Big Data

Worksheet 2

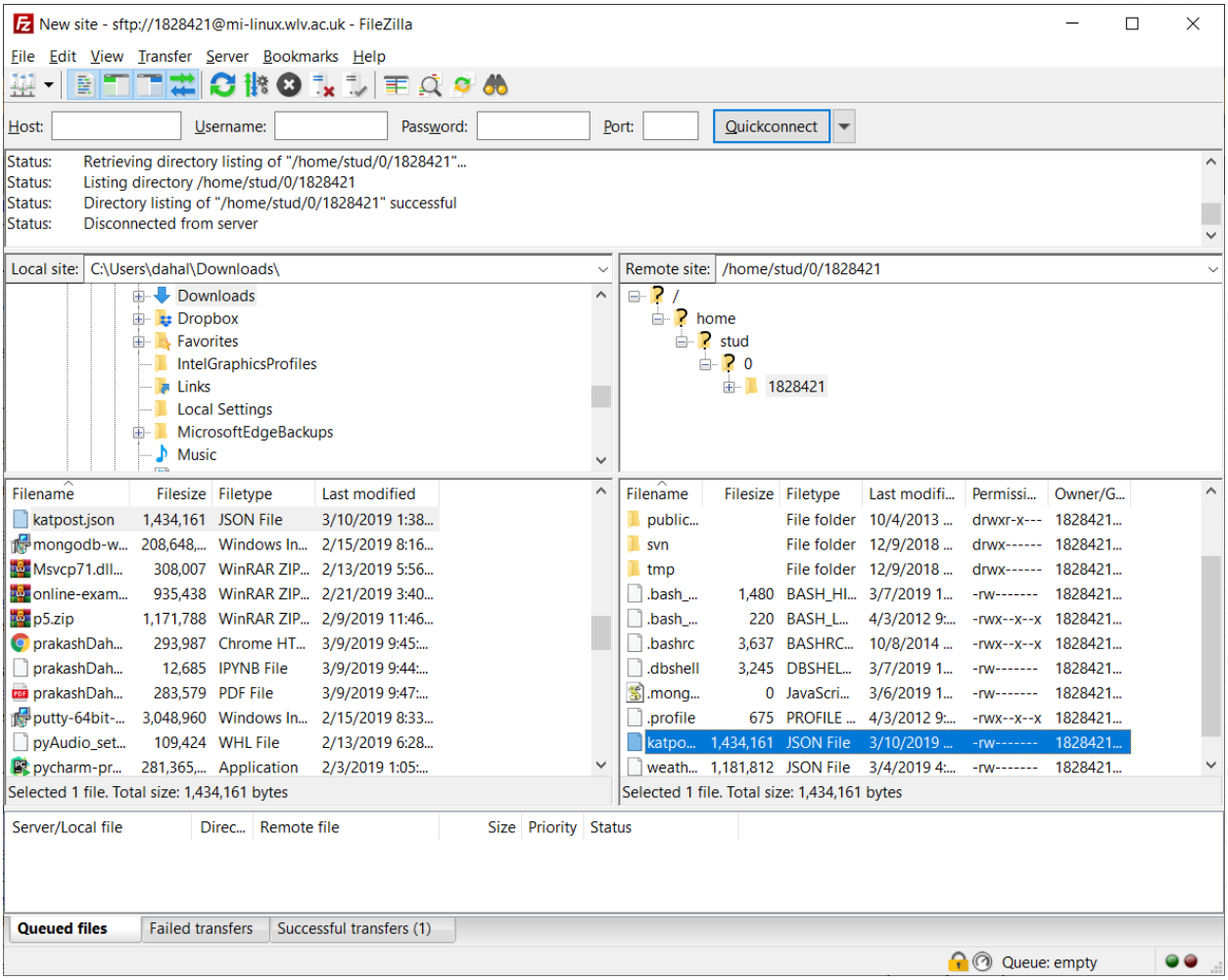
Name: Prakash Dahal

Student ID: 1828421

Remainder: 2 (katpost)

