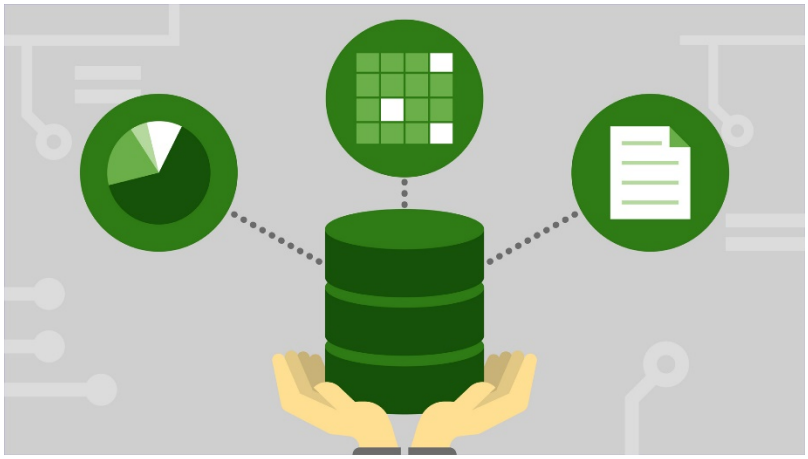
به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر



آزمایشگاه پایگاه‌داده

دستورکار شماره ۶

MongoDB

مهلت تحویل : ۱۴۰۱/۱۰/۳۰

مجتبی بنائی

دستور کار شماره ۶

هدف اصلی از این تمرین، آشنایی با **مانگودی‌بی** 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 فراخواهیم گرفت.

پیش‌نیاز و شروع به کار با مانگودی‌بی

در این قسمت ، به آشنایی با مانگودی‌بی به کمک فیلم های آموزشی سایت رسمی مانگودی‌بی به آدرس

https://university.mongodb.com/courses/M001/about

خواهیم پرداخت.

بنابراین، با مرور فیلم‌های آموزشی مانگودی‌بی و انجام بخشی از تمرینات انجام شده آن، از دستوراتی که نوشته‌اید و خروجی هایی که تولید کرده اید(انتخاب آنها بر عهده خودتان خواهد بود)، اسکرین شات گرفته ، برای هر یک، توضیحی کوتاه بدهید. (شش کوئری و خروجی آنها برای گزارش کار کافی است.)

دقت کنید که مجموع طول دوره آموزشی فوق، هشت ساعت است اما در همان حدی از فیلم‌ها را مشاهده و به صورت عملی کار کنید که آماده انجام ادامه‌ی دستور کار شوید. (هر چند توصیه می‌کنم شش فصل این دوره را حتما مشاهده و به صورت عملی انجام دهید)

نصب و راه‌اندازی مانگودی‌بی

در ادامه برای یادگیری نحوه‌ی نصب و راه‌اندازی دیتابیس مانگودی‌بی، مقاله آشنایی با مانگو دی بی سایت مهندسی داده[1] را مطالعه نموده و 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 پیش نیاز و دست گرمی محسوب می‌شود و نیاز به آوردن آن در گزارش نهایی دستورکار 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 جی سان دریافت میکنیم (با پستمن با 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 توئیت‌های فارسی خواهیم پرداخت.

برای دریافت اطلاعات میتوانید از کد زیر استفاده کنید:

[1] www.bigdata.ir/?p=214

```
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)
```

نصب مانگو و ساخت کالکشن توییت ها

مانگو دی بی را نصب کرده و کالکشن **tweets** را در دیتابیس **sahamyab** (این دیتابیس هم باید ایجاد شود) بسازید. میتوانید از خط فرمان مانگودی‌بی یا ابزارهای گرافیکی رایج مانند **MongoDB Compass**[1] برای این منظور استفاده کنید.

کتاب کوچک **The Little MongoDB**[2] می‌تواند 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/twiter/list?v=0.1 ده توییت آخر را دریافت کرده و **به صورت دستی در مانگو ذخیره کنید** (از خروجی پستمن هم می‌توانید در این مرحله استفاده کنید و نیاز به کدنویسی نخواهد بود) و بررسی کنید چه فیلدهایی توسط خود مانگو به صورت خودکار به داده‌ها افزوده میشود. (هر توئیت را به عنوان یک داکیومنت ذخیره کنید یعنی با فراخوانی کدفوق، ده توئیت را ذخیره خواهیم کرد.)

سپس با استفاده از کتابخانه **pymongo**[3] کد دریافت اطلاعات فوق را به گونه‌ای تغییر دهید که **هر یک دقیقه یکبار**، توئیت‌های جدید را دریافت کرده و همزمان با دریافت توییت‌ها، آنها را در مانگو هم ذخیره کند.(دقت کنید که هر توئیت باید جداگانه ذخیره شود و توییت‌های تکراری بر اساس فیلد id هم باید حذف شوند که البته می‌توانید upsert کنید)

کد نوشته شده را تا زمانی اجرا کنید که حداقل ۵۰۰ توئیت منحصربفرد در مانگو ذخیره شده باشند. با دستور count، مطمئن شوید که ۵۰۰ توییت ذخیره شده باشد.

خروجی گام اول

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

[1] https://www.mongodb.com/products/compass
[2] https://openmymind.net/mongodb.pdf
[3] https://pymongo.readthedocs.io/en/stable/tutorial.html

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

1. نام کاربرانی که **mediaContentType** توییت آنها **image/png** هستند و **parentId** آنها مقدار دارد را بیابید.
2. **senderUsername** و **content** توییت آنهایی که در یک بازه ۱۵ دقیقه‌ای دلخواه (از بازه توییت‌های دریافتیتان) توییت فرستاده‌اند را بیابید.
3. قصد داریم **senderName** و **senderProfileImage** کسانی که در یک روز خاص و یک بازه‌ی یک ساعته که داده‌های آن در کالکشن شما موجود باشد و بیش از یک توییت کرده‌اند را پیدا کنیم، 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 بر گرداند.
2. تعداد توییت های هر هشتگ را بشمارید و به صورت نزولی رتبه بندی کنید.
3. برای توییت‌هایی که **parentId** دارند، فیلد **parentContent** 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 ورودی های کوئری خواهد بود).
6. فعالترین کاربر هر روز را به همراه تعداد توئیت‌های انجام شده، پیدا کنید.

خروجی گام چهارم

دستور نوشته شده، خروجی و زمان اجرا