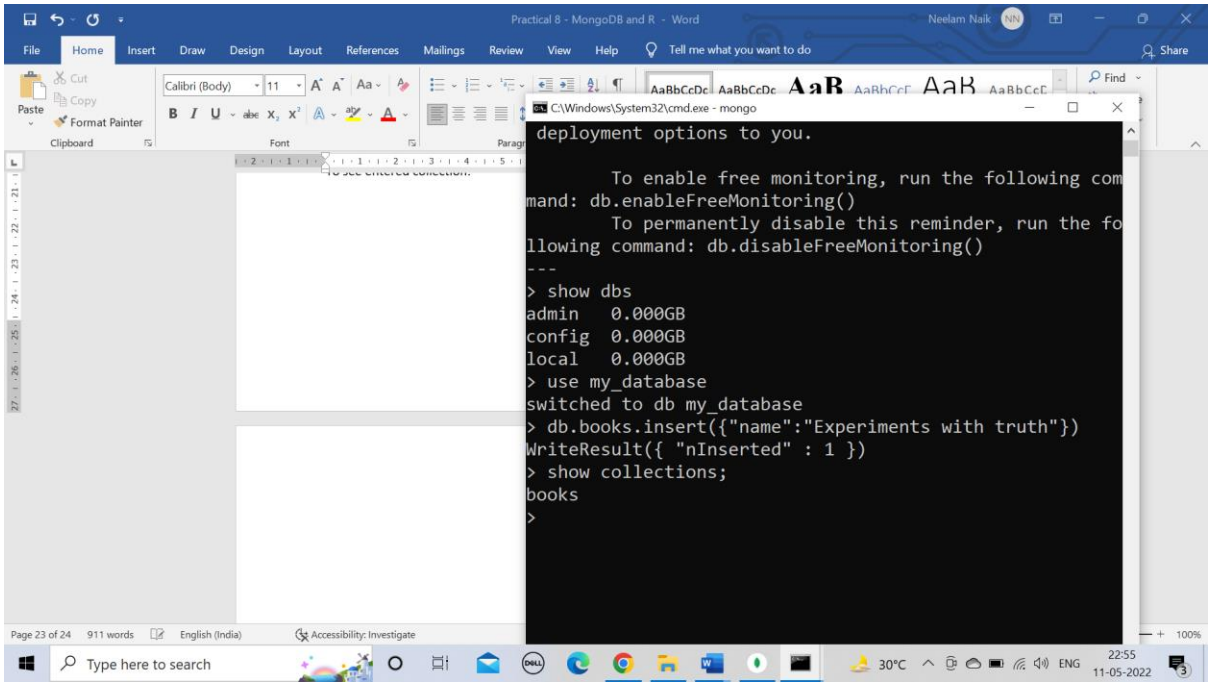
To create collection in the database:

And insert json values into it:

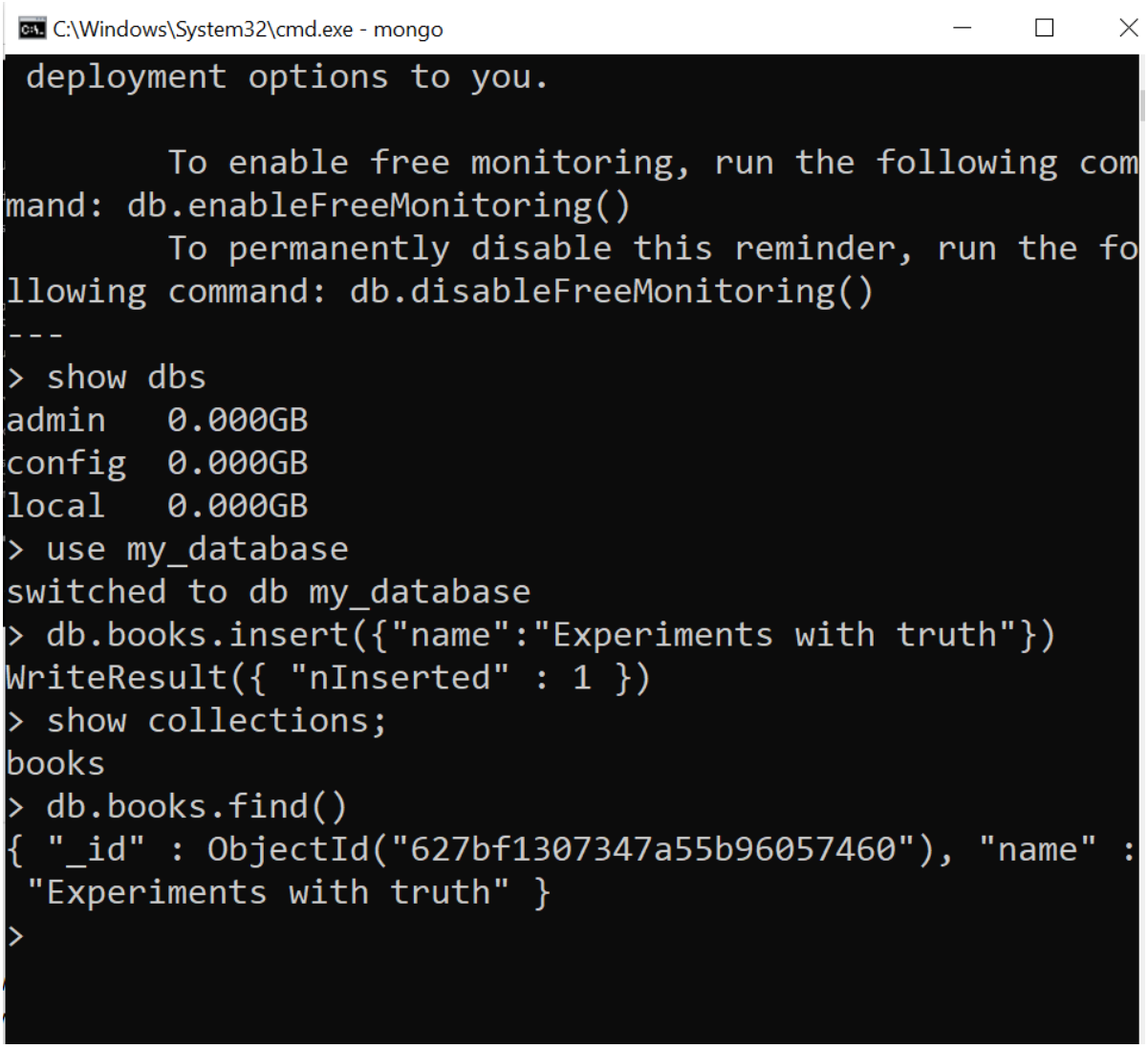




To see entered collection:



To see all the documents in the collection:

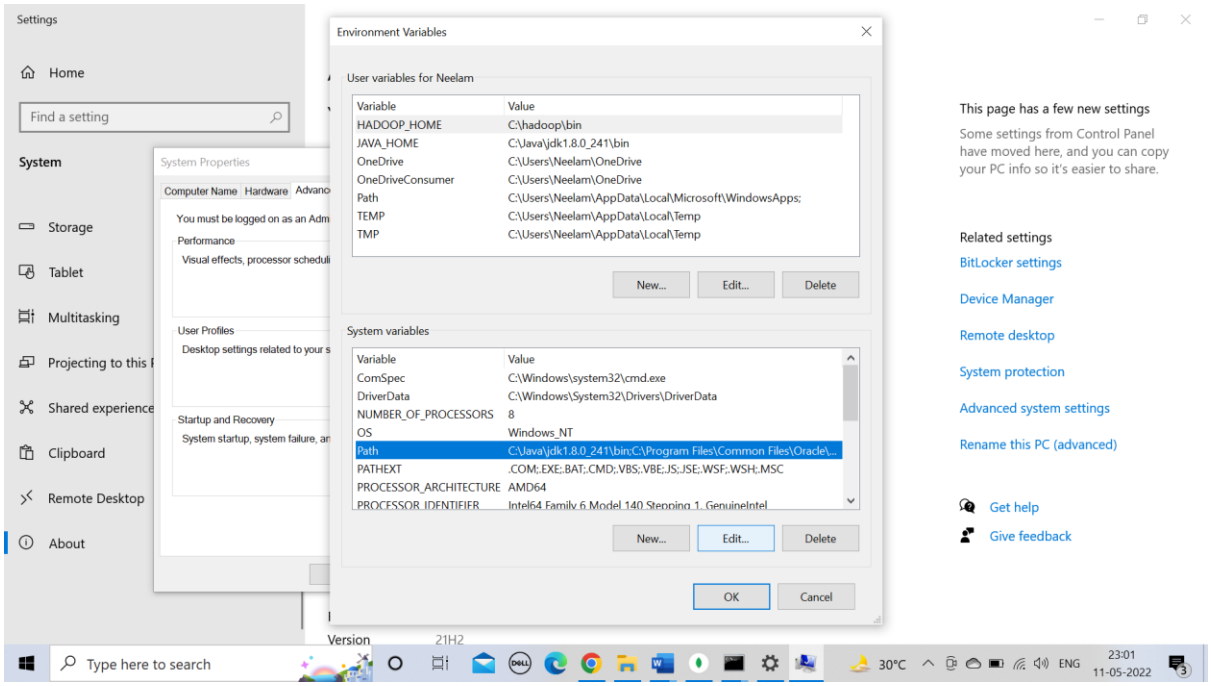


To set path of MongoDB server and shell

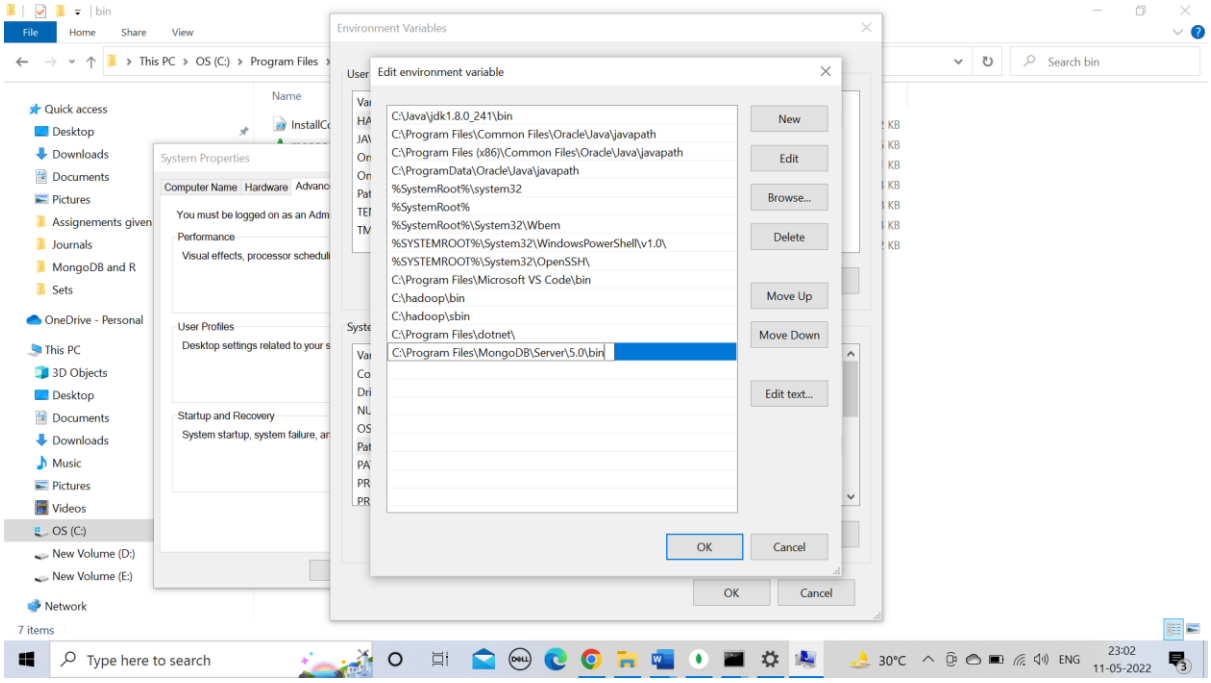
Copy in clipboard: C:\Program Files\MongoDB\Server\5.0\bin