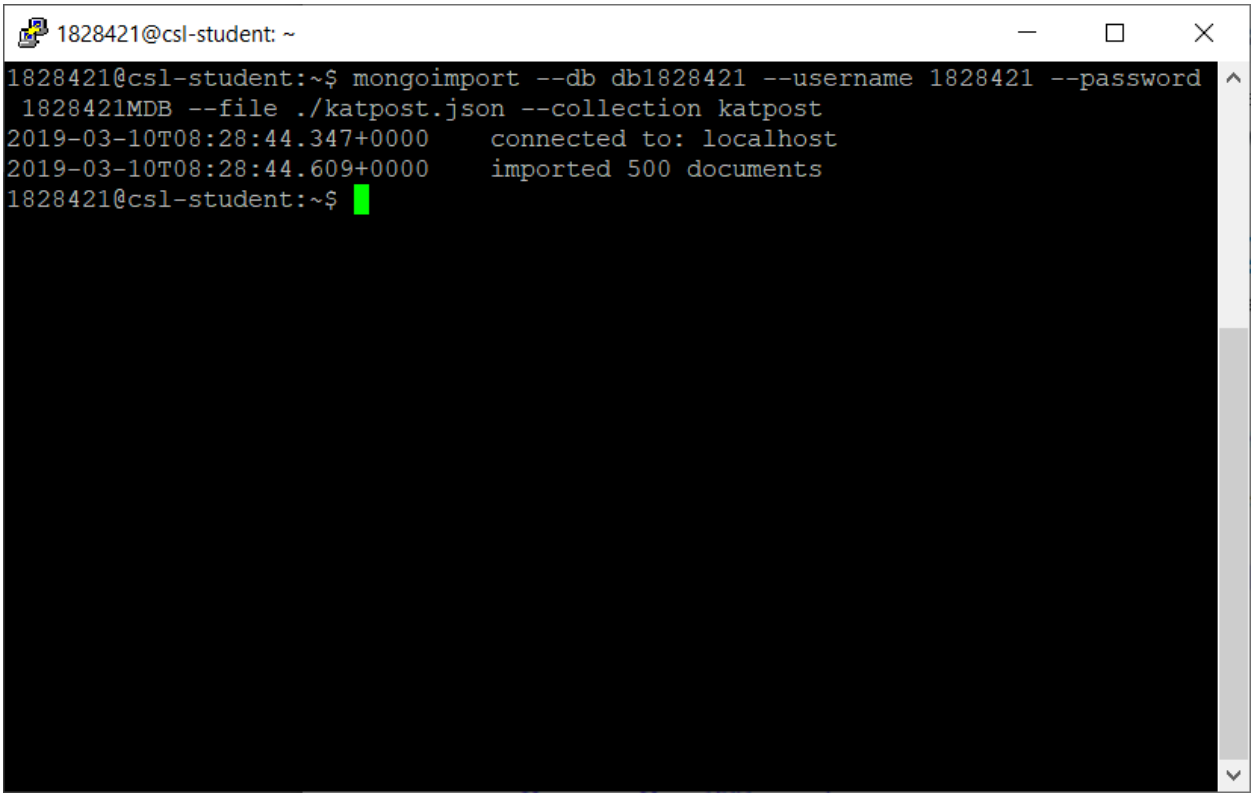
Name: Prakash Dahal Student ID: 1828421

Storing JSON file (katpost) through FileZilla data in university Linux software.

