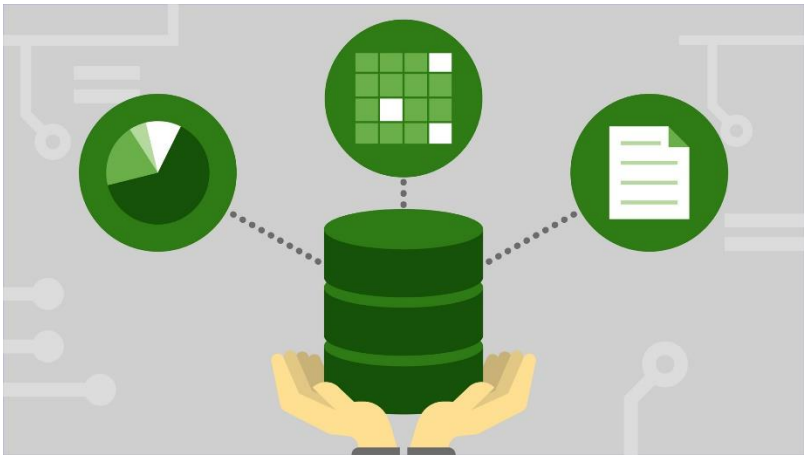
به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر



آزمایشگاه پایگاه‌داده

دستورکار شماره ۶

شماره دانشجویی

۸۱۰۱۹۶۴۴۳

اردیبهشت ۱۴۰۰

هومان چمنی

گزارش فعالیت‌های انجام شده

گام اول:

داده‌ها به صورت غیردستی و اتومات از طریق یک اسکریپ پایتون به دیتابیس وارد شده‌اند. عکس کد زده شده در زیر قابل مشاهده می‌باشد:

C: > Users > hoomo > Downloads > Term 8 > db lab > Projects > dblab6 > get_tweets.py > ...

```python
import time
import requests, json
from pymongo import MongoClient

client = MongoClient()
db = client.tweets
url = 'https://www.sahamyab.com/guest/twiter/list?v=0.1'
count_needed, sleep_time = 1000, 60
count_sofar = db.tweets.count_documents({})

while count_sofar < count_needed:
    response = requests.request('GET', url, headers={'User-Agent': 'Chrome/61'})
    result = response.status_code
    if result == requests.codes.ok:
        data = response.json()['items']
        for d in data:
            try:
                db.tweets.replace_one({"id": d["id"]}, d, upsert=True)
            except Exception as e:
                print("Upsert exception: " + str(e))
        count_sofar = db.tweets.count_documents({})
    else:
        print("Response code error: " + str(result))
    print(f'Count of fetched tweets is {count_sofar}')
    time.sleep(sleep_time)

count_sofar = db.tweets.count_documents({})
```

نحوه اطمینان از درج شدن حداقل ۱۰۰۰ ورودی نیز به این صورت است که یک متغیر count_sofar تعریف شده که نگه‌دارنده تعداد توییت‌هایی است که تا کنون جمع و ذخیره شده. این متغیر در هربار گردش حلقه بروزرسانی شده و شرط خروج از حلقه نیز این است که متغیر ذکرشده حداقل مقدار ۱۰۰۰ را داشته باشد.

فیلدهایی نیز که اضافه شده‌اند به صورت زیر می‌باشند:

_id, id, sendTime, sendTimePersian, retwitSendTime, retwitSendTimePersian, retwitSenderName, retwitSenderUsername, retwitSenderProfileImage, senderName, senderUsername, senderProfileImage, content, lastLikeNickName, likeCount, retwitCount, type, scoredPostDate, retwitId, finalPullDatePersian

البته لازم به ذکر است که توییت‌هایی که از جنس retweet نمی‌باشند فیلدهای مربوط به retweet را نیز ندارند.

گام دوم:

از کد زیر استفاده کردیم تا هشتگ‌های درون متن را به عنوان یک فیلد جدا اضافه کنیم:

```
import re, time
from pymongo import MongoClient

client = MongoClient()
start_time = time.time()

for item in client.tweets.tweets.find({}):
    item['hashtags'] = re.findall(r"#(\w+)", item['content'])
    client.tweets.tweets.update(
        {"_id" : item['_id']}, item)

end_time = time.time()
print(end_time - start_time)
```

زمان اجرای آن نیز طبق خروجی داده شده مقدار 0.2976 ذکر شده است.

