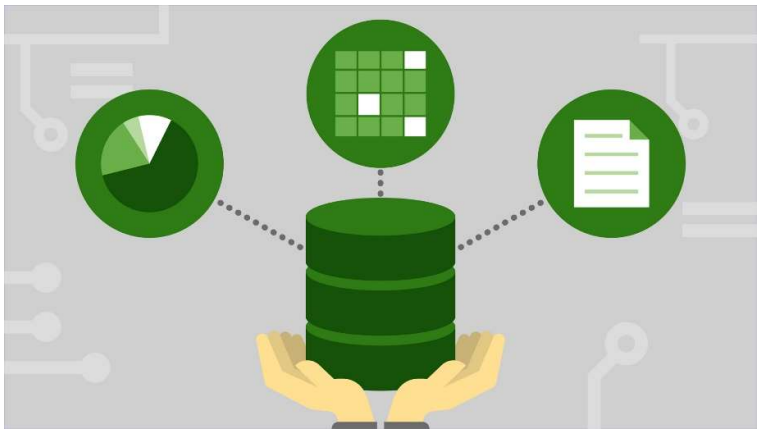
به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر



آزمایشگاه پایگاه‌داده

دستورکار شماره ۶

MongoDB

مهلت تحویل : ۱۴۰۰/۲/۳۱

مجتبی بنائی

دستور کار شماره ۷

هدف اصلی از این تمرین، آشنایی با **مانگودی‌بی** به عنوان یکی از رایجترین دیتابیس‌های غیررابطه‌ای (NoSQL) دنیاست.در قسمت اول که در حقیقت دستور کار شماره هفتم است، به آشنایی با مانگو، نحوه نصب و کار با آن خواهیم پرداخت.

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

برای این منظور کافی است مقاله آشنایی با مانگو دی بی سایت مهندسی داده[1] 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 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 در 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:

```
import requests, json
response = requests.get('https://www.sahamyab.com/guest/twiter/list?v=0.1', headers={'User-Agent': 'Chrome/61'})
data = json.loads(response.text)
```

نصب مانگو و ساخت کالکشن توییت ها

مانگو دی بی را نصب کرده[2] 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**[3] برای این منظور استفاده کنید.

علاوه بر کتاب درسی معرفی شده، کتاب کوچک **The Little MongoDB**[4] می‌تواند راهنمای سریع شما برای کار با مانگو در این تمرین باشد.

[1] www.bigdata.ir/?p=214
[2] https://bit.ly/2XWSqM7
[3] https://www.mongodb.com/products/compass
[4] https://openmymind.net/mongodb.pdf

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**[1] 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 هزار توئیت منحصر بفرد در مانگو ذخیره شده باشند. با دستور count، مطمئن شوید که هزار 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 هزار توییت 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

1. نام کاربرانی که **mediaContentType** توییت آنها **image/jpeg** هستند و **parentId** آنها مقدار دارد را بیابید.
2. **senderUsername** و **type** توییت آنها که در یک بازه ۱۵ دقیقه‌ای دلخواه (از بازه توییت‌های دریافتیتان) توییت فرستاده‌اند را بیابید.
3. قصد داریم به کسانی که در بازه ساعت نه تا ده صبح، توییت کردند جایزه بدهیم **senderName** و **senderProfileImage** این کاربران را بیابید.(در صورت نبود توییت در این بازه، بازه دلخواه دیگری انتخاب کنید.)

خروجی گام سوم

دستور نوشته شده، خروجی و زمان اجرا

[1] https://pymongo.readthedocs.io/en/stable/tutorial.html

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 type را حذف کنید.
4. پرتکرارترین و کم‌تکرارترین هشتگ را بیابید.
5. ده هشتگ 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

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