Dist Data Engineering

Assignment 01

Muhammad Usman Malik	19i-0402
Shaheer Asif	19i-2152
Musa Ishaq	19i-0731

Instructor: Dr.Adnan Tariq

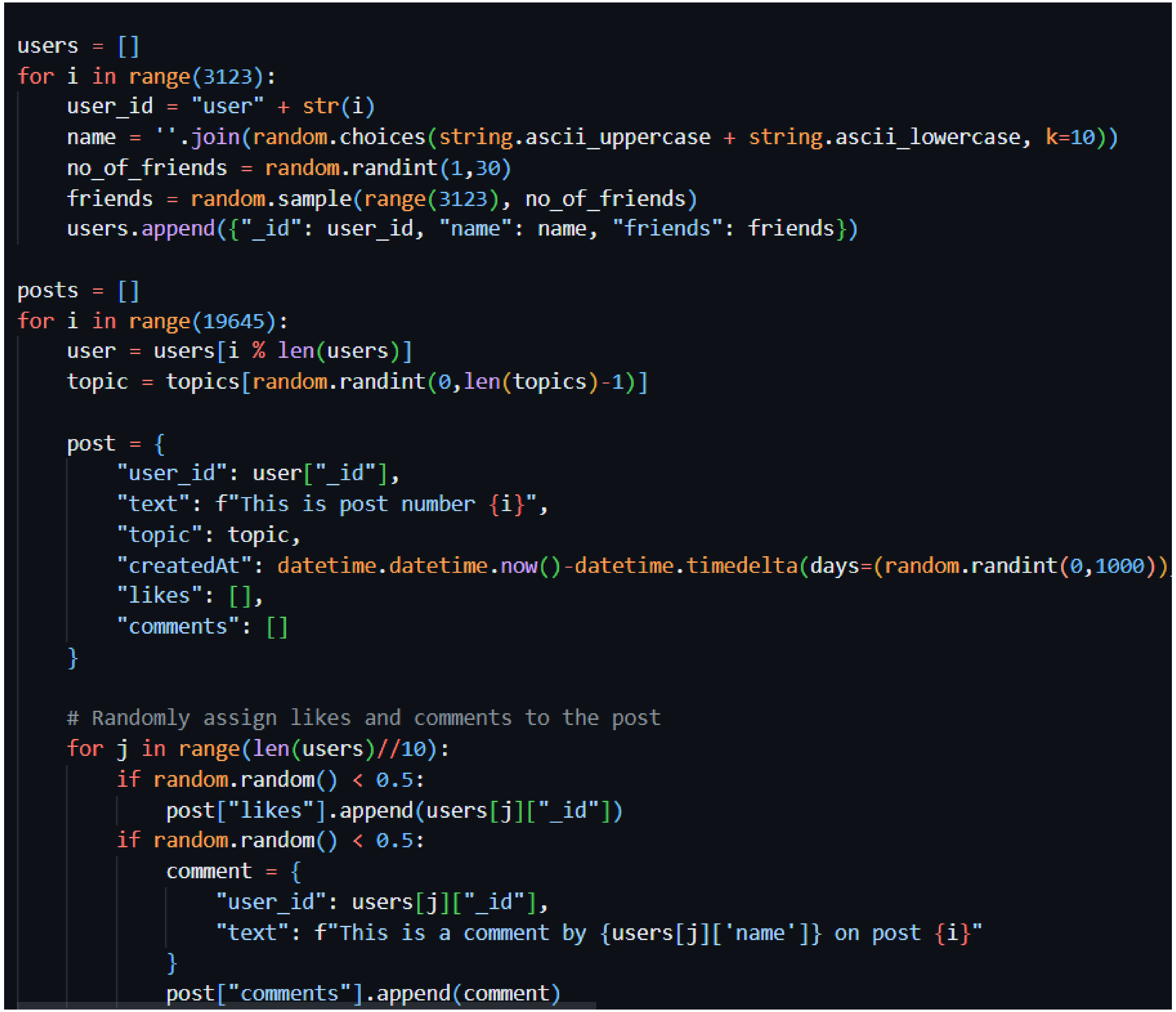
https://github.com/GhostByteX/Dist_Data_Eng_Assignment01

Database Design

- Created 3 shards via Atlas Interface
- Inserted records automatically in the shards
- Deployed using MongoDB Atlas
- M40 40GB Database
- Two collections (Users and Posts)
- Users contain userID, name, no of friends and friend id
- Posts contain userID, text, topic, date, likes and comments
- Likes contain the userID of users liking the post
- Comments contain userID and the text
- We created indexes on userID, topic, name and date

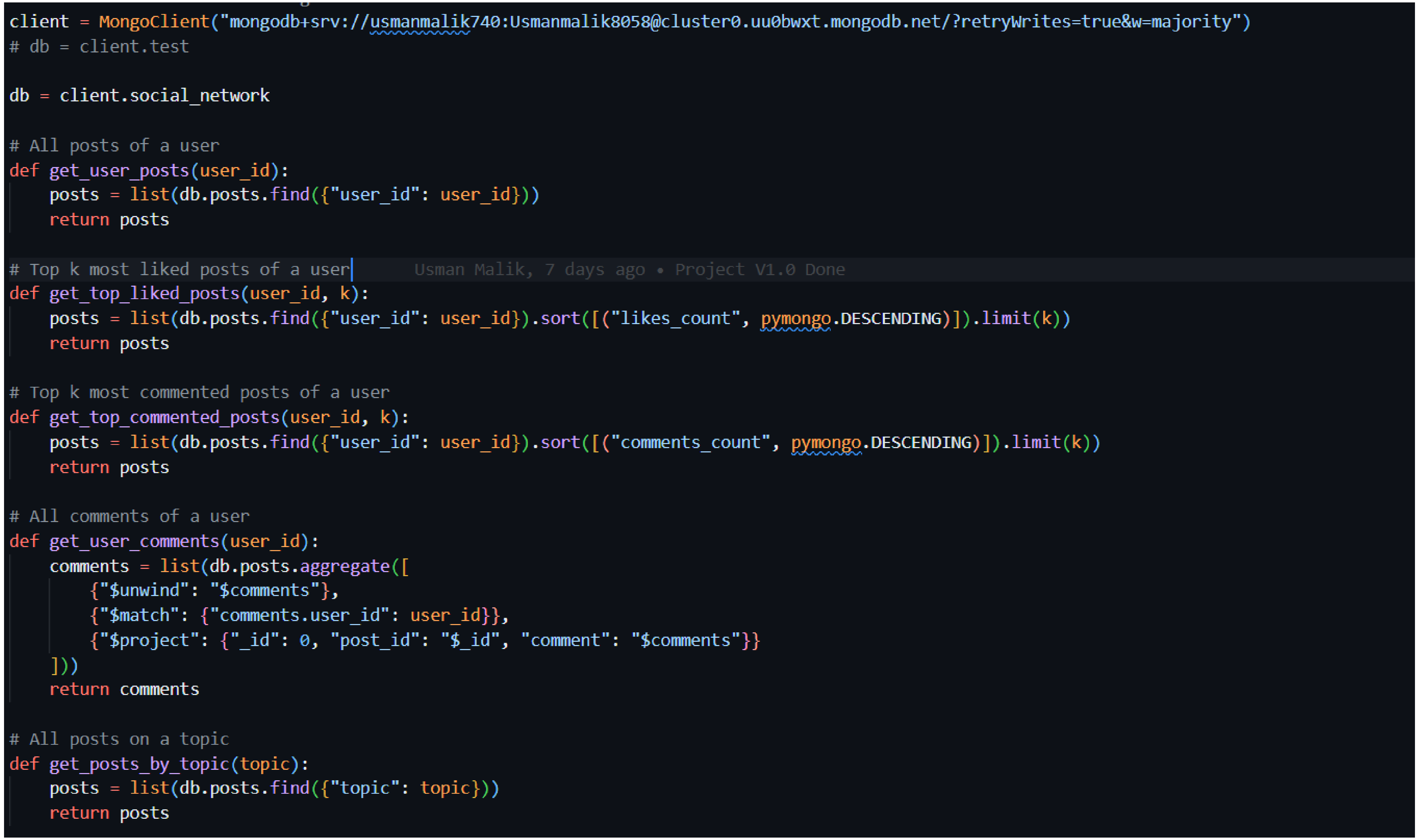
Database Design

- userID: This field is used to identify the user who created the post. If we frequently need to retrieve all posts made by a specific user, an index on this field can speed up those queries.
- topic: This field is used to store the topic(s) associated with each post. If we frequently need to retrieve all posts that belong to a specific topic, an index on this field can speed up those queries.
- date: This field is used to store the date and time when each post was created. If we frequently need to retrieve posts within a specific date range, an index on this field can speed up those queries.
- name: This field is not strictly necessary, but if we frequently need to sort posts by the name of the user who created them, an index on this field can speed up those queries.