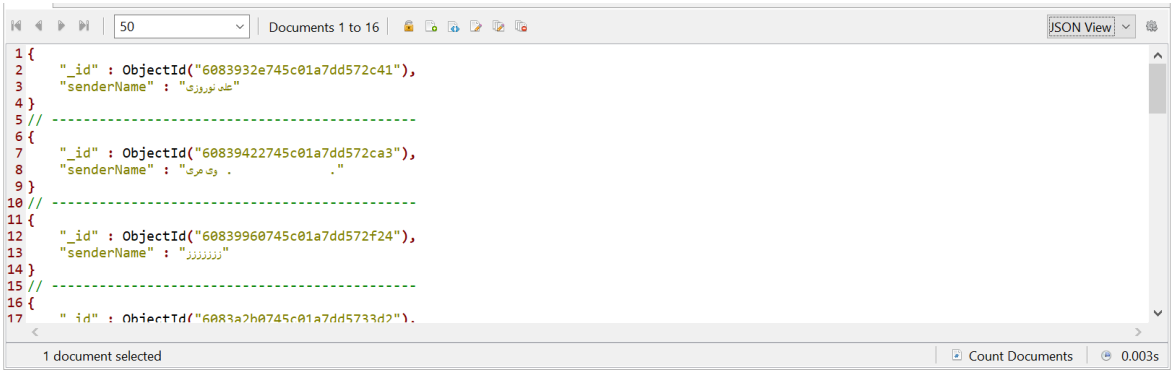
همان‌طور که در تصویر زیر نیز مشاهده می‌شود یک توییت که دارای هشتگ برکت بود به عنوان مثال آپدیت شده است:


```
{
    "_id": {
        "$oid": "608390d2745c01a7dd572b3a"
    },
    "id": "273717511",
    "sendTime": "2021-04-23T20:49:25Z",
    "sendTimePersian": "1400/02/04 01:19",
    "retwitSendTime": "2021-04-24T03:29:25Z",
    "retwitSendTimePersian": "1400/02/04 07:59",
    "retwitSenderName": "BARCA",
    "retwitSenderUsername": "campnou",
    "retwitSenderProfileImage": "default",
    "senderName": "Saham joo",
    "senderUsername": "sahamkha",
    "senderProfileImage": "8597f2ad-e509-4614-adb8-d2f3dd6efdc1",
    "content": "#برکت \nدارد.و شخصا انتظار دارم سهم از یکی دو روز آینده راه سبز برایش میسر شود
    "lastLikeNickName": "1500آرمان",
    "likeCount": "11",
    "retwitCount": "3",
    "type": "retwit",
    "scoredPostDate": "1619212540075",
    "retwitId": "273730567",
    "finalPullDatePersian": "",
    "hashtags": ["برکت"]
}
```

گام سوم:

بخش اول:

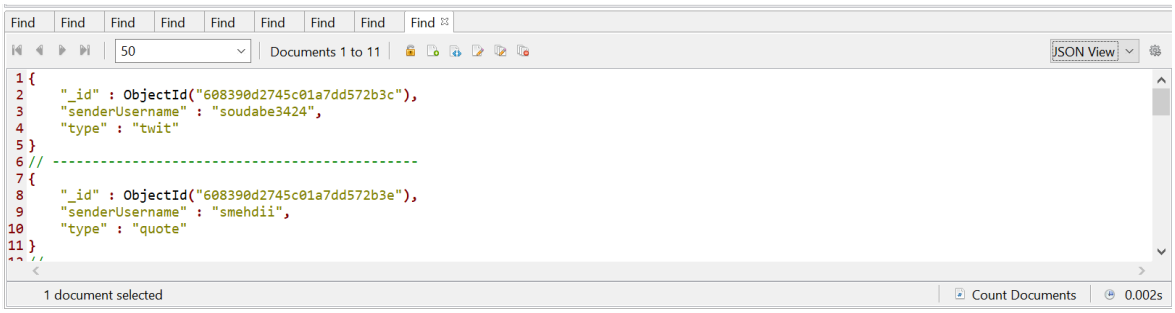
```
// Part3: section1
db.tweets.find(
    {"mediaContentType":"image/jpeg",
     "parentId":{"$exists" : true}},
     {"senderName":1})

```



زمان هم همان‌طور که در گوشه تصویر بالا معلوم است 0.003 می‌باشد.