Go to environment variables

Add mongodb path to system path variable



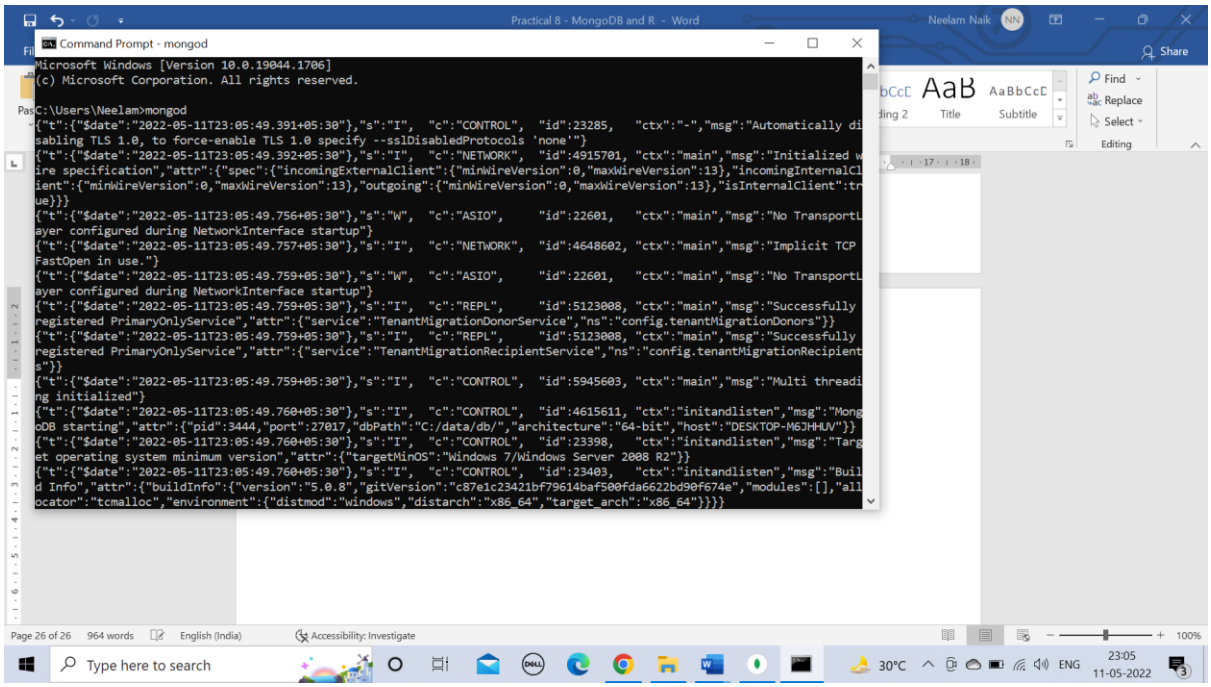
Click on New

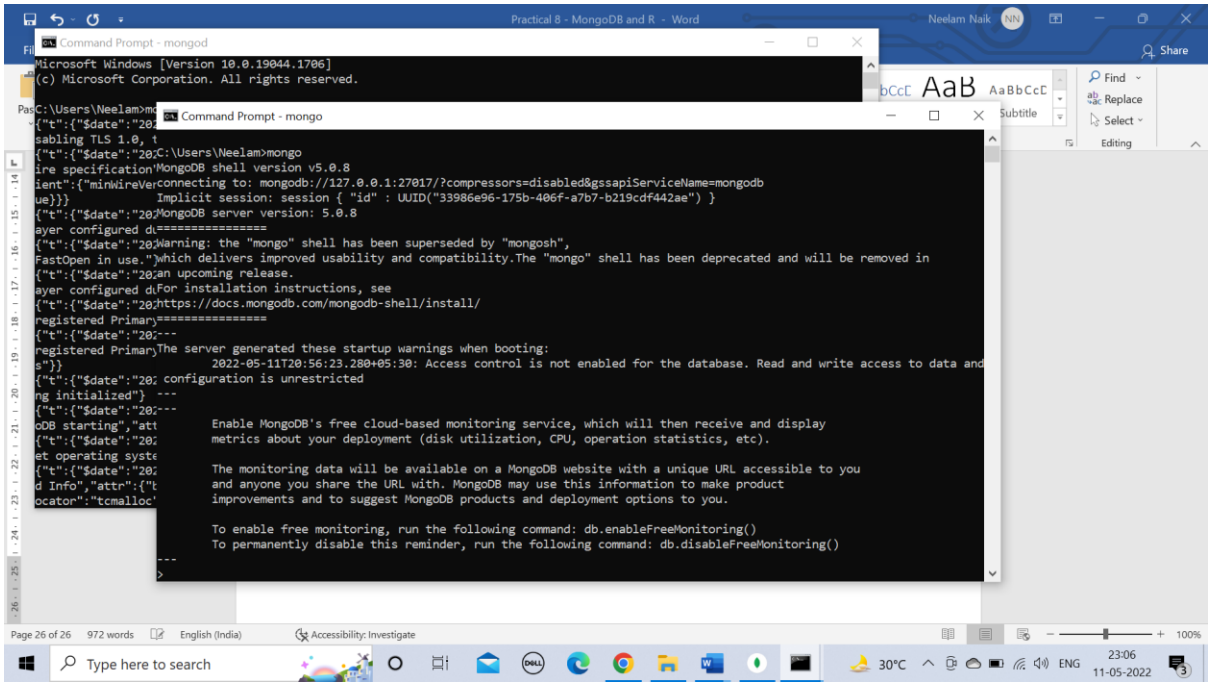


Paste the path

Click on ok..ok

Run mongod and mongo command from command prompt once again from anywhere, it will start mongo server and mongo shell





To get the effect of running mongod as service, Restart your windows operating system.