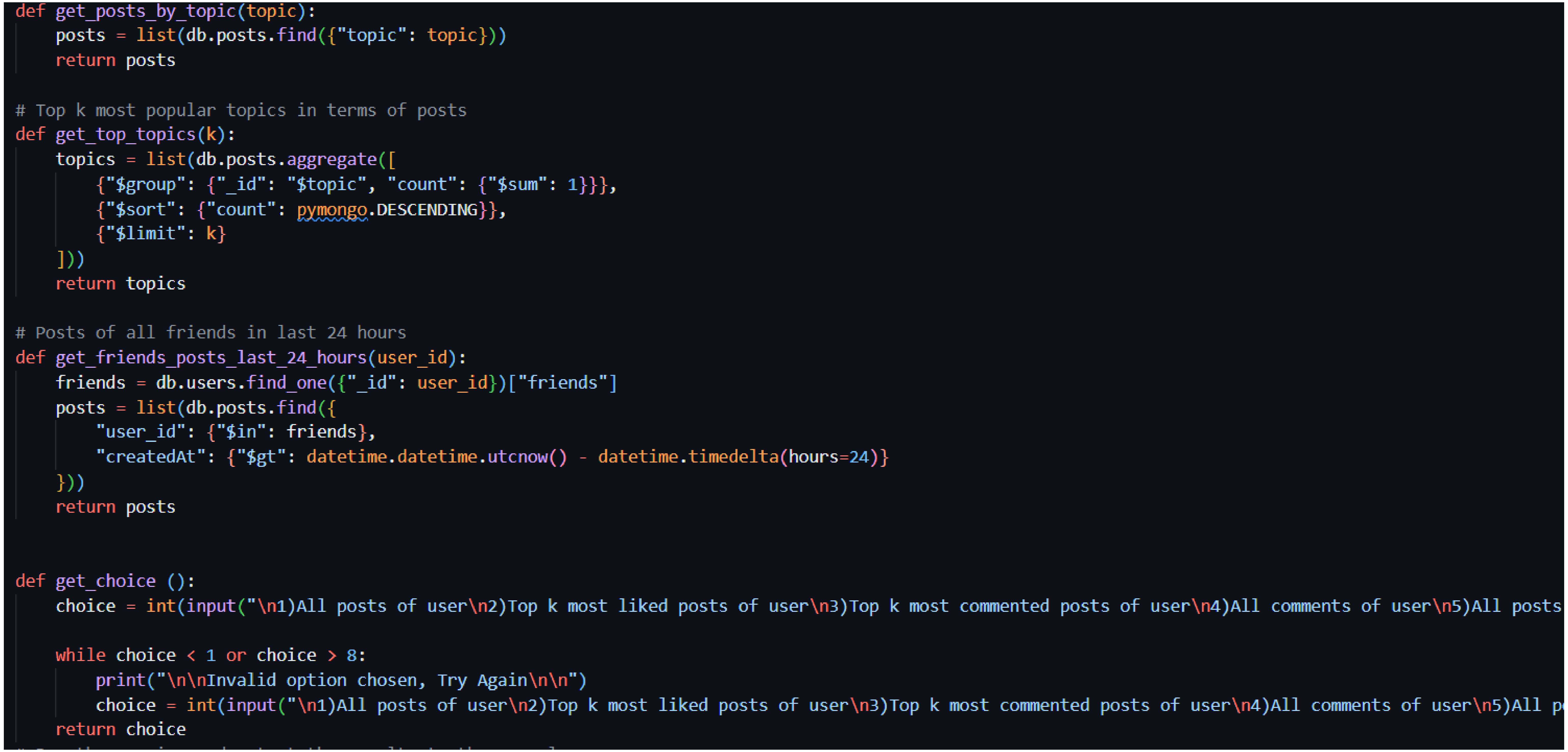
```python
users = []
for i in range(3123):
    user_id = "user" + str(i)
    name = ''.join(random.choices(string.ascii_uppercase + string.ascii_lowercase, k=10))
    no_of_friends = random.randint(1,30)
    friends = random.sample(range(3123), no_of_friends)
    users.append({"_id": user_id, "name": name, "friends": friends})

posts = []
for i in range(19645):
    user = users[i % len(users)]
    topic = topics[random.randint(0,len(topics)-1)]

    post = {
        "user_id": user["_id"],
        "text": f"This is post number {i}",
        "topic": topic,
        "createdAt": datetime.datetime.now()-datetime.timedelta(days=(random.randint(0,1000))
        "likes": [],
        "comments": []
    }

    # Randomly assign likes and comments to the post
    for j in range(len(users)//10):
        if random.random() < 0.5:
            post["likes"].append(users[j]["_id"])
        if random.random() < 0.5:
            comment = {
                "user_id": users[j]["_id"],
                "text": f"This is a comment by {users[j]['name']} on post {i}"
            }
            post["comments"].append(comment)
```