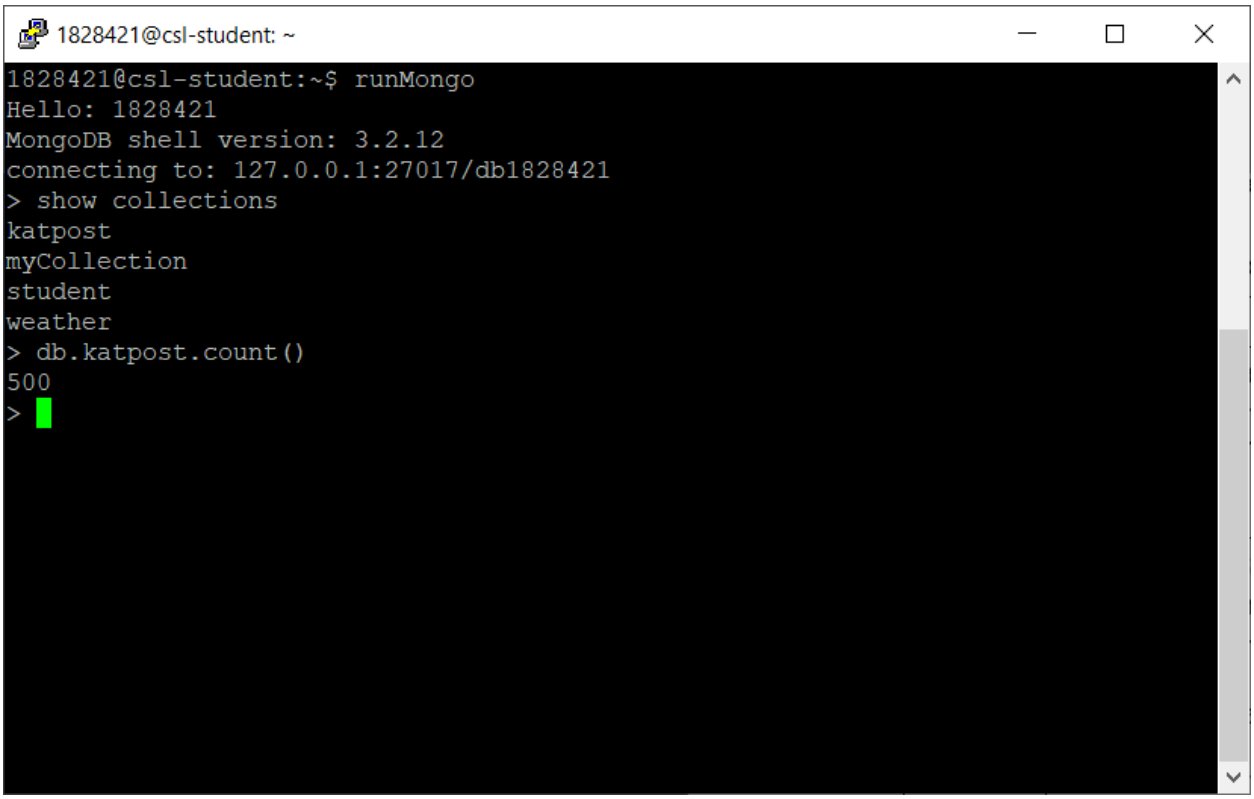


A. Importing Data:

a) Now importing Katpost.json file into the database



b) Showing all the collections available in the database. Katpost is a new collection which has just been imported from the above import. It has got 500 documents in the collection.



B. Analyze the data:

a) Showing single document of the collection: There are two ways of doing it. They are as follows:

1. Using limit (1) function in find() :

Query:

db.katpost.find().pretty().limit(1)

```
1828421@csl-student: ~
> db.katpost.find().pretty().limit(1)
{
        "_id" : ObjectId("5c5d6988be9a1c0096797d8a"),
        "created_at" : "Fri Feb 08 11:01:57 +0000 2019",
        "id" : NumberLong("1093827265941639168"),
        "id_str" : "1093827265941639168",
        "text" : "Nepal Communist Party lawmaker Mahesh Basnet has demanded the government introduce a system to return the plastic p… https://t.co/aumOutM3w6",
        "truncated" : true,
        "entities" : {
                "hashtags" : [ ],
                "symbols" : [ ],
                "user_mentions" : [ ],
                "urls" : [
                        {
                                "url" : "https://t.co/aumOutM3w6",
                                "expanded_url" : "https://twitter.com/i/web/status/1093827265941639168",
                                "display_url" : "twitter.com/i/web/status/1…",
                                "indices" : [
                                        117,
                                        140
                                ]
                        }
                ]
        },
        "source" : "<a href=\"https://about.twitter.com/products/tweetdeck\" rel=\"nofollow\">TweetDeck</a>",
        "in_reply_to_status_id" : null,
        "in_reply_to_status_id_str" : null,
        "in_reply_to_user_id" : null,
        "in_reply_to_user_id_str" : null,
        "in_reply_to_screen_name" : null,
        "user" : {
                "id" : 625760052,
                "id_str" : "625760052",
                "name" : "The Kathmandu Post",
                "screen_name" : "kathmandupost",
                "location" : "Kathmandu",
                "description" : "Nepal's leading national daily. Follow us for the latest news, analysis, and opinion. Founded in 1993.",
                "url" : "http://t.co/83LrI7hUfU",
                "entities" : {
                        "url" : {
                                "urls" : [
                                        {
                                                "url" : "http://t.co/83LrI7hUfU",
                                                "expanded_url" : "http://kathmandupost.ekantipur.com",
                                                "display_url" : "kathmandupost.ekantipur.com",
                                                "indices" : [
                                                        0,
                                                        22
                                                ]
                                        }
                                ]
                        },
                        "description" : {
                                "urls" : [ ]
                        }
                },
                "protected" : false,
                "followers_count" : 408509,
                "friends_count" : 61,
                "listed_count" : 644,
                "created_at" : "Tue Jul 03 16:21:08 +0000 2012",
                "favourites_count" : 61,
                "utc_offset" : null,
                "time_zone" : null,
                "geo_enabled" : true,
                "verified" : true,
                "statuses_count" : 82338,
                "lang" : "en",
                "contributors_enabled" : false,
                "is_translator" : false,
                "is_translation_enabled" : false,
                "profile_background_color" : "9AE4E8",
                "profile_background_image_url" : "http://abs.twimg.com/images/themes/theme1/bg.png",
                "profile_background_image_url_https" : "https://abs.twimg.com/images/themes/theme1/bg.png",
                "profile_background_tile" : true,
                "profile_image_url" : "http://pbs.twimg.com/profile_images/665467018028711936/0xZ68kA9_normal.jpg",
                "profile_image_url_https" : "https://pbs.twimg.com/profile_images/665467018028711936/0xZ68kA9_normal.jpg",
                "profile_banner_url" : "https://pbs.twimg.com/profile_banners/625760052/1439924720",
                "profile_link_color" : "DD2E44",
                "profile_sidebar_border_color" : "FFFFFF",
                "profile_sidebar_fill_color" : "DDEEF6",
                "profile_text_color" : "333333",
                "profile_use_background_image" : true,
                "has_extended_profile" : false,
                "default_profile" : false,
                "default_profile_image" : false,
                "following" : false,
                "follow_request_sent" : false,
                "notifications" : false,
                "translator_type" : "none"
        },
        "geo" : null,
        "coordinates" : null,
        "place" : null,
        "contributors" : null,
        "is_quote_status" : false,
        "retweet_count" : 0,
        "favorite_count" : 4,
        "favorited" : false,
        "retweeted" : false,
        "possibly_sensitive" : false,
        "lang" : "en"
}
>
```