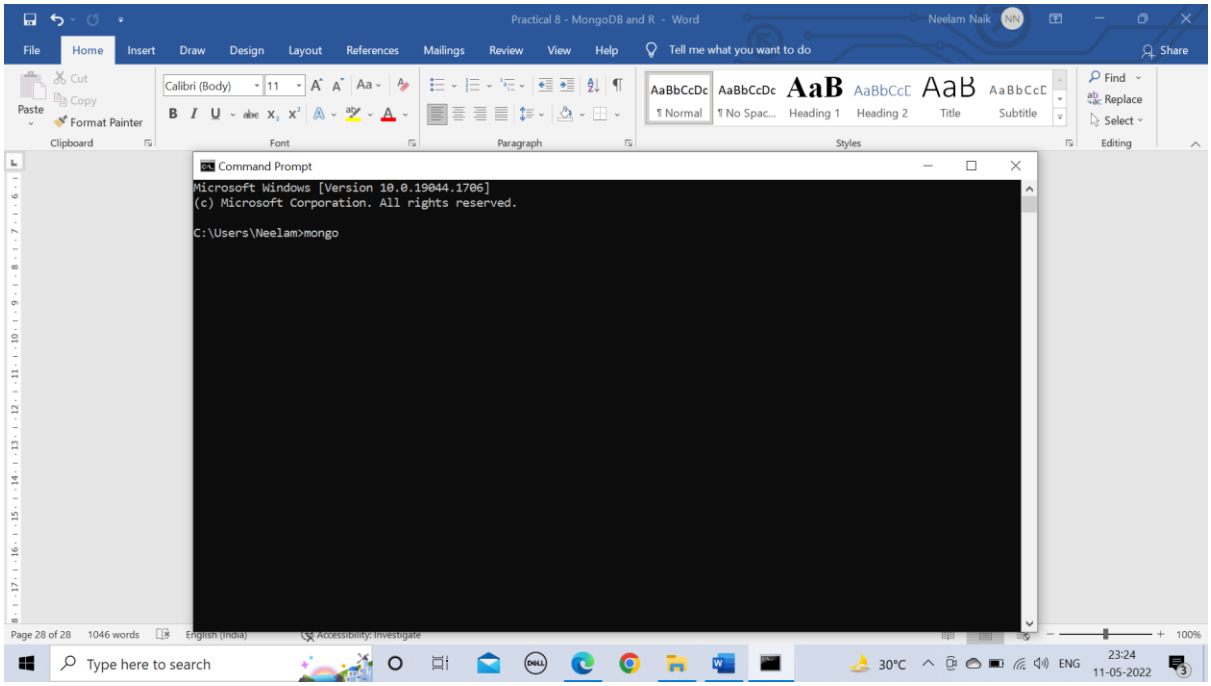
Restarted the machine

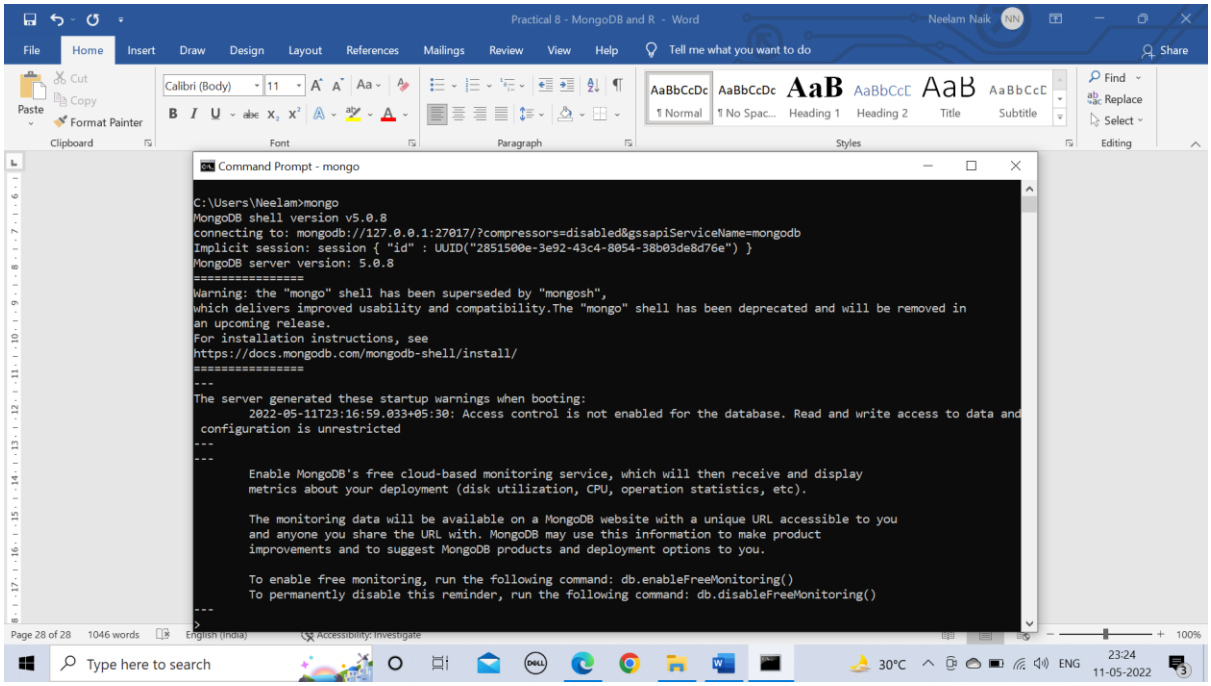
To check mongod service is running automatically:

Open command prompt

Give mongo command without running mongod command in another terminal.

It will start mongod server.