Queries



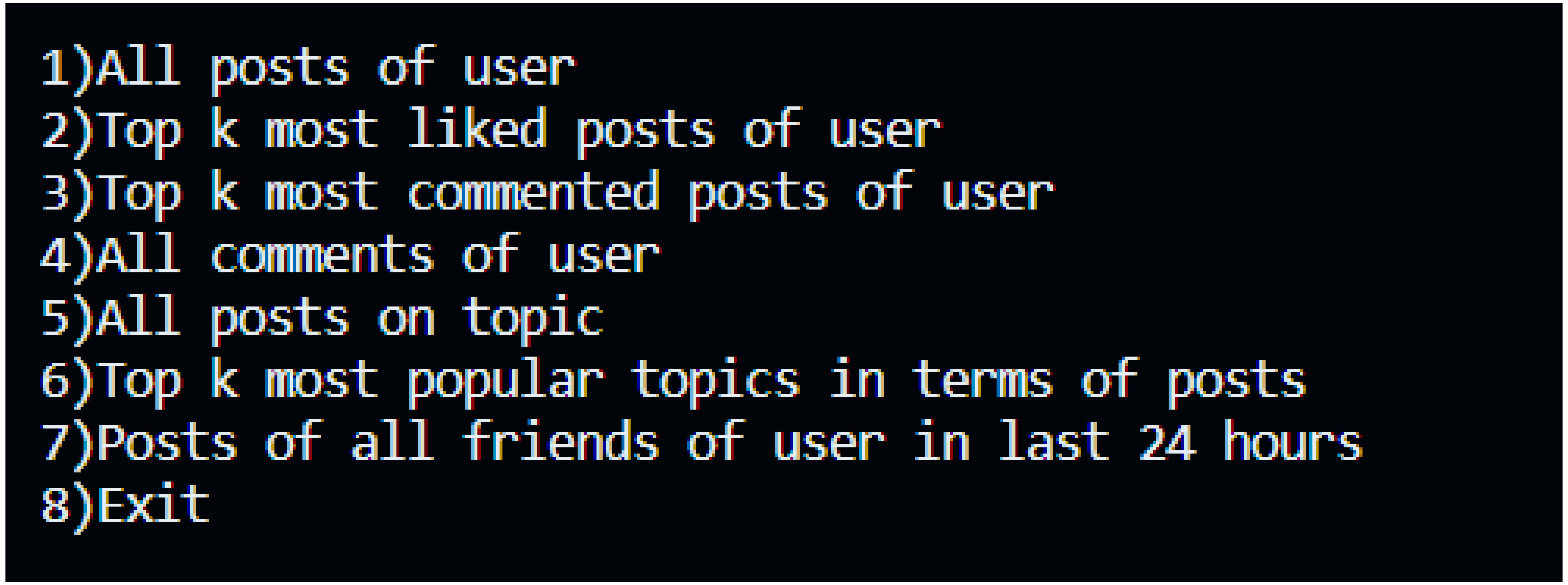
```
client = MongoClient("mongodb+srv://usmanmalik740:Usmanmalik8058@cluster0.uu0bwxt.mongodb.net/?retryWrites=true&w=majority")
# db = client.test

db = client.social_network

# All posts of a user
def get_user_posts(user_id):
    posts = list(db.posts.find({"user_id": user_id}))
    return posts

# Top k most liked posts of a user      Usman Malik, 7 days ago • Project V1.0 Done
def get_top_liked_posts(user_id, k):
    posts = list(db.posts.find({"user_id": user_id}).sort([("likes_count", pymongo.DESCENDING)]).limit(k))
    return posts

# Top k most commented posts of a user
def get_top_commented_posts(user_id, k):
    posts = list(db.posts.find({"user_id": user_id}).sort([("comments_count", pymongo.DESCENDING)]).limit(k))
    return posts

# All comments of a user
def get_user_comments(user_id):
    comments = list(db.posts.aggregate([
        {"$unwind": "$comments"},
        {"$match": {"comments.user_id": user_id}},
        {"$project": {"_id": 0, "post_id": "$_id", "comment": "$comments"}}
    ]))
    return comments

# All posts on a topic
def get_posts_by_topic(topic):
    posts = list(db.posts.find({"topic": topic}))
    return posts
```

Queries

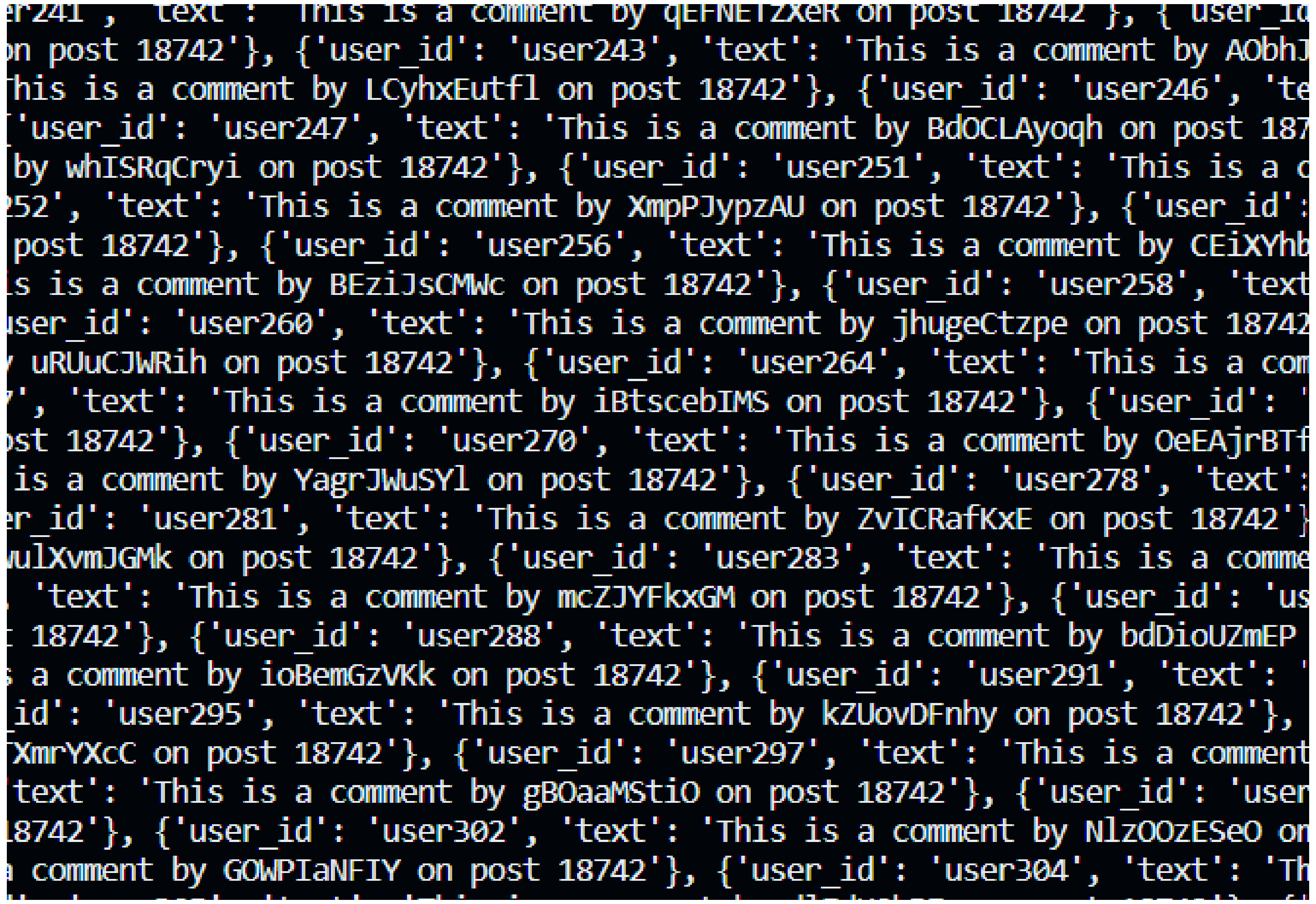


```python
def get_posts_by_topic(topic):
    posts = list(db.posts.find({"topic": topic}))
    return posts

# Top k most popular topics in terms of posts
def get_top_topics(k):
    topics = list(db.posts.aggregate([
        {"$group": {"_id": "$topic", "count": {"$sum": 1}}},
        {"$sort": {"count": pymongo.DESCENDING}},
        {"$limit": k}
    ]))
    return topics

# Posts of all friends in last 24 hours
def get_friends_posts_last_24_hours(user_id):
    friends = db.users.find_one({"_id": user_id})["friends"]
    posts = list(db.posts.find({
        "user_id": {"$in": friends},
        "createdAt": {"$gt": datetime.datetime.utcnow() - datetime.timedelta(hours=24)}
    }))
    return posts


def get_choice ():
    choice = int(input("\n1)All posts of user\n2)Top k most liked posts of user\n3)Top k most commented posts of user\n4)All comments of user\n5)All posts

    while choice < 1 or choice > 8:
        print("\n\nInvalid option chosen, Try Again\n\n")
        choice = int(input("\n1)All posts of user\n2)Top k most liked posts of user\n3)Top k most commented posts of user\n4)All comments of user\n5)All p
    return choice
```

