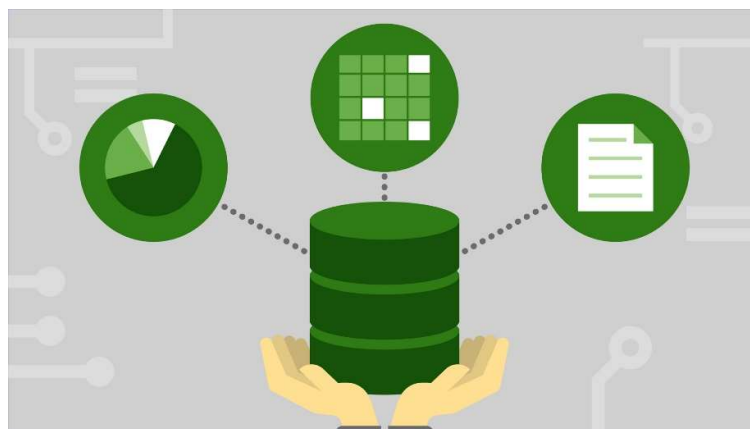


به نام خدا



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر



آزمایشگاه پایگاه داده

دستورکار شماره ۶

MongoDB

مهلت تحویل : ۱۴۰۰/۱۰/۰۱

مجتبی بنائی

دستور کار شماره ۶

هدف اصلی از این تمرین، آشنایی با مانگودی بی به عنوان یکی از رایجترین دیتابیس‌های غیررابطه‌ای (NoSQL) دنیاست.

در این تمرین، برای ذخیره توییت‌های سایت سهامیاب از مانگو استفاده میکنیم و بعد از ذخیره اطلاعات، با انجام چند پرس و جوی ساده، نحوه کار با این دیتابیس محبوب را فرا خواهیم گرفت.

پیش‌نیاز و شروع به کار با مانگودی بی

برای این منظور کافی است مقاله آشنایی با مانگو دی بی سایت مهندسی داده^۱ را که نصب و راه‌اندازی این دیتابیس هم توضیح داده شده است را مطالعه نموده، بخش پیاده سازی دیتابیس **retrogames** آنرا طبق دستوراتی که داده شده است در خط فرمان (پاورشل یا خط فرمان لینوکس) انجام دهید. سپس همین دستورات را در محیط **Robo3T** و یا **MongoDB Compass** اجرا کنید.

این بخش از کار، به عنوان پیش‌نیاز و دست‌گرمی محسوب می‌شود و نیاز به آوردن آن در گزارش نهایی دستور کار نخواهد بود.

دریافت اطلاعات

با استفاده از <https://www.sahamyab.com/guest/twiter/list?v=0.1> ده توییت آخر سایت سهامیاب با تمامی مشخصات را با فرمت جی‌سان دریافت میکنیم (با پستمن یا روش GET می‌توانید خروجی را تست کنید). توییت‌ها در فیلد **items** پاسخ، قابل مشاهده هستند. برای این تمرین، به کمک API فوق به جمع‌آوری و پردازش توییت‌های فارسی خواهیم پرداخت. برای دریافت اطلاعات می‌توانید از کد زیر استفاده کنید:

```

Import requests, time
url =
"https://www.sahamyab.com/guest/twiter/list?v=0.1"
delay = 60
while True:
    response = requests.request('GET', url,
headers={'User-Agent': 'Chrome/61'})
    if response.status_code == requests.codes.ok:
        items = response.json()['items']
        for item in items:
            print(item)
            time.sleep(60)

```

^۱ www.bigdata.ir/?p=214

نصب مانگو و ساخت کالکشن تویییت ها

مانگو دی بی را نصب کرده^۱ و کالکشن **tweets** را در دیتابیس **sahamyab** (این دیتابیس هم باید ایجاد شود) بسازید. میتوانید از خط فرمان مانگودی بی یا ابزارهای گرافیکی رایج مانند **MongoDB Compass**^۲ برای این منظور استفاده کنید.

کتاب کوچک **The Little MongoDB**^۳ می تواند راهنمای سریع شما برای کار با مانگو در این تمرین باشد.