بخش دوم:

```
db.tweets.find(
    {"sendTimePersian":{"$gt":"1400/02/04 07:45", "$lt":"1400/02/04 08:00"}},
    {"senderUsername":1,
        "type":1})
```



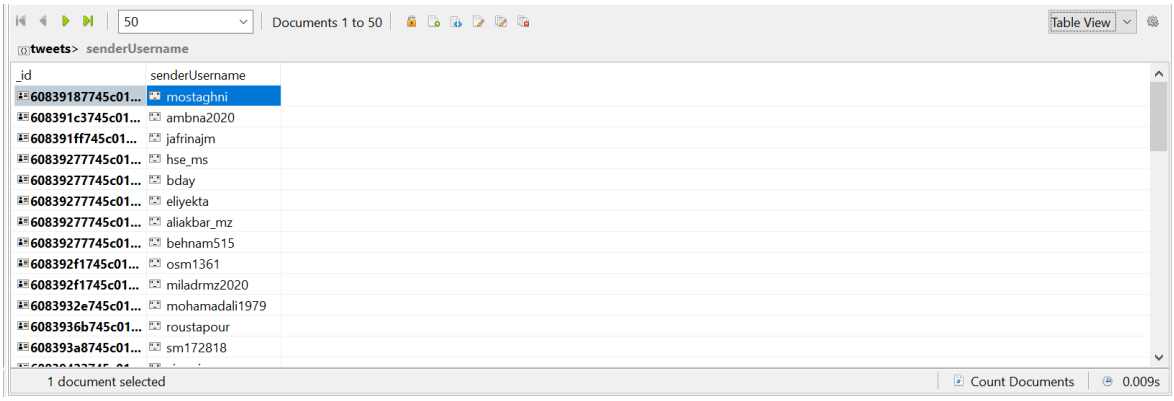
همان‌طور که در شکل بالا مشخص است حدود ۱۱ نفر سحرخیز! داریم که بین بازه ۷.۴۵ تا ۸ صبح توییت کرده‌اند و زمان اجراهم 0.002 می‌باشد.

بخش سوم:

بازه زمانی که شروع آن ۸ صبح است را در نظر گرفته‌ایم.

```
//// Part3: section3
db.tweets.find(
    {"sendTimePersian":{"$regex":".*/.*/.{1,2} 08:.*"}},
    {"senderUsername":1})
```

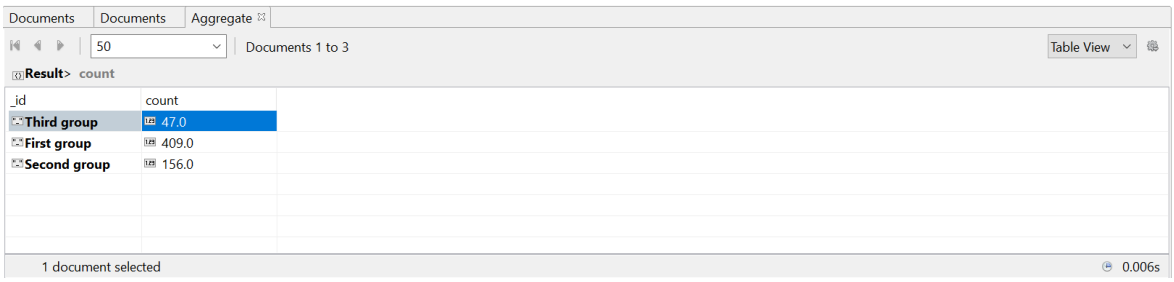
نتیجه ران کردن کووری نیز در زیر قابل مشاهده است. زمان اجرا هم 0.009 ثانیه است:



گام چهارم:

بخش اول:

```
db.tweets.aggregate(
    [
        {
            "$group" : {
                _id : "$senderUsername",
                "count" : {"$sum" : 1}
            }
        },
        {
            "$group" : {
                _id : {
                    "$cond" : {
                        "if" : {"$lt" : ["$count",2]},
                        "then" : "First group",
                        "else" : {
                            "$cond" : {
                                "if" : {"$lte" : ["$count",3]},
                                "then" : "Second group",
                                "else" : {
                                    "$cond" : {
                                        "if" : {"$gte" : ["$count",4]},
                                        "then" : "Third group",
                                        "else" : "No group!"
                                    }
                                }
                            }
                        }
                    }
                },
                "count" : {"$sum" : 1}
```