2. Using findOne() function:

Query:

db.katpost.findOne()

1828421@csl-student: ~

```
> db.katpost.findOne()
{
        "_id" : ObjectId("5c5d6988be9a1c0096797d8a"),
        "created_at" : "Fri Feb 08 11:01:57 +0000 2019",
        "id" : NumberLong("1093827265941639168"),
        "id_str" : "1093827265941639168",
        "text" : "Nepal Communist Party lawmaker Mahesh Basnet has demanded the government introduce a system to return the plastic p… https://t.co/aumOutM3w6",
        "truncated" : true,
        "entities" : {
                "hashtags" : [ ],
                "symbols" : [ ],
                "user_mentions" : [ ],
                "urls" : [
                        {
                                "url" : "https://t.co/aumOutM3w6",
                                "expanded_url" : "https://twitter.com/i/web/status/1093827265941639168",
                                "display_url" : "twitter.com/i/web/status/1…",
                                "indices" : [
                                        117,
                                        140
                                ]
                        }
                ]
        },
        "source" : "<a href=\"https://about.twitter.com/products/tweetdeck\" rel=\"nofollow\">TweetDeck</a>",
        "in_reply_to_status_id" : null,
        "in_reply_to_status_id_str" : null,
        "in_reply_to_user_id" : null,
        "in_reply_to_user_id_str" : null,
        "in_reply_to_screen_name" : null,
        "user" : {
                "id" : 625760052,
                "id_str" : "625760052",
                "name" : "The Kathmandu Post",
                "screen_name" : "kathmandupost",
                "location" : "Kathmandu",
                "description" : "Nepal's leading national daily. Follow us for the latest news, analysis, and opinion. Founded in 1993.",
                "url" : "http://t.co/83LrI7hUfU",
                "entities" : {
                        "url" : {
                                "urls" : [
                                        {
                                                "url" : "http://t.co/83LrI7hUfU",
                                                "expanded_url" : "http://kathmandupost.ekantipur.com",
                                                "display_url" : "kathmandupost.ekantipur.com",
                                                "indices" : [
                                                        0,
                                                        22
                                                ]
                                        }
```

b) Showing unique values of one filed.

For this, entities.user_mentions.name is selected to get unique values.