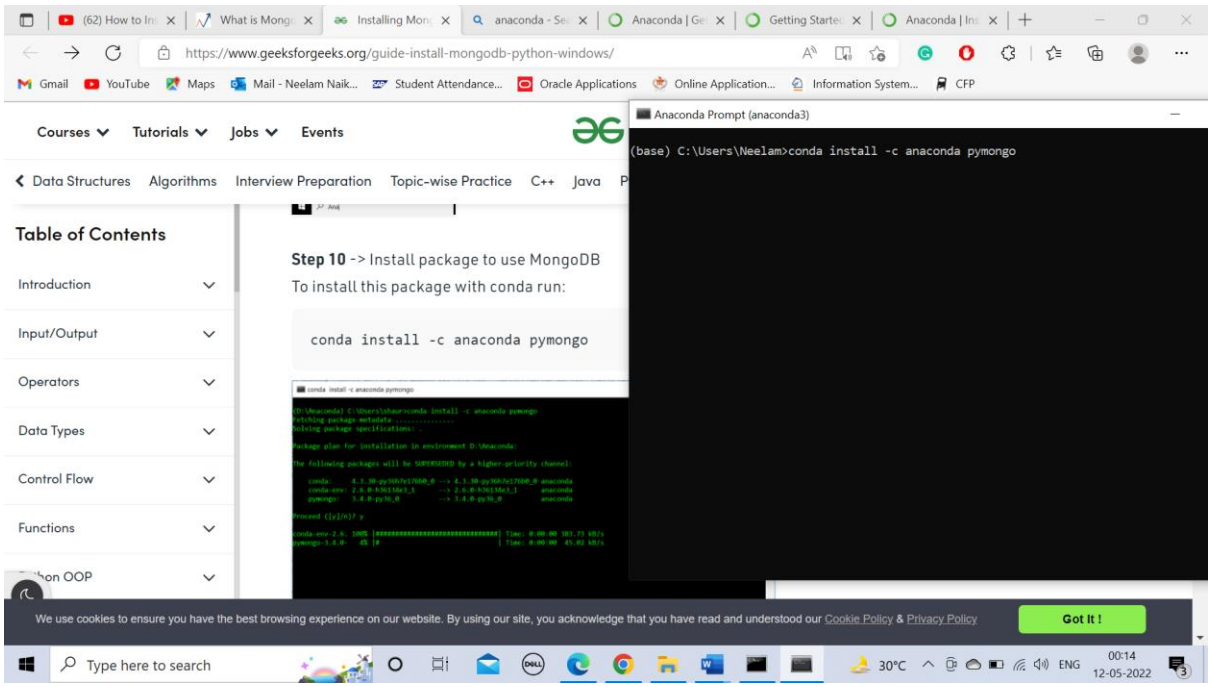
Step 4: Install MongoDB Python on Windows

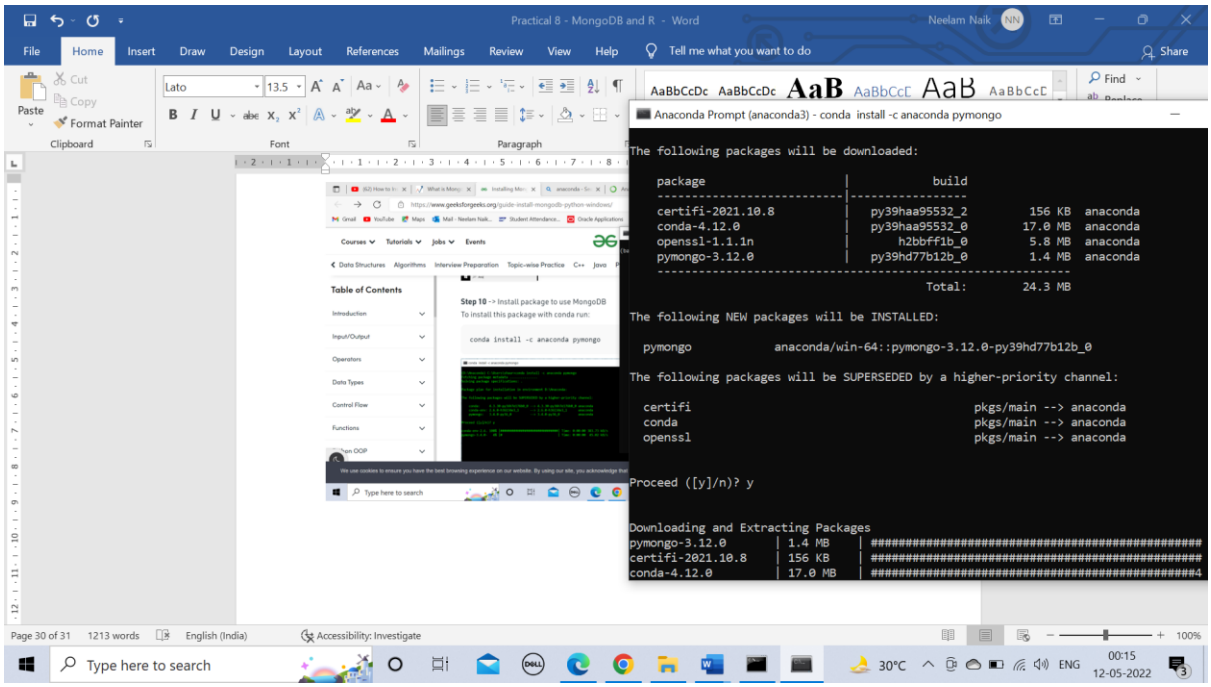
We will be performing a few key basic operations on a MongoDB database in Python using the PyMongo library.