Results



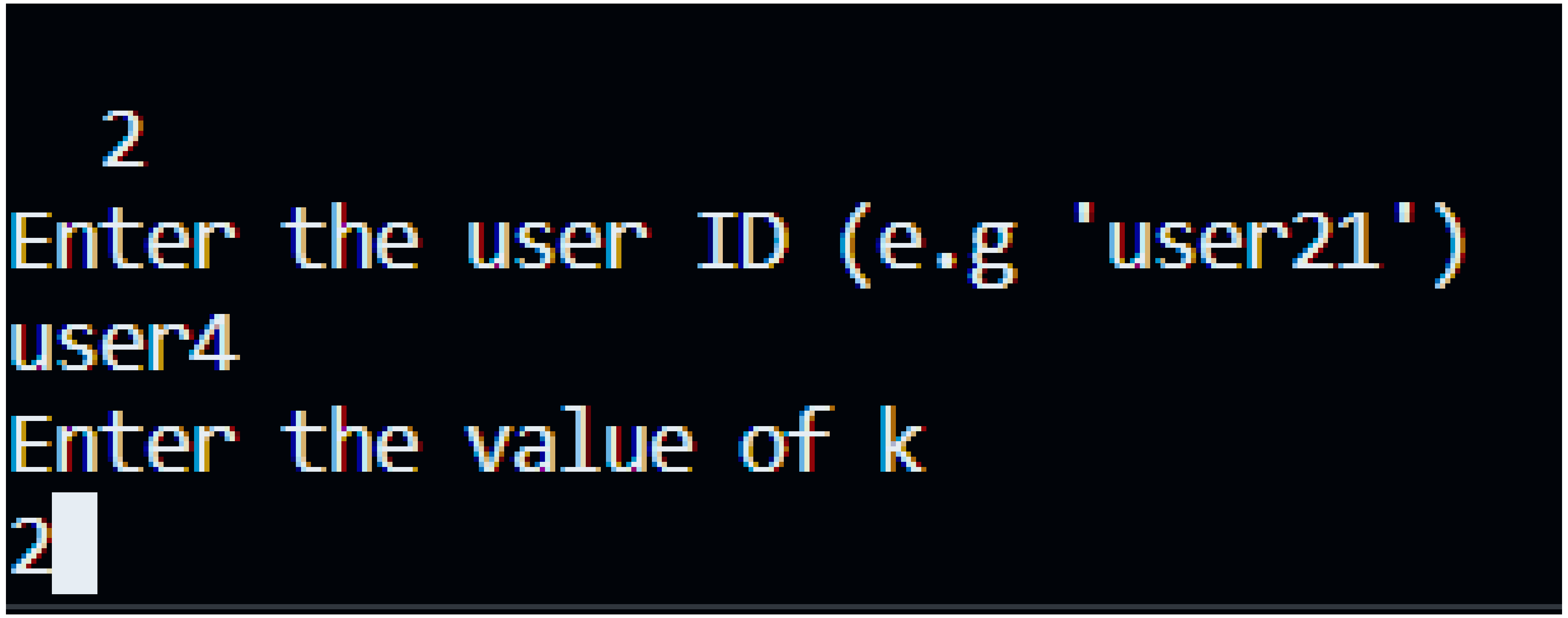
```
1)All posts of user
2)Top k most liked posts of user
3)Top k most commented posts of user
4)All comments of user
5)All posts on topic
6)Top k most popular topics in terms of posts
7)Posts of all friends of user in last 24 hours
8)Exit
```



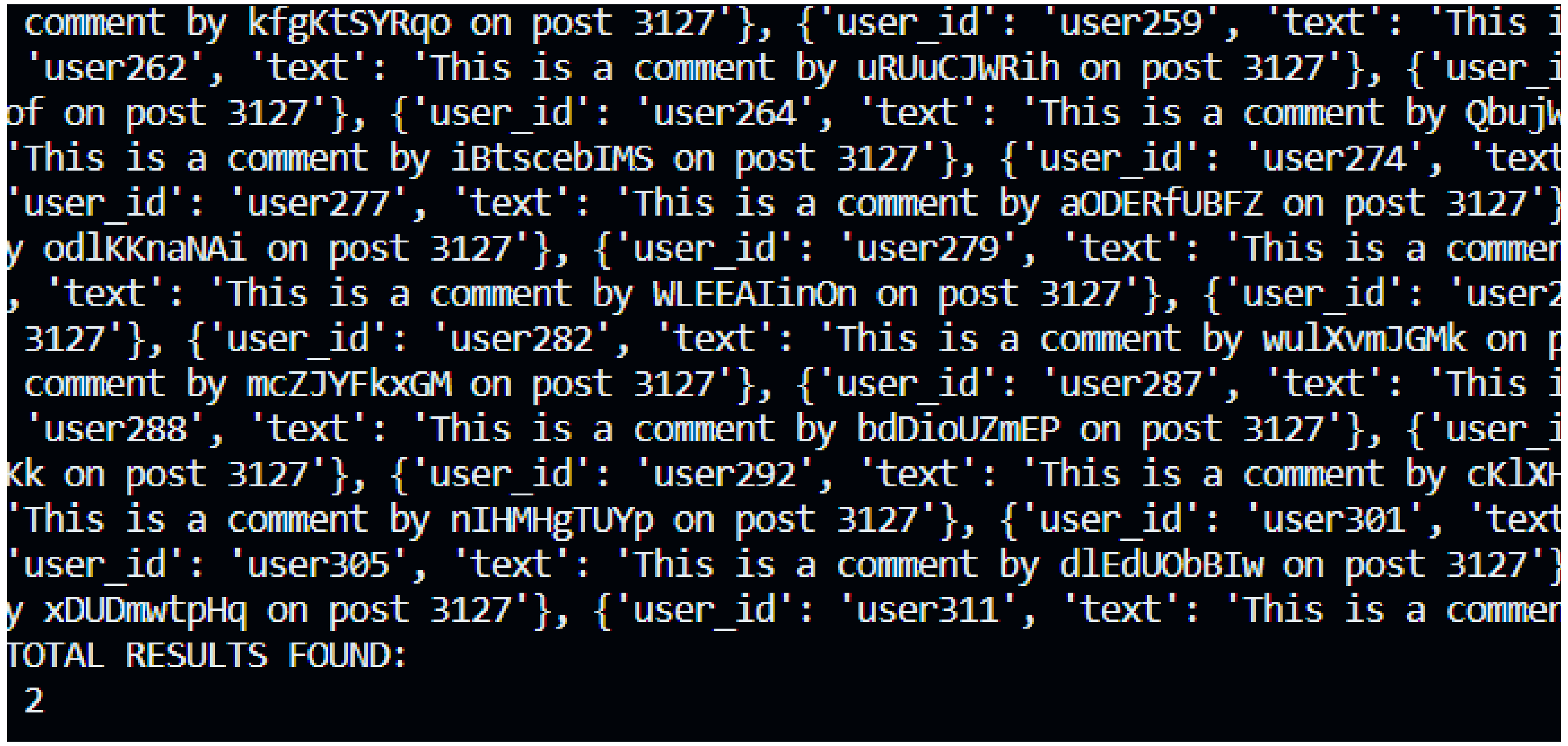
```
er241', 'text': 'This is a comment by qEFNETZXeR on post 18742'}, {'user_id
on post 18742'}, {'user_id': 'user243', 'text': 'This is a comment by AObhJ
This is a comment by LCyhxEutfl on post 18742'}, {'user_id': 'user246', 'te
{'user_id': 'user247', 'text': 'This is a comment by BdOCLAyoqh on post 187
 by whISRqCryi on post 18742'}, {'user_id': 'user251', 'text': 'This is a c
252', 'text': 'This is a comment by XmpPJypzAU on post 18742'}, {'user_id':
 post 18742'}, {'user_id': 'user256', 'text': 'This is a comment by CEiXYhb
is is a comment by BEziJsCMWc on post 18742'}, {'user_id': 'user258', 'text
user_id': 'user260', 'text': 'This is a comment by jhugeCtzpe on post 18742
y uRUuCJWRih on post 18742'}, {'user_id': 'user264', 'text': 'This is a com
', 'text': 'This is a comment by iBtscebIMS on post 18742'}, {'user_id': '
ost 18742'}, {'user_id': 'user270', 'text': 'This is a comment by OeEAjrBTf
 is a comment by YagrJWuSYl on post 18742'}, {'user_id': 'user278', 'text':
er_id': 'user281', 'text': 'This is a comment by ZvICRafKxE on post 18742'}
wulXvmJGMk on post 18742'}, {'user_id': 'user283', 'text': 'This is a comme
, 'text': 'This is a comment by mcZJYFkxGM on post 18742'}, {'user_id': 'us
t 18742'}, {'user_id': 'user288', 'text': 'This is a comment by bdDioUZmEP
s a comment by ioBemGzVKk on post 18742'}, {'user_id': 'user291', 'text': '
_id': 'user295', 'text': 'This is a comment by kZUovDFnhy on post 18742'},
TXmrYXcC on post 18742'}, {'user_id': 'user297', 'text': 'This is a comment
'text': 'This is a comment by gBOaaMStiO on post 18742'}, {'user_id': 'user
18742'}, {'user_id': 'user302', 'text': 'This is a comment by NlzOOzESeO on
a comment by GOWPIaNFIY on post 18742'}, {'user_id': 'user304', 'text': 'Th
```