Query:

db.katpost.distinct("entities.user_mentions.name")

1828421@csl-student: ~

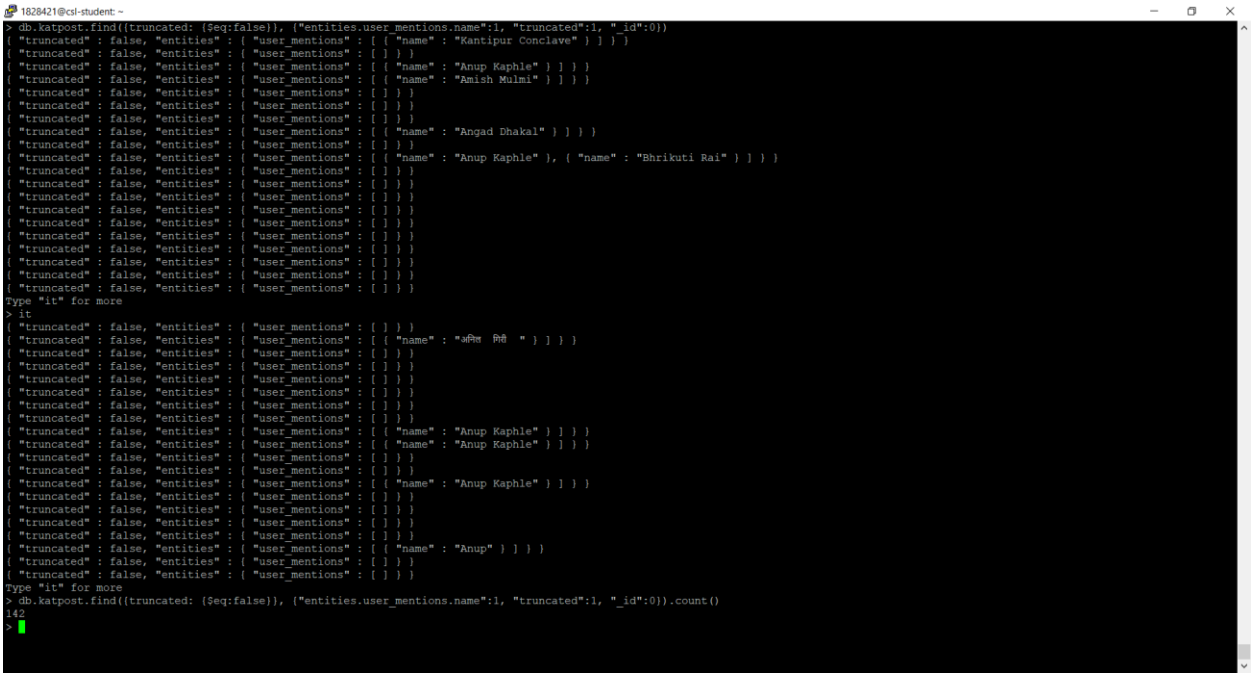
```
> db.katpost.distinct("entities.user_mentions.name")
[
        "Kantipur Conclave",
        "Anup Kaphle",
        "Amish Mulmi",
        "Angad Dhakal",
        "Bhrikuti Rai",
        "अनिल  गिरी  ",
        "Anup",
        "हेलो  सरकार  HelloSarkar",
        "keshav thapa",
        "Sanjog Manandhar",
        "Amitava Kumar",
        "Avasna  Pandey",
        "KHORUNGA",
        "Timothy Aryal",
        "Dinesh Kafle",
        "The Kathmandu Post",
        "Thomas Heaton",
        "Tsering Ngodup Lama",
        "Pramod Mishra",
        "Alisha Sijapati",
        "David Kainee"
]
>
```

c) Showing some set of documents based on criteria.

Here, showing two fields of document's; entities.user_mentions.name and truncated which has truncated false. 142 result is obtained from this query.

Query:

db.katpost.find ({truncated: {$eq:false}}, {"entities.user_mentions.name":1, "truncated":1, "_id":0}).count()



```
1828421@csl-student: ~
> db.katpost.find({truncated: {$eq:false}}, {"entities.user_mentions.name":1, "truncated":1, "_id":0})
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Kantipur Conclave" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Anup Kaphle" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Amish Mulmi" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Angad Dhakal" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Anup Kaphle" }, { "name" : "Bhrikuti Rai" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
Type "it" for more
> it
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "अनिल  मिसी  " } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Anup Kaphle" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Anup Kaphle" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Anup Kaphle" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ { "name" : "Anup" } ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
{ "truncated" : false, "entities" : { "user_mentions" : [ ] } }
Type "it" for more
> db.katpost.find({truncated: {$eq:false}}, {"entities.user_mentions.name":1, "truncated":1, "_id":0}).count()
142
>
```

d) Using regular expression to find word “rural” in field **text** and displaying default _id, text, truncated and entities.user_mentions.name fields. The word rural is case insensitive which means this query can display result even if rural is in capital letters like Rural:

Query:

db.katpost.find({'text':{$regex:/rural/i}},{'text':1,'truncated':1, 'entities.user_mentions.name':1}).pretty()

```
1828421@csl-student: ~
> db.katpost.find({'text':{$regex:/rural/i}},{'text':1,'truncated':1, 'entities.user_mentions.name':1}).pretty().count()
5
> db.katpost.find({'text':{$regex:/rural/i}},{'text':1,'truncated':1, 'entities.user_mentions.name':1}).pretty()
{
        "_id" : ObjectId("5c5d6988be9a1c0096797d8c"),
        "text" : "Most women in rural areas of Achham district do not divulge information or talk about their reproductive health con… https://t.co/tfCyoTK538",
        "truncated" : true,
        "entities" : {
                "user_mentions" : [ ]
        }
}
{
        "_id" : ObjectId("5c5d6997be9a1c0096797e25"),
        "text" : "A suspension bridge constructed over Mahakali River in Lali of Lekam Rural Municipality-3 in Darchula district one-… https://t.co/EAVJfdfok6",
        "truncated" : true,
        "entities" : {
                "user_mentions" : [ ]
        }
}
{
        "_id" : ObjectId("5c5d69a6be9a1c0096797ec3"),
        "text" : "Municipalities, rural municipalities are among the most corrupt agencies, CIAA survey shows \n\nhttps://t.co/BglrmS5nIe",
        "truncated" : false,
        "entities" : {
                "user_mentions" : [ ]
        }
}
{
        "_id" : ObjectId("5c5d69a7be9a1c0096797ec9"),
        "text" : "Economic assistance can be used to prevent migration by rebuilding rural economies, Mahendra P Lama\n\nhttps://t.co/NtrExM83G7",
        "truncated" : false,
        "entities" : {
                "user_mentions" : [ ]
        }
}
{
        "_id" : ObjectId("5c5d69a9be9a1c0096797ee3"),
        "text" : "#Opinion Economic assistance can be used to prevent migration by rebuilding rural economies, by Mahendra P Lama\n\nhttps://t.co/NtrExM83G7",
        "truncated" : false,
        "entities" : {
                "user_mentions" : [ ]
        }
}
>
```

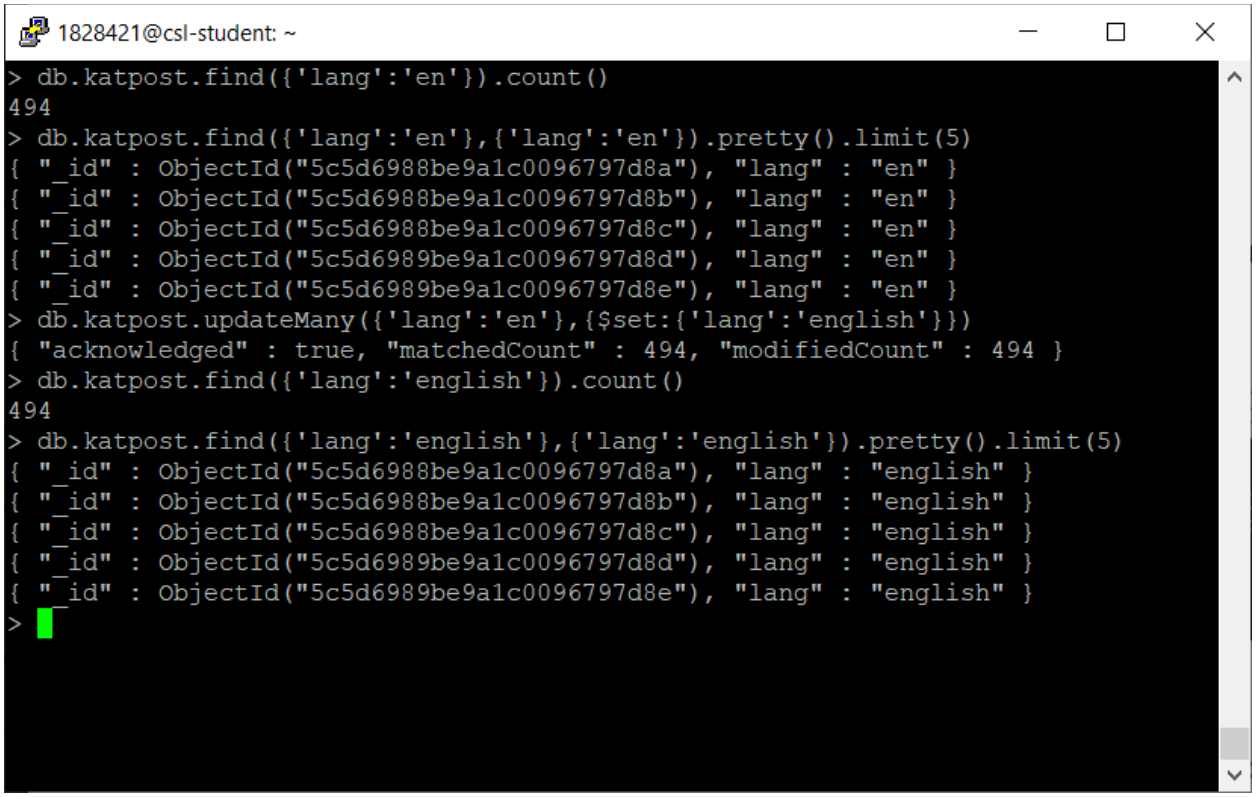
C. Reshape the collection:

