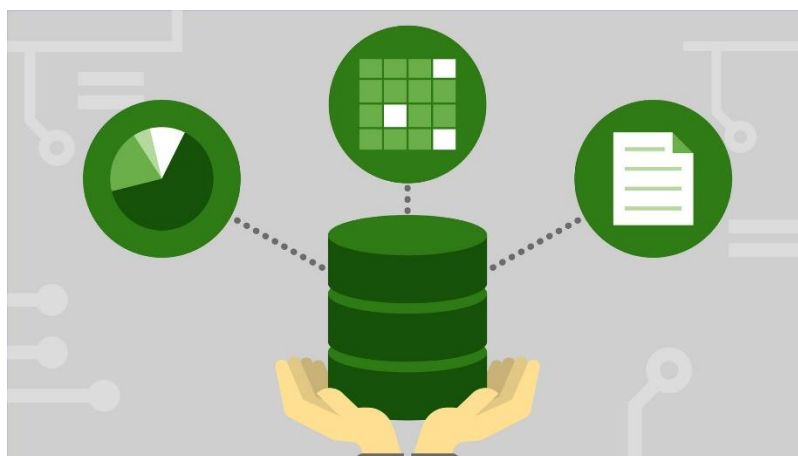


به نام خدا



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر



آزمایشگاه پایگاه داده

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 > ...
1 import time
2 import requests, json
3 from pymongo import MongoClient
4
5 client = MongoClient()
6 db = client.tweets
7 url = 'https://www.sahamyab.com/guest/twiter/list?v=0.1'
8 count_needed, sleep_time = 1000, 60
9 count_sofar = db.tweets.count_documents({})
10
11 while count_sofar < count_needed:
12     response = requests.request('GET', url, headers={'User-Agent': 'Chrome/61'})
13     result = response.status_code
14     if result == requests.codes.ok:
15         data = response.json()['items']
16         for d in data:
17             try:
18                 db.tweets.replace_one({"id": d["id"]}, d, upsert=True)
19             except Exception as e:
20                 print("Upsert exception: " + str(e))
21             count_sofar = db.tweets.count_documents({})
22     else:
23         print("Response code error: " + str(result))
24         print(f'Count of fetched tweets is {count_sofar}')
25         time.sleep(sleep_time)
26
27 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 هر بار 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 هشتک‌های 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 هشتک 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": "دارد.و شخصا انتظار دارم سهم از یکی دو روز آینده راه سبز برایش میسر شود\nerکت",
  "lastLikeNickName": "آرمان1500",
  "likeCount": "11",
  "retwitCount": "3",
  "type": "retwit",
  "scoredPostDate": "1619212540075",
  "retwitId": "273730567",
  "finalPullDatePersian": "",
  "hashtags": ["برکت"]
}
```

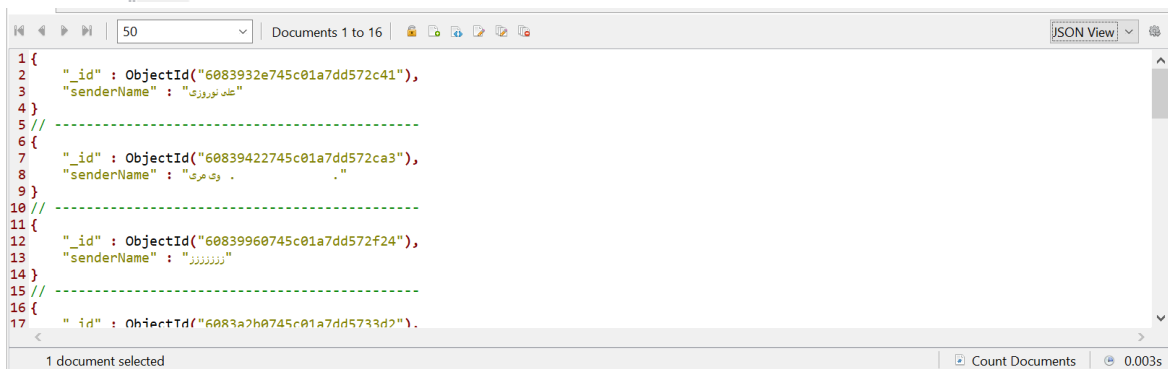
گام سوم:

بخش اول:

```

1 // Part3: section1
2 db.tweets.find(
3   {"mediaContentType":"image/jpeg"},[
4     "parentId":{"$exists" : true}},
5   {"senderName":1})
6

```



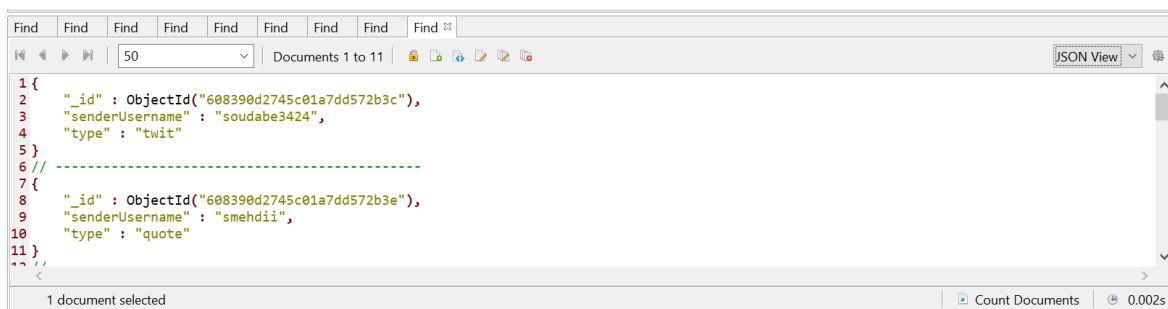
زمان هم همان طور که در گوشه تصویر بالا معلوم است 0.003 می باشد.

بخش دوم:

```

3 db.tweets.find(
4   {"sendTimePersian":{"$gt":"1400/02/04 07:45", "$lt":"1400/02/04 08:00"}},
5   {"senderUsername":1,
6     "type":1})
7

```



همان طور که در شکل بالا مشخص است حدود ۱۱ نفر سحرخیز! داریم که بین بازه ۷.۴۵ تا ۸ صبح توییت کرده اند و زمان اجرا هم 0.002 می باشد.

بخش سوم:

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

```

1 // Part3: section3
2 db.tweets.find(
3   {"sendTimePersian":{"$regex":".*/.*/.{1,2} 08:.*"}},
4   {"senderUsername":1})
5

```

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

_id	senderUsername
60839187745c01...	mostaghni
608391c3745c01...	ambna2020
608391ff745c01...	jafrinajm
60839277745c01...	hse_ms
60839277745c01...	bday
60839277745c01...	eliyekta
60839277745c01...	aliakbar_mz
60839277745c01...	behnam515
608392f1745c01...	osm1361
608392f1745c01...	miladrmz2020
6083932e745c01...	mohamadali1979
6083936b745c01...	roustapour
608393a8745c01...	sm172818

گام چهارم:

بخش اول:

```

7 db.tweets.aggregate(
8   [
9     {
10      "$group" : {
11        "_id" : "$senderUsername",
12        "count" : {"$sum" : 1}
13      }
14    },
15    {
16      "$group" : {
17        "_id" : {
18          "$cond" : {[
19            "if" : {"$lt" : ["$count",2]},
20            "then" : "First group",
21            "else" : {
22              "$cond" : {
23                "if" : {"$lte" : ["$count",3]},
24                "then" : "Second group",
25                "else" : {
26                  "$cond" : {
27                    "if" : {"$gte" : ["$count",4]},
28                    "then" : "Third group",
29                    "else" : "No group!"
30                  }
31                }
32              }
33            }
34          }
35        },
36        "count" : {"$sum" : 1}
37      }

```

Documents

Documents

Aggregate

50

Documents 1 to 3

Table View

Result

count

id	count
Third group	47.0
First group	409.0
Second group	156.0

1 document selected

0.006s

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

بخش دوم:

```

58 //// Part4: section2
59 db.tweets.aggregate(
60   [
61     {
62       "$unwind" : "$hashtags"
63     },
64     /* Deconstructs an array field from the input documents
65      to output a document for each element. Each output document is the input document
66      with the value of the array field replaced by the element. */
67   ],
68   {
69     "$group" : {
70       "_id" : "$hashtags",
71       "count" : {"$sum" : 1}
72     }
73   },
74   {
75     "$sort" : { "count" : -1}
76   }
77 ]
78 )

```

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 داده ایم که نتیجه در شکل زیر قابل مشاهده است:

_id	count
برکت	143.0
شاخص بورس	64.0
خودرو	46.0
کاما	22.0
شینا	21.0
گشان	20.0
خسایا	17.0
پالایش	15.0
خمصرکه	15.0
ستران	15.0
فملی	14.0
شستا	14.0
ویملت	14.0
سمگا	12.0
پترول	12.0
شپلی	12.0
کرمان	10.0
غگا	9.0

زمان اجرای کووری نیز 0.004 می باشد.

بخش سوم:

```

8  ///// Part4: section3
9  db.tweets.aggregate([
10   {
11     "$match" : {
12       "parentId" : {"$exists" : true}
13     }
14   },
15   {
16     "$unset" : "type"
17   }
18 ])

```

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	وینصادر صیدی دو دلاری کجاست خبری ارزش نیست	String
senderName	Mehdi	String
senderUsername	smehdii	String
senderProfileImage	default	String
content	بجاش صیدی یک ستنی اومده	String
finalPullDatePersian		String
hashtags	[0 elements]	Array
(2) { _id : 608390d2745c01a7dd572b4c }	{ 17 fields }	Document

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

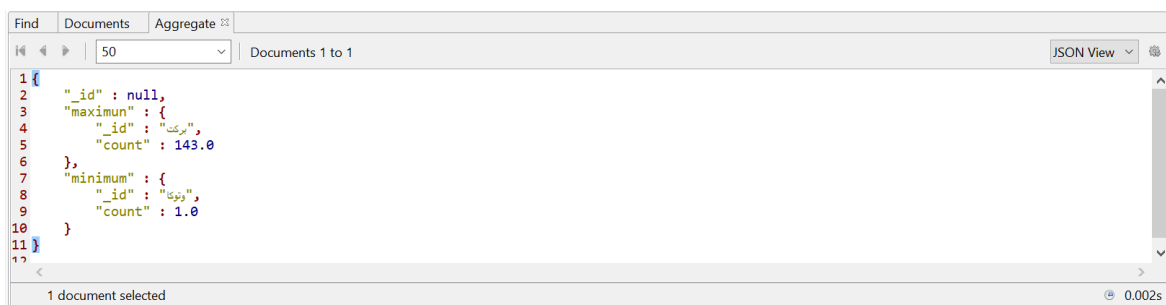
بخش چهارم:

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

```

13 //// Part4: section4
14 db.tweets.aggregate(
15   [
16     {
17       "$unwind" : "$hashtags"
18     },
19     {
20       "$group" : {
21         "_id" : "$hashtags",
22         "count" : {"$sum" : 1}
23       }
24     },
25     {
26       "$sort" : { "count" : -1}
27     },
28     {
29       "$group" : {
30         "_id" : "$hashtags",
31         "maximum" : { "$first" : "$$ROOT"},
32         "minimum" : { "$last" : "$$ROOT"}
33       }
34     }
35   ]
36 )
37

```



زمان اجرای آن هم طبق عکس بالا 0.002 میباشد.

بخش پنجم:

```

2
3 //// Part4: section5
4 db.tweets.aggregate(
5   [
6     {
7       "$match" : { "sendTimePersian" : {"$regex" : "1400/02/04 .*:.*"}}
8     },
9     {
10      "$unwind" : "$hashtags"
11
12      /* Deconstructs an array field from the input documents
13      to output a document for each element. Each output document is the input document
14      with the value of the array field replaced by the element. */
15    },
16    {
17      "$group" : {
18        _id : "$hashtags",
19        "count" : {"$sum" : 1}
20      }
21    },
22    {
23      "$sort" : { "count" : -1},
24    },
25    {
26      "$limit" : 10
27    }
28  ]
29 )

```

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

Aggregate	
Aggregate	
50	Documents 1 to 10
Table View	
Result	count
پرکت	124.0
شاخص بورس	59.0
خودرو	45.0
کاما	22.0
گشان	20.0
شپنا	20.0
ستران	15.0
خمجرکه	15.0
پالایش	15.0
خسایا	15.0

زمان اجرای کووری نیز 0.004 می‌باشد.

بخش ششم:

```

.62 //// Part4: section6
.63 db.tweets.aggregate(
.64   [
.65     {
.66       "$match" : { "sendTimePersian" : {"$regex" : "1400/02/04 .*:.*"}}
.67     },
.68     {
.69       "$unwind" : "$hashtags"
.70
.71       /* Deconstructs an array field from the input documents
.72        to output a document for each element. Each output document is the input document
.73        with the value of the array field replaced by the element. */
.74     },
.75     {
.76       "$group" : {
.77         "_id" : "$senderUsername",
.78         "tweet_count" : {"$sum" : 1}
.79       }
.80     },
.81     {
.82       "$sort" : { "tweet_count" : -1},
.83     },
.84     {
.85       "$limit" : 1
.86     }
.87   ]
.88 )

```

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:

Aggregate	Aggregate	Aggregate
50	Documents 1 to 1	Table View
Result> tweet_count		
_id	tweet_count	
jafar0023	9.0	
1 document selected		
0.003s		

زمان اجرای کووری نیز 0.003 می‌باشد.

مشکلات و توضیحات تکمیلی

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

آنچه آموختم

کار کردن با کووری ها در مونگو