گام اول تمرین

در این گام، با فراخوانی آدرس <https://www.sahamyab.com/guest/twitter/list?v=0.1> ده تویییت آخر را دریافت کرده و به صورت دستی در مانگو ذخیره کنید (از خروجی پستمن هم می توانید در این مرحله استفاده کنید و نیاز به کدنویسی نخواهد بود) و بررسی کنید چه فیلدهایی توسط خود مانگو به صورت خودکار به داده ها افزوده میشود. (هر تویییت را به عنوان یک داکومننت ذخیره کنید یعنی با فراخوانی کد فوق، ده تویییت را ذخیره خواهیم کرد.)

سپس با استفاده از کتابخانه **pymongo**^۴ کد دریافت اطلاعات فوق را به گونه ای تغییر دهید که هر یک دقیقه یکبار، تویییت های جدید را دریافت کرده و همزمان با دریافت تویییت ها، آنها را در مانگو هم ذخیره کند. (دقت کنید که هر تویییت باید جداگانه ذخیره شود و تویییت های تکراری بر اساس فیلد **id** هم باید حذف شوند که البته می توانید **upsert** کنید)

کد نوشته شده را تا زمانی اجرا کنید که حداقل ۵۰۰ تویییت منحصر بفرد در مانگو ذخیره شده باشند. با دستور **count**، مطمئن شوید که ۵۰۰ تویییت ذخیره شده باشد.

خروجی گام اول

نحوه ورود دستی داده ها در مانگو و فیلدهای اضافه شده، کدهای نوشته شده برای درج اطلاعات و نحوه اطمینان از درج ۵۰۰ تویییت در گزارش آورده شود.

گام دوم - پیش پردازش داده

در این گام با استفاده از **Regex** هشتگ های استفاده شده کاربر در فیلد **content** را پیدا کرده و سپس با استفاده از دستور **update** در فیلدی به نام **hashtags** به صورت **Array** ذخیره کنید.

خروجی گام دوم

دستور نوشته شده، خروجی و زمان اجرا

گام سوم - دستورات اصلی

۱. نام کاربرانی که **mediaContentType** تویییت آنها **image/jpeg** هستند و **parentId** آنها مقدار دارد را بیابید.

۲. **senderUsername** و **type** تویییت آنهایی که در یک بازه ۱۵ دقیقه ای دلخواه (از بازه تویییت های دریافتیتان)

تویییت فرستاده اند را بیابید.

^۱<https://bit.ly/2XWSqM7>

^۲<https://www.mongodb.com/products/compass>

^۳<https://openmymind.net/mongodb.pdf>

^۴<https://pymongo.readthedocs.io/en/stable/tutorial.html>

3. قصد داریم به کسانی که در بازه ساعت نه تا ده صبح (یا یک ساعت خاص که داده‌های آن در کالکشن شما موجود باشد)، بیش از یک توییت کرده‌اند (به ازای یک روز خاص) جایزه بدهیم `senderName` و `senderProfileImage` این کاربران را بیابید.

خروجی گام سوم

دستور نوشته شده، خروجی و زمان اجرا

گام چهارم - دستورات تجمعی و آماری (Aggregate Functions)

1. می‌خواهیم کاربران را بر اساس فعالیتشان دسته بندی کنیم. کاربران را به سه دسته به صورت زیر تقسیم کنید:
کاربرانی با یک توییت، کاربرانی با دو تا سه توییت، کاربرانی با بیش از سه توییت
دستوری بنویسید که تعداد هر گروه را برگرداند.
2. تعداد توییت‌های هر هشتگ را بشمارید و به صورت نزولی رتبه بندی کنید.
3. برای توییت‌هایی که `parentId` دارند، فیلد `type` را حذف کنید.
4. پرتکرارترین و کم‌تکرارترین هشتگ را بیابید.
5. ده هشتگ (که معمولا همان نماد بورسی است) پر استفاده هر روز را بیابید. (بازه زمانی جزء ورودی‌های کوئری خواهد بود).
6. فعالترین کاربر هر روز را به همراه تعداد توییت‌های انجام شده، پیدا کنید.

خروجی گام چهارم

دستور نوشته شده، خروجی و زمان اجرا

برای انجام این دستورکار، می‌توانید از این راهنمای فارسی yun.ir/tqhhsa استفاده کنید.