- All queries are executed via this menu
- All running successfully
- The name of a user is a random string that is reflected in the results below

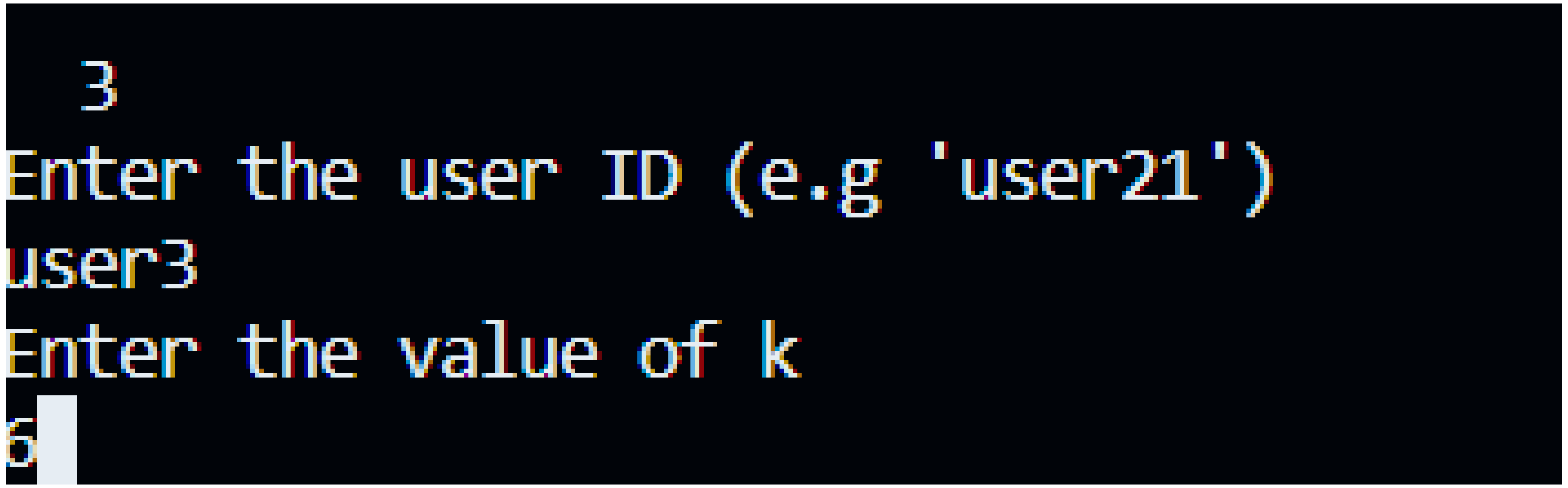
Results



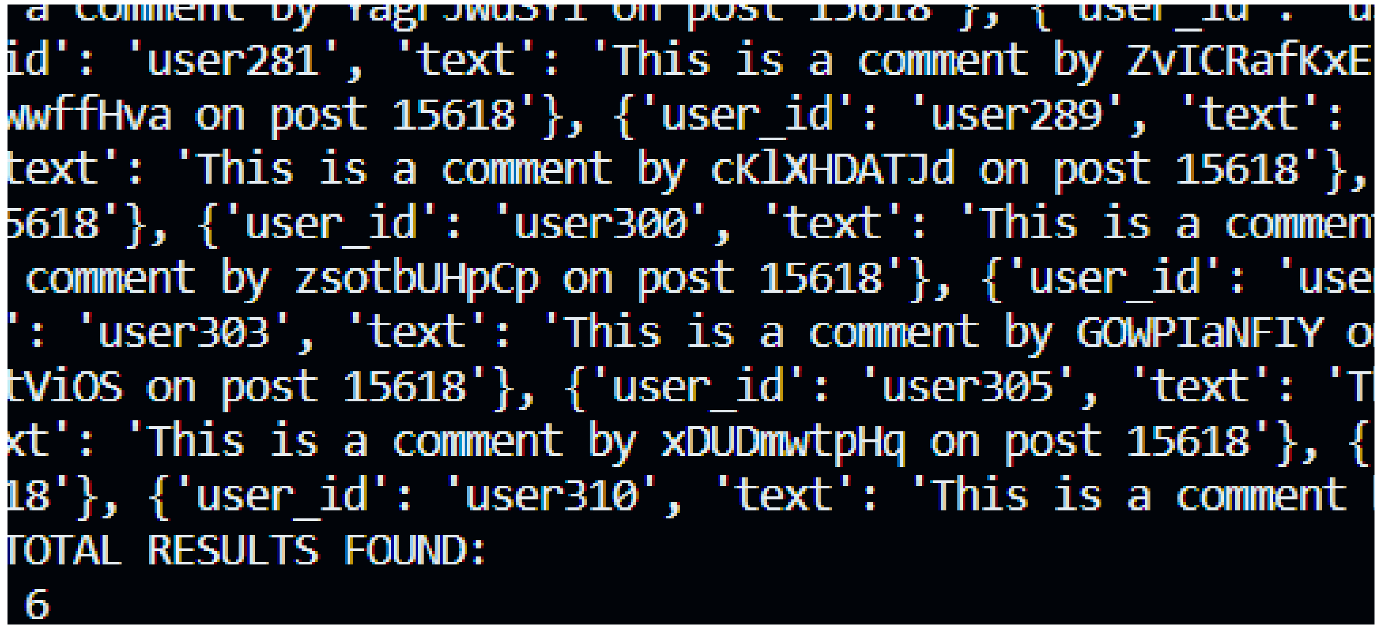
```
 2
Enter the user ID (e.g 'user21')
user4
Enter the value of k
2
```



```
 comment by kfgKtSYRqo on post 3127'}, {'user_id': 'user259', 'text': 'This i
 'user262', 'text': 'This is a comment by uRUuCJWRih on post 3127'}, {'user_i
of on post 3127'}, {'user_id': 'user264', 'text': 'This is a comment by QbujW
'This is a comment by iBtscebIMS on post 3127'}, {'user_id': 'user274', 'text
'user_id': 'user277', 'text': 'This is a comment by aODERfUBFZ on post 3127'}
y odlKKnaNAi on post 3127'}, {'user_id': 'user279', 'text': 'This is a commen
, 'text': 'This is a comment by WLEEAIinOn on post 3127'}, {'user_id': 'user2
 3127'}, {'user_id': 'user282', 'text': 'This is a comment by wulXvmJGMk on p
 comment by mcZJYFkxGM on post 3127'}, {'user_id': 'user287', 'text': 'This i
 'user288', 'text': 'This is a comment by bdDioUZmEP on post 3127'}, {'user_i
Kk on post 3127'}, {'user_id': 'user292', 'text': 'This is a comment by cKlXH
'This is a comment by nIHMHgTUYp on post 3127'}, {'user_id': 'user301', 'text
'user_id': 'user305', 'text': 'This is a comment by dlEdUObBIw on post 3127'}
y xDUDmwtpHq on post 3127'}, {'user_id': 'user311', 'text': 'This is a commen
TOTAL RESULTS FOUND:
 2
```