Install package to use MongoDB

To install this package with conda run:

conda install -c anaconda pymongo





Step 5: Verify MongoDB Python Connection

To retrieve the data from a MongoDB database, we will first connect to it. Write and execute the below code in your spider anaconda

```
import pymongo
mongo_uri = "mongodb://localhost:27017/"
client = pymongo.MongoClient(mongo_uri)
```

Let’s see the available databases:

```
print(client.list_database_names())
```

We will use the my_database database for our purpose. Let’s set the cursor to the same database:

```
db = client.my_database
```

connect to analysis database

The list_collection_names command shows the names of all the available collections:

```
print(db.list_collection_names())
```

Let’s see the number of books we have. We will connect to the customers collection and then print the number of documents available in that collection:

```
table=db.books
print(table.count_documents({}) )    #gives the number of documents in the table
```

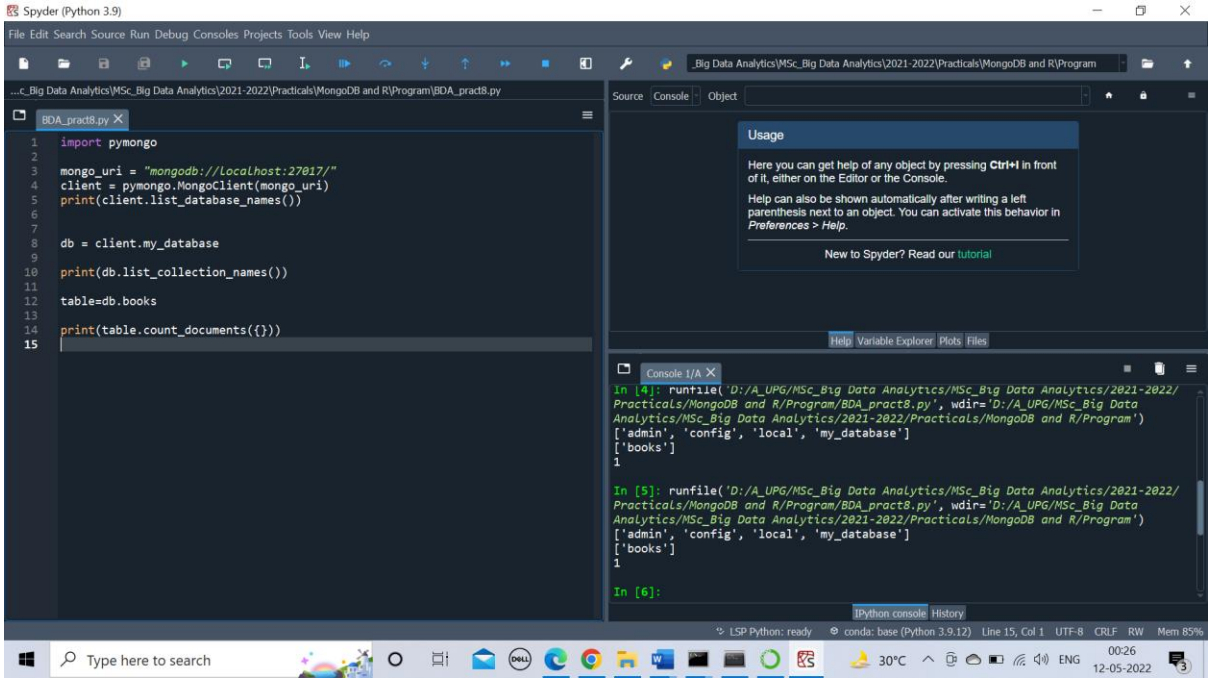
Code:

```
import pymongo

mongo_uri = "mongodb://localhost:27017/"
client = pymongo.MongoClient(mongo_uri)
print(client.list_database_names())
db = client.my_database
print(db.list_collection_names())
table=db.books
```

zprint(table.count_documents({}))

Output:



References:

https://www.analyticsvidhya.com/blog/2020/02/mongodb-in-python-tutorial-for-beginners-using-pymongo/

https://www.youtube.com/watch?v=FwMwO8pXfq0&t=3s&ab_channel=ProgrammingKnowledge