Data reshaping and storing some required in new collection as well:

a) Updating a field within the collection. The updating field is lang. Updated value is “en” of key “lang” as “english”. Total 494 documents have field lang as eng which is changed to english and displayed 5 documnts out of 494.

Query:

db.katpost.updateMany({'lang':'en'},{$set:{'lang':'english'}})

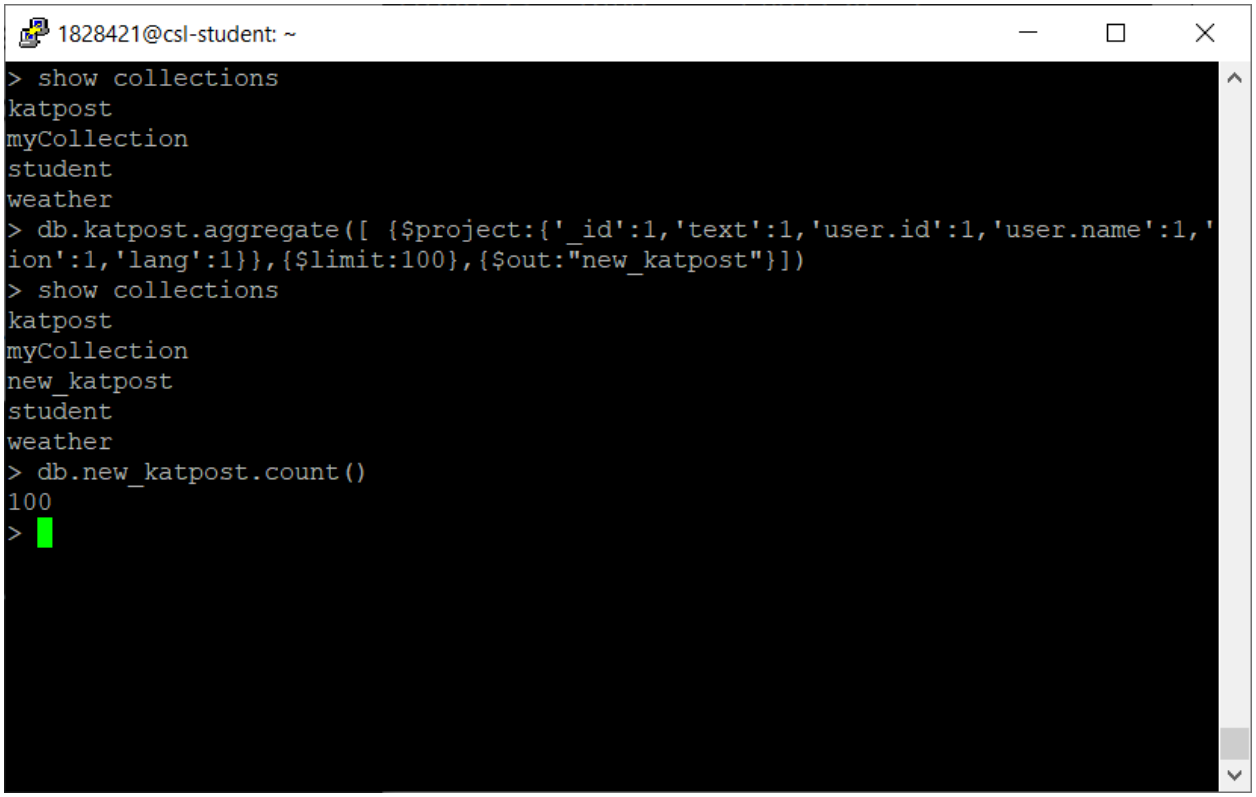


```
1828421@csl-student: ~
> db.katpost.find({'lang':'en'}).count()
494
> db.katpost.find({'lang':'en'},{'lang':'en'}).pretty().limit(5)
{ "_id" : ObjectId("5c5d6988be9a1c0096797d8a"), "lang" : "en" }
{ "_id" : ObjectId("5c5d6988be9a1c0096797d8b"), "lang" : "en" }
{ "_id" : ObjectId("5c5d6988be9a1c0096797d8c"), "lang" : "en" }
{ "_id" : ObjectId("5c5d6989be9a1c0096797d8d"), "lang" : "en" }
{ "_id" : ObjectId("5c5d6989be9a1c0096797d8e"), "lang" : "en" }
> db.katpost.updateMany({'lang':'en'},{$set:{'lang':'english'}})
{ "acknowledged" : true, "matchedCount" : 494, "modifiedCount" : 494 }
> db.katpost.find({'lang':'english'}).count()
494
> db.katpost.find({'lang':'english'},{'lang':'english'}).pretty().limit(5)
{ "_id" : ObjectId("5c5d6988be9a1c0096797d8a"), "lang" : "english" }
{ "_id" : ObjectId("5c5d6988be9a1c0096797d8b"), "lang" : "english" }
{ "_id" : ObjectId("5c5d6988be9a1c0096797d8c"), "lang" : "english" }
{ "_id" : ObjectId("5c5d6989be9a1c0096797d8d"), "lang" : "english" }
{ "_id" : ObjectId("5c5d6989be9a1c0096797d8e"), "lang" : "english" }
>
```