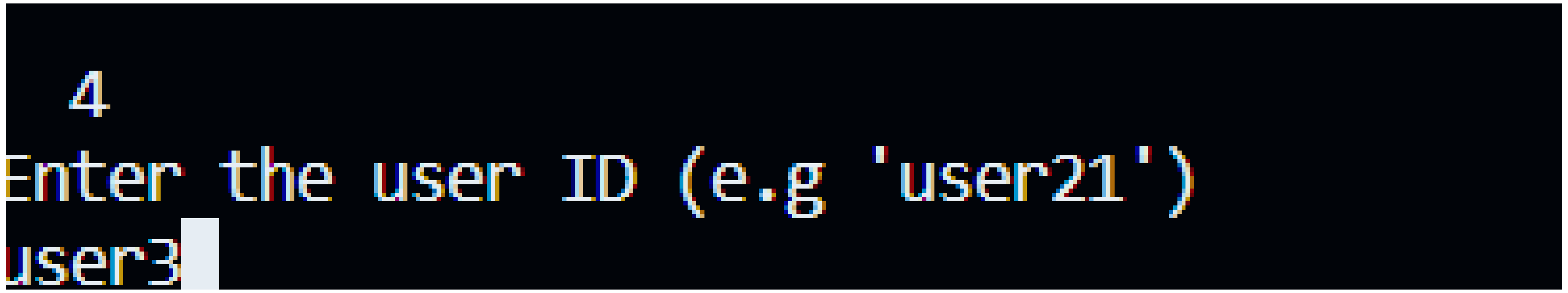


```
 3
Enter the user ID (e.g 'user21')
user3
Enter the value of k
6
```

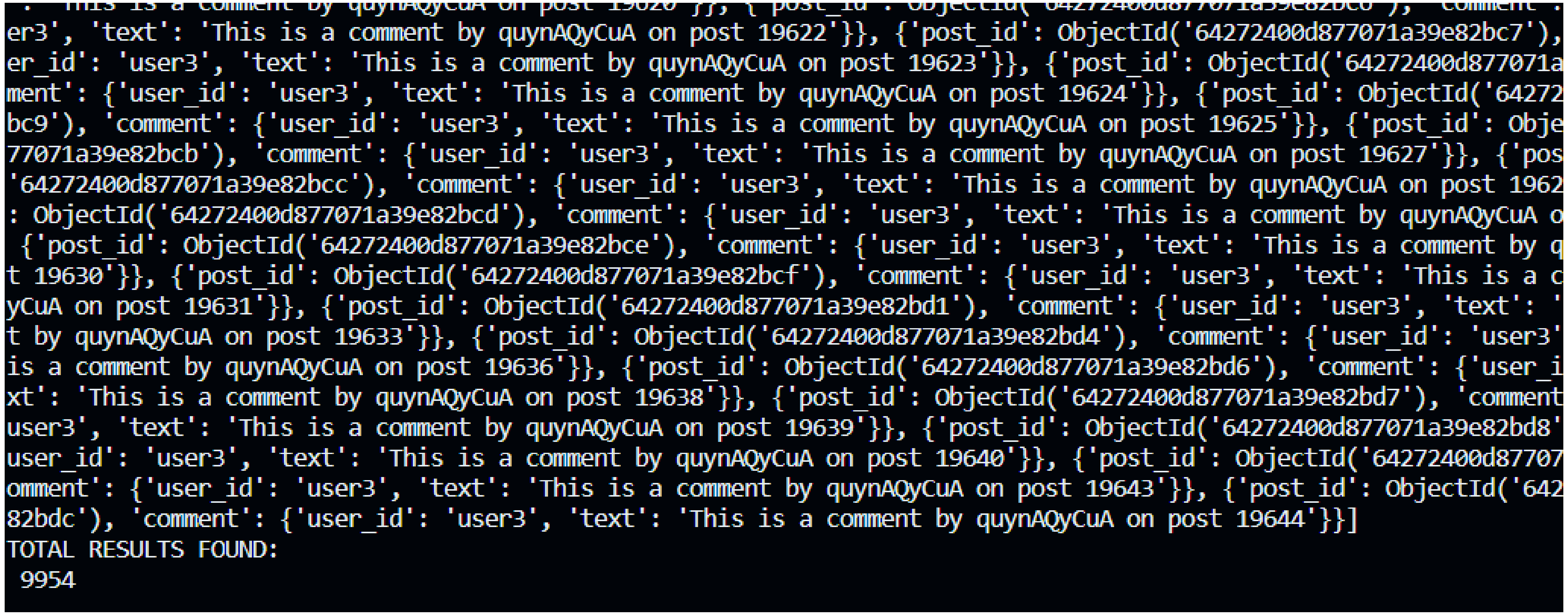


```
 a comment by YagrJWdSYI on post 15618'}, {'user_id': 'u
id': 'user281', 'text': 'This is a comment by ZvICRafKxE
wwffHva on post 15618'}, {'user_id': 'user289', 'text':
text': 'This is a comment by cKlXHDATJd on post 15618'},
5618'}, {'user_id': 'user300', 'text': 'This is a commen
 comment by zsotbUHpCp on post 15618'}, {'user_id': 'use
': 'user303', 'text': 'This is a comment by GOWPIaNFIY o
tViOS on post 15618'}, {'user_id': 'user305', 'text': 'T
xt': 'This is a comment by xDUDmwtpHq on post 15618'}, {
18'}, {'user_id': 'user310', 'text': 'This is a comment
TOTAL RESULTS FOUND:
 6
```

Results

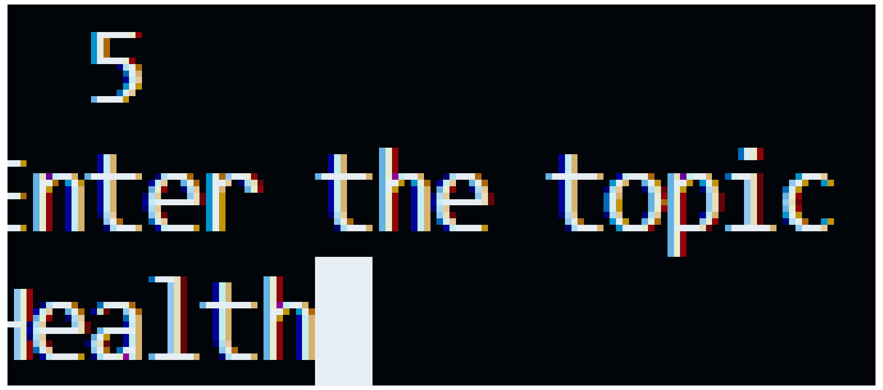


```
 4
Enter the user ID (e.g 'user21')
user3
```