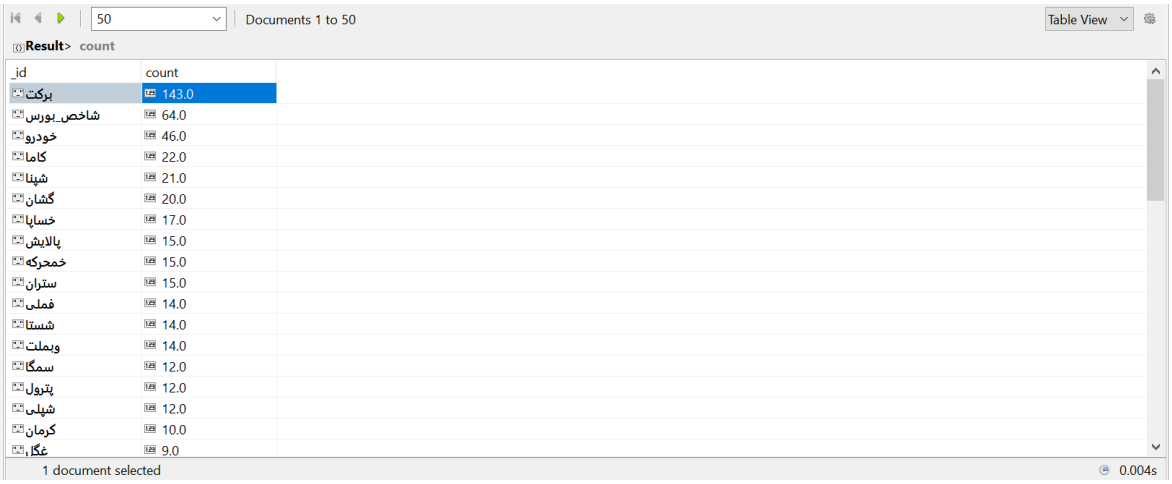


طبق عکس بالا نیز از دسته اول ۴۰۹ کاربرو از دسته دوم ۱۵۶ و از دسته آخر هم فقط ۴۷ کاربر داریم. زمان اجرای کووری نیز 0.006 می‌باشد.

بخش دوم:

```
//// Part4: section2
db.tweets.aggregate(
    [
        {
            "$unwind" : "$hashtags"

            /* Deconstructs an array field from the input documents
            to output a document for each element. Each output document is the input document
            with the value of the array field replaced by the element. */
        },
        {
            "$group" : {
                _id : "$hashtags",
                "count" : {"$sum" : 1}
            }
        },
        {
            "$sort" : { "count" : -1}
        }
    ]
)
```

در کووری بالا ابتدا از دستور unwind که توضیح آن زیر آن درج شده استفاده کردیم و در آخر هم سورت طبق تعداد به صورت نزولی انجام داده‌ایم که نتیجه در شکل زیر قابل مشاهده است:



زمان اجرای کووری نیز 0.004 می‌باشد.

بخش سوم:

```
//// Part4: section3
db.tweets.aggregate([
    {
        "$match" : {
            "parentId" : {"$exists" : true}
        }
    },
    {
        "$unset" : "type"
    }
])
```

Find | Documents | Aggregate | Aggregate

50 | Documents 1 to 50 | Tree View

Key	Value	Type
(1) { _id : 608390d2745c01a7dd572b3e}	{ 17 fields }	Document
_id	608390d2745c01a7dd572b3e	ObjectId
id	273730532	String
sendTime	2021-04-24T03:28:54Z	String
sendTimePersian	1400/02/04 07:58	String
parentSendTime	2021-04-21T16:09:46Z	String
parentSendTimePersian	1400/02/01 20:39	String
parentId	272562624	String
parentSenderName	ali	String
parentSenderUsername	ali9806	String
parentSenderProfileImage	default	String
parentContent	#خبری ارزش نیست وبصادر صیدی دو دلاری کجاست	String
senderName	Mehdi	String
senderUsername	smehdii	String
senderProfileImage	default	String
content	بجاش صیدی یک سنتی اومده	String
finalPullDatePersian		String
hashtags	[0 elements]	Array
(2) { _id : 608390d2745c01a7dd572b4c}	{ 17 fields }	Document

1 item selected 0.005s

طبق عکس بالا مشاده میشود که دیگر فیلدی به نام type در بین اعضا وجود ندارد. زمان اجرای کووری نیز 0.005 میباشد.

بخش چهارم:

از کووری که در قسمت ۲ زدیم استفاده کردیم با این تفاوت که در آخر آن یک گروه‌بندی دیگر کردیم تا کلیدها با مقادیر ماکس و مین مشخص شوند:

```
//// Part4: section4
db.tweets.aggregate(
    [
        {
            "$unwind" : "$hashtags"
        },
        {
            "$group" : {
                _id : "$hashtags",
                "count" : {"$sum" : 1}
            }
        },
        {
            "$sort" : { "count" : -1}
        },
        {
            "$group" : {
                _id : "$hashtags",
                "maximun" : { "$first" : "$$ROOT"},
                "minimum" : { "$last" : "$$ROOT"}
            }
        }
    ]
)
```

زمان اجرای آن هم طبق عکس بالا 0.002 میباشد.

بخش پنجم:

```
//// Part4: section5
db.tweets.aggregate(
    [
        {
            "$match" : { "sendTimePersian" : {"$regex" : "1400/02/04 .*:.*"}}
        },
        {
            "$unwind" : "$hashtags"

            /* Deconstructs an array field from the input documents
            to output a document for each element. Each output document is the input document
            with the value of the array field replaced by the element. */
        },
        {
            "$group" : {
                _id : "$hashtags",
                "count" : {"$sum" : 1}
            }
        },
        {
            "$sort" : { "count" : -1},
        },
        {
            "$limit" : 10
        }
    ]
)
```

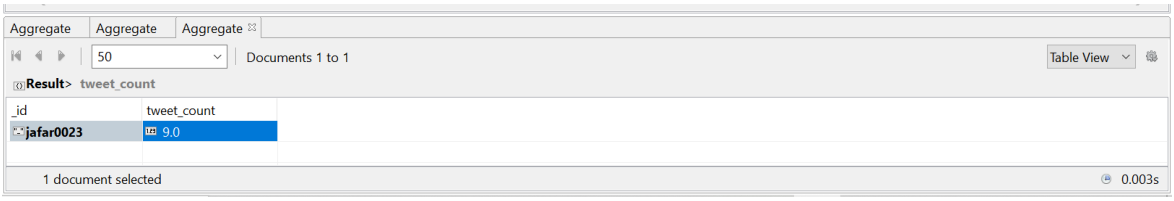
این کووری مشابه کووری‌های قبلی بوده با این تفاوت که ابتدا صرفا همه عناصری که در تاریخ ۴ اردیبهشت بوده‌اند را در نظر گرفته‌ایم و در آخر هم نتایجی که حاصل شده را صرفا ۱۰ نتیجه اول را خروجی داده‌ایم.

زمان اجرای کووری نیز 0.004 می‌باشد.

بخش ششم:

```
//// Part4: section6
db.tweets.aggregate(
    [
        {
            "$match" : { "sendTimePersian" : {"$regex" : "1400/02/04 .*:.*"}}
        },
        {
            "$unwind" : "$hashtags"

            /* Deconstructs an array field from the input documents
            to output a document for each element. Each output document is the input document
            with the value of the array field replaced by the element. */
        },
        {
            "$group" : {
                _id : "$senderUsername",
                "tweet_count" : {"$sum" : 1}
            }
        },
        {
            "$sort" : { "tweet_count" : -1},
        },
        {
            "$limit" : 1
        }
    ]
)
```

این کووری مشابه کووری قبلی بوده با این تفاوت که چون فعال‌ترین کاربر را می‌خواهیم محدوده را به ۱ عوض کرده و در گروه‌بندی هم طبق یوزرنیم افراد گروه‌بندی می‌کنیم. نتیجه در شکل زیر قابل مشاهده است:



زمان اجرای کووری نیز 0.003 می‌باشد.

مشکلات و توضیحات تکمیلی

برخی از مواردی که در داکیومنتیشن مونگو وجود داشتند در زدن کووری با 3T با مشکل مواجه شده و نیاز به اندکی تغییر داشتند.

آنچه آموختم

کار کردن با کووری‌ها در مونگو