b) Our katpost collection has 500 documents. Now selecting required fields and storing some of the rows into new collection. The name for the new collection is new_katpost. Out command creates new collection if it does not exist.

Query:

db.katpost.aggregate([
{$project:{'_id':1,'text':1,'user.id':1,'user.name':1,'user.description':1,'lang':1}},{$limit:100},{$out:"new_katpost"}])



c) Displaying 3 documents of new_katpost collection:

```
1828421@csl-student: ~
> db.new_katpost.find().limit(3).pretty()
{
        "_id" : ObjectId("5c5d6988be9a1c0096797d8a"),
        "text" : "Nepal Communist Party lawmaker Mahesh Basnet has demanded the government introduce a sy
stem to return the plastic p… https://t.co/aumOutM3w6",
        "user" : {
                "id" : 625760052,
                "name" : "The Kathmandu Post",
                "description" : "Nepal's leading national daily. Follow us for the latest news, analysis,
 and opinion. Founded in 1993."
        },
        "lang" : "english"
}
{
        "_id" : ObjectId("5c5d6988be9a1c0096797d8b"),
        "text" : "The sister of Thailand’s king entered the race to become prime minister on Friday as th
e candidate of a populist pa… https://t.co/f4FwkAN43b",
        "user" : {
                "id" : 625760052,
                "name" : "The Kathmandu Post",
                "description" : "Nepal's leading national daily. Follow us for the latest news, analysis,
 and opinion. Founded in 1993."
        },
        "lang" : "english"
}
{
        "_id" : ObjectId("5c5d6988be9a1c0096797d8c"),
        "text" : "Most women in rural areas of Achham district do not divulge information or talk about t
heir reproductive health con… https://t.co/tfCyoTK538",
        "user" : {
                "id" : 625760052,
                "name" : "The Kathmandu Post",
                "description" : "Nepal's leading national daily. Follow us for the latest news, analysis,
 and opinion. Founded in 1993."
        },
        "lang" : "english"
}
>
```

D. Advantage and Disadvantage:

The current approach we are using for Big-Data is MongoDB. Everything has its both merit and demerits sites. One advantage and one disadvantage are listed and explained below:

a. Advantage:

i. Flexibility:

MongoDB is a schema-less database which mean any type of data can be stored in one document. It is highly flexible because all the fields may not have same types of data.

For example;

We have courses document of Students,

Student1{science, math, computer}

Student2{environment, biology, math}

Student3{science, environment, math, computer, biology, geology}

So, all students may not enroll on same subject so different student may have different fields which can be implemented in big-data easily. Therefore, it is highly flexible.

On the other hand, since data are stored in JSON format, it is easy to extract and store data.

b. Disadvantage:

ii. Memory Limitation:

MongoDB uses mapped records and OS handles the caching. The maximum BSON document size is 16 MB. It does not give direct functionality of joining. Therefore, there can be redundant data in the same collection which uses unnecessary memory.