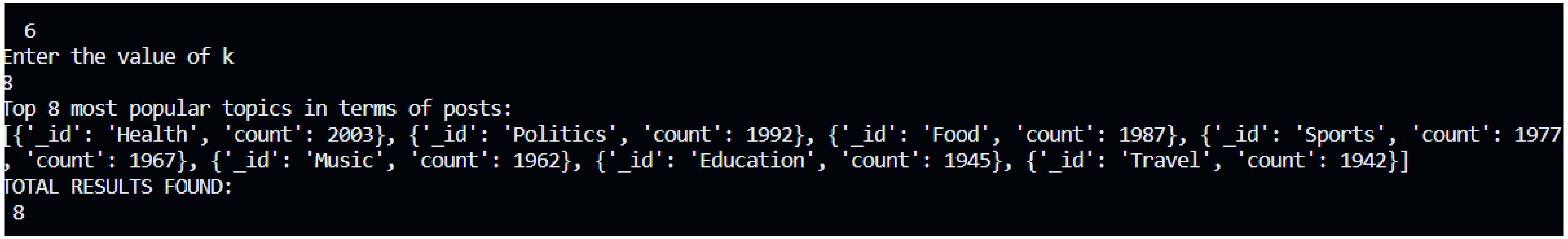


```
: 'This is a comment by quynAQyCuA on post 19620'}}, {'post_id': ObjectId('64272400d877071a39e82bc6'), 'comment':
er3', 'text': 'This is a comment by quynAQyCuA on post 19622'}}, {'post_id': ObjectId('64272400d877071a39e82bc7'),
er_id': 'user3', 'text': 'This is a comment by quynAQyCuA on post 19623'}}, {'post_id': ObjectId('64272400d877071a
ment': {'user_id': 'user3', 'text': 'This is a comment by quynAQyCuA on post 19624'}}, {'post_id': ObjectId('64272
bc9'), 'comment': {'user_id': 'user3', 'text': 'This is a comment by quynAQyCuA on post 19625'}}, {'post_id': Obje
77071a39e82bcb'), 'comment': {'user_id': 'user3', 'text': 'This is a comment by quynAQyCuA on post 19627'}}, {'pos
'64272400d877071a39e82bcc'), 'comment': {'user_id': 'user3', 'text': 'This is a comment by quynAQyCuA on post 1962
: ObjectId('64272400d877071a39e82bcd'), 'comment': {'user_id': 'user3', 'text': 'This is a comment by quynAQyCuA o
 {'post_id': ObjectId('64272400d877071a39e82bce'), 'comment': {'user_id': 'user3', 'text': 'This is a comment by q
t 19630'}}, {'post_id': ObjectId('64272400d877071a39e82bcf'), 'comment': {'user_id': 'user3', 'text': 'This is a c
yCuA on post 19631'}}, {'post_id': ObjectId('64272400d877071a39e82bd1'), 'comment': {'user_id': 'user3', 'text': '
t by quynAQyCuA on post 19633'}}, {'post_id': ObjectId('64272400d877071a39e82bd4'), 'comment': {'user_id': 'user3'
is a comment by quynAQyCuA on post 19636'}}, {'post_id': ObjectId('64272400d877071a39e82bd6'), 'comment': {'user_i
xt': 'This is a comment by quynAQyCuA on post 19638'}}, {'post_id': ObjectId('64272400d877071a39e82bd7'), 'comment
user3', 'text': 'This is a comment by quynAQyCuA on post 19639'}}, {'post_id': ObjectId('64272400d877071a39e82bd8'
user_id': 'user3', 'text': 'This is a comment by quynAQyCuA on post 19640'}}, {'post_id': ObjectId('64272400d87707
omment': {'user_id': 'user3', 'text': 'This is a comment by quynAQyCuA on post 19643'}}, {'post_id': ObjectId('642
82bdc'), 'comment': {'user_id': 'user3', 'text': 'This is a comment by quynAQyCuA on post 19644'}}]
TOTAL RESULTS FOUND:
 9954
```

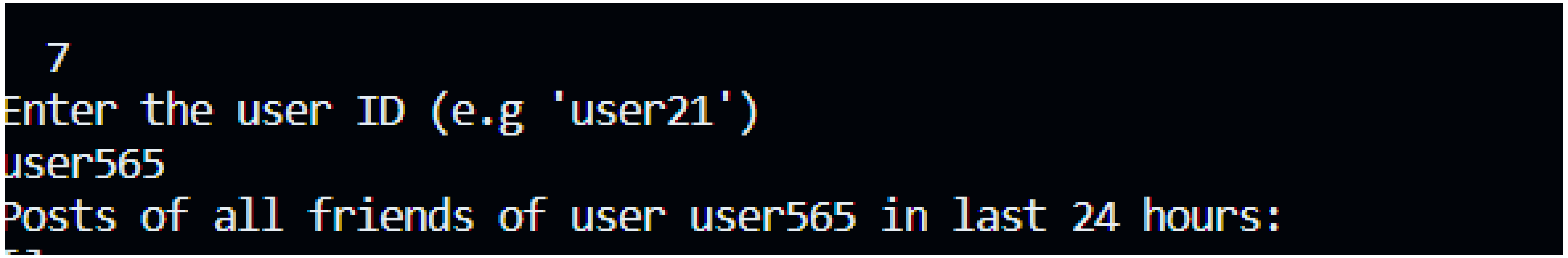
Results



```
  5
Enter the topic
Health
```

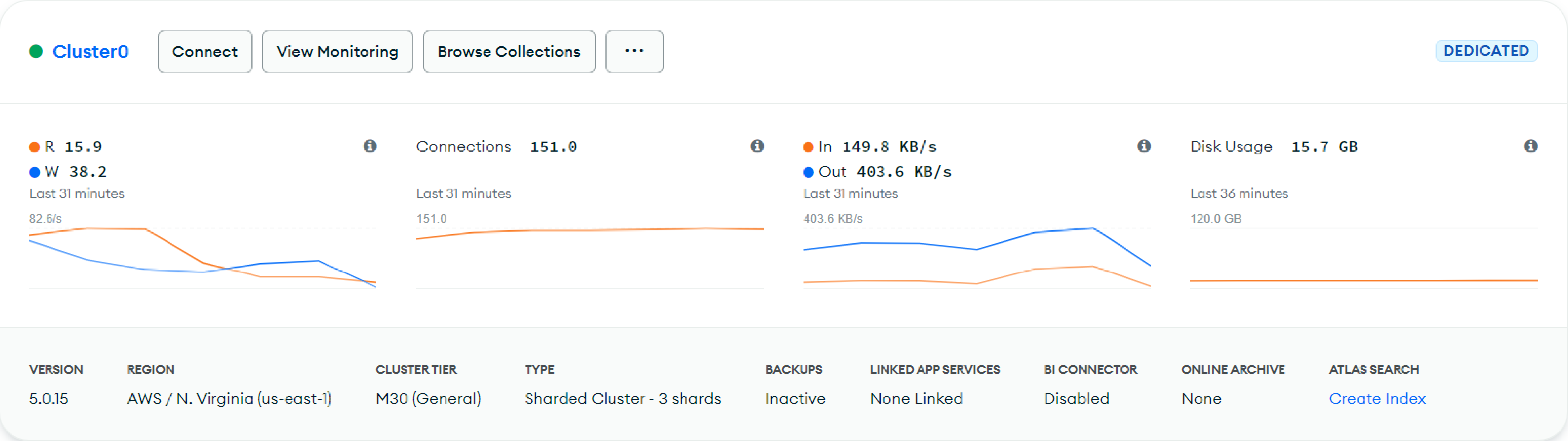


```
  6
Enter the value of k
8
Top 8 most popular topics in terms of posts:
[{'_id': 'Health', 'count': 2003}, {'_id': 'Politics', 'count': 1992}, {'_id': 'Food', 'count': 1987}, {'_id': 'Sports', 'count': 1977
, 'count': 1967}, {'_id': 'Music', 'count': 1962}, {'_id': 'Education', 'count': 1945}, {'_id': 'Travel', 'count': 1942}]
TOTAL RESULTS FOUND:
 8
```



```
  7
Enter the user ID (e.g 'user21')
user565
Posts of all friends of user user565 in last 24 hours:
```

Results



● Cluster0

DEDICATED

● R 15.9
● W 38.2
Last 31 minutes

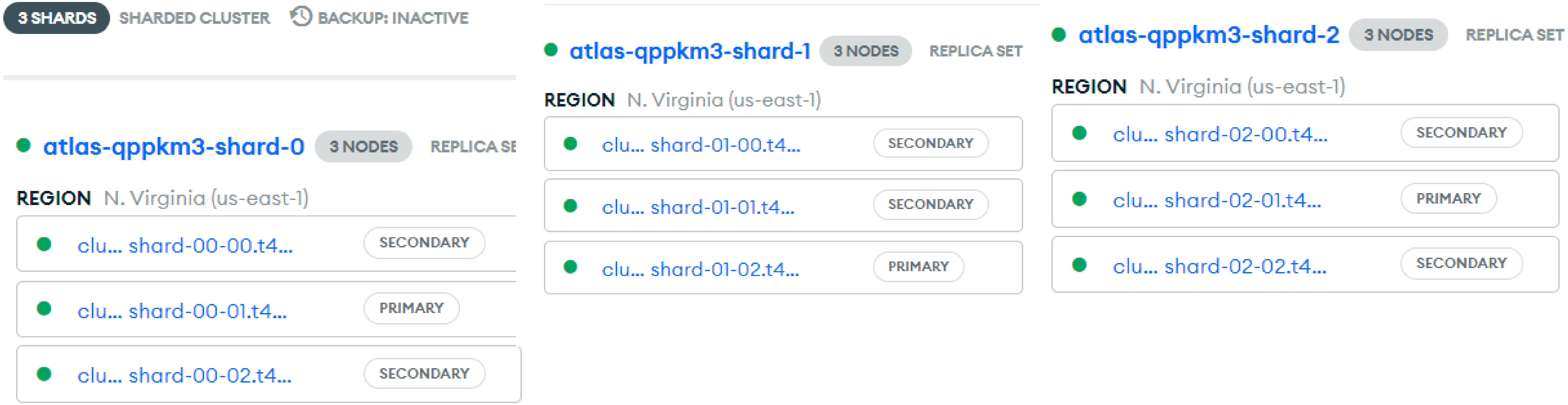
Connections 151.0
Last 31 minutes

● In 149.8 KB/s
● Out 403.6 KB/s
Last 31 minutes

Disk Usage 15.7 GB
Last 36 minutes

VERSION	REGION	CLUSTER TIER	TYPE	BACKUPS	LINKED APP SERVICES	BI CONNECTOR	ONLINE ARCHIVE	ATLAS SEARCH
5.0.15	AWS / N. Virginia (us-east-1)	M30 (General)	Sharded Cluster - 3 shards	Inactive	None Linked	Disabled	None	Create Index

Results



3 SHARDS SHARDED CLUSTER BACKUP: INACTIVE

● atlas-qppkm3-shard-0 3 NODES REPLICA SE

REGION N. Virginia (us-east-1)

● clu... shard-00-00.t4... SECONDARY

● clu... shard-00-01.t4... PRIMARY

● clu... shard-00-02.t4... SECONDARY

● atlas-qppkm3-shard-1 3 NODES REPLICA SET

REGION N. Virginia (us-east-1)

● clu... shard-01-00.t4... SECONDARY

● clu... shard-01-01.t4... SECONDARY

● clu... shard-01-02.t4... PRIMARY

● atlas-qppkm3-shard-2 3 NODES REPLICA SET

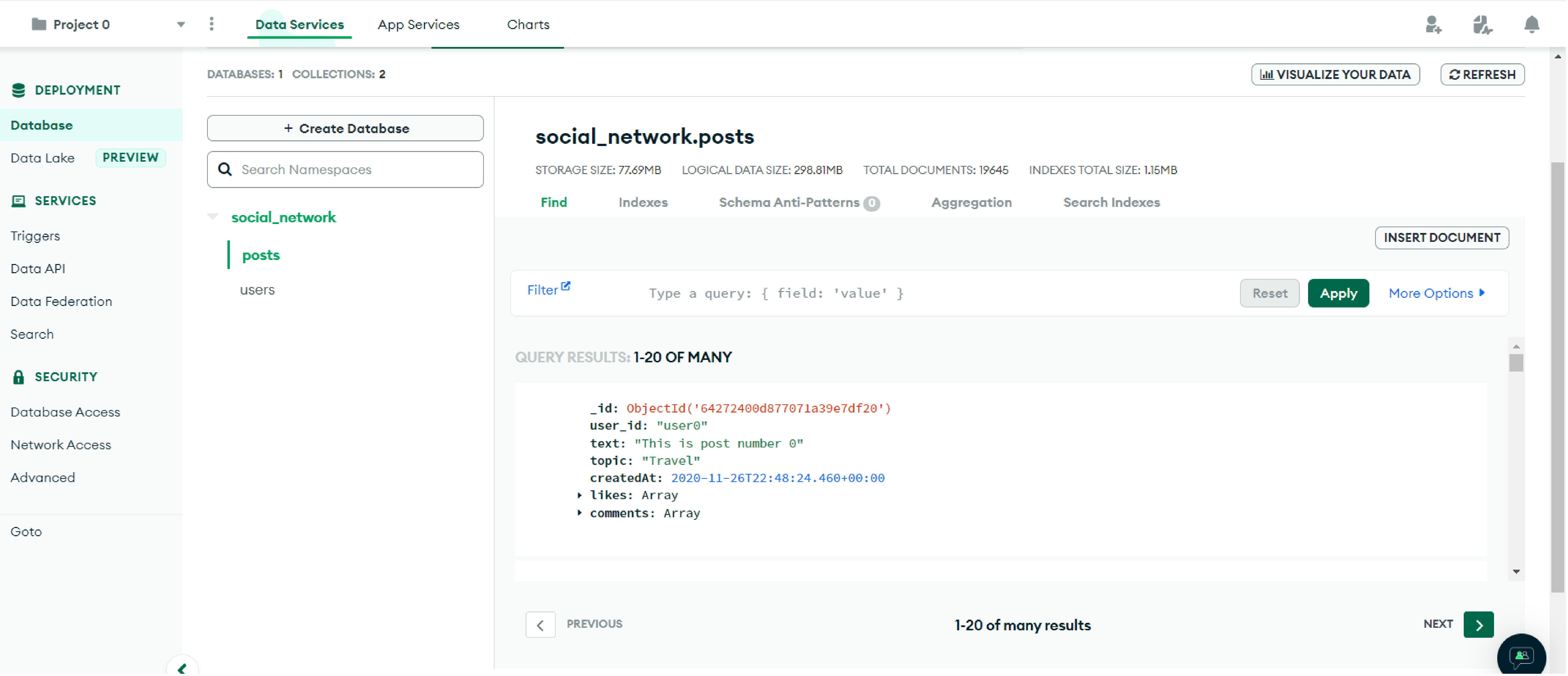
REGION N. Virginia (us-east-1)

● clu... shard-02-00.t4... SECONDARY

● clu... shard-02-01.t4... PRIMARY

● clu... shard-02-02.t4... SECONDARY

Results



DATABASES: 1 COLLECTIONS: 2

social_network.posts

STORAGE SIZE: 77.69MB LOGICAL DATA SIZE: 298.81MB TOTAL DOCUMENTS: 19645 INDEXES TOTAL SIZE: 1.15MB

QUERY RESULTS: 1-20 OF MANY

```
_id: ObjectId('64272400d877071a39e7df20')
user_id: "user0"
text: "This is post number 0"
topic: "Travel"
createdAt: 2020-11-26T22:48:24.460+00:00
▸ likes: Array
▸ comments: Array
```

1-20 of many results

THANK YOU