



مبانی بازیابی اطلاعات و جستجوی وب

نیم سال دوم ۱۴۰۳-۱۴۰۲

تمرین ۴

اهداف:

آشنایی با مفاهیم Elasticsearch و Kibana

نمایه سازی و تجسم داده‌ها در Elasticsearch و Kibana

پیاده سازی موتور جست‌وجو پیشرفته

پیاده سازی سیستم توصیه‌گر

مهلت تحویل:

چهارشنبه ۱۰ تیر ۱۴۰۳

توضیحات تمرین

به طور کلی در ادامه می‌خواهیم مباحث زیر را مورد بررسی قرار دهیم:

مباحث: خزش داده، نمایه سازی داده ها در الستیک ، ساخت پرس و جوها ، تجسم داده ها در کیبانا، پیاده سازی موتور جستجو و نهایتا پیاده سازی سیستم توصیه گر

قسمت اول - خزش داده(Data Crawling):

در این تمرین می‌خواهیم از خروجی داده های بخش قبلی(Crawling) به عنوان ورودی این بخش استفاده کنیم، پس خروجی قسمت قبل به فرمت درست ذخیره کرده تا بتوانیم از آنها به عنوان ورودی الستیک استفاده کنیم.

قسمت دوم - نمایه سازی داده ها در الستیک (Indexing Data into Elasticsearch):

برای انجام نمایه سازی داده های ورودی به الستیک راه های مختلفی وجود دارد که می‌توانید استفاده کنید:

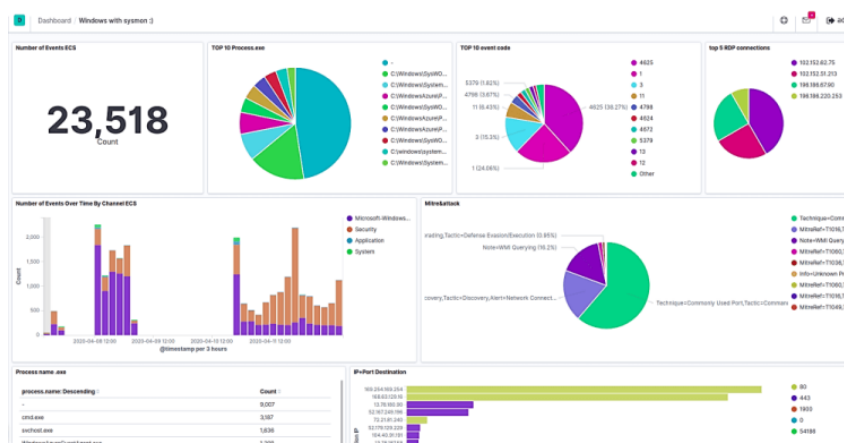
- استفاده از Elasticsearch API، نقشه ها(mappings) را تعریف کنند، اسناد فهرست بندی کنند و عملیات اولیه CRUD را انجام دهند.

- همچنین می‌توانید از کد insert.py که در سامانه ویو بارگذاری شده است و نحوه استفاده از آن را در تمرین کلاسی فرا گرفته اید بهره ببرید.

تجسم داده ها در کیبانا(Visualizing Data in Kibana):

به واسطه قابلیت های موجود در کیبانا الگوهای شاخص را ایجاد کرده، داده ها را با استفاده از انواع مختلف نمودارها (مانند نمودار میله ای، نمودار خطی، نمودار دایره ای) که در داشبورد کیبانا وجود دارد تجسم کنید و در نهایتا ساخت داشبورد برای ارائه نتایج جستجو.

*حداقل سه حالت مختلف برای تجسم داده های وارد شده و نتایج جستجو را در کیبانا پیاده سازی کنید.



تصویر ۱: نمونه ای از داشبورد کیبانا

ساخت پرس و جوها (Building Queries):

به منظور ساخت پرس و جو دو رویکرد مد نظر است که انتظار می‌رود هر دو روش را پیاده سازی کنید:

۱. ساخت پرس و جوی (query) به صورت دستی به واسطه یکی از زبان های برنامه نویسی و ذخیره سازی خروجی ها
۲. ایجاد پرس و جو در کیبانا، ارائه نتایج در داشبورد ایجاد شده بخش پیشین و ذخیره سازی خروجی ها

```
4 queries = [  
5   {  
6     "match": {  
7       "reviews": "awesome"  
8     }  
9   },  
10  {  
11    "range": {  
12      "price": {  
13        "gte": 50,  
14        "lte": 60  
15      }  
16    }  
17  },
```

تصویر ۲: نمونه ای از ایجاد پرس و جوی دستی در پایتون

قسمت سوم - پیاده سازی موتور جستجو (Search Engine Implementation):

در این بخش می‌بایست یک رابط وب اولیه بسازید که در آن کاربران می‌توانند پرس و جویهای جستجو را وارد کنند و نتایج جستجو را با استفاده از تجسم Kibana در داشبورد ساخته شده بخش پیشین نمایش دهید.

* در نظر داشته باشید که رابط اولیه شما می‌بایست مشابه با بخش Advanced Search در IEEE باشد.

Advanced Search ?

Advanced Search	Command Search	Citation Search
-----------------	----------------	-----------------

Enter keywords and select fields.

in

All Metadata

?

RS

AND

 in

All Metadata

↑

×

+

تصویر ۳: نمونه رابط وب اولیه به منظور پرس و جو

قسمت چهارم - پیاده سازی سیستم توصیه گر (Recommended System)

در این بخش شما می‌بایست با استفاده از داده های جمع آوری شده بخش پیشین، یک سیستم توصیه گر طراحی کنید، به این صورت که در رابط اولیه وب شما می‌بایست یک گزینه با عنوان RS وجود داشته باشد که با انتخاب آن یک فیلدی ورودی داشته باشید که می تواند عنوان یا DOI مقاله به عنوان ورودی از کاربر دریافت کرده و سپس یک فیلد برای انتخاب تعداد مقالات مشابه داشته باشد (شامل ۵، ۱۰، ۱۵، ۲۰، ۳۰) که متناسب با ورودی، مقالات مشابه با مقاله مد نظر را پیشنهاد می‌دهد.

راهنمایی: شما برای طراحی سیستم آزاد هستید اما به منظور پیشنهاد و به عنوان یکی از سادهترین روش‌ها می‌توانید تشابه بین عنوان، چکیده مقاله، کلمات کلیدی، کلمات کلیدی نویسنده مقاله انتخاب شده را بسنجید و آنها را گزارش کنید (استفاده از tf-idf و Cosine similarity)

شیوه تحویل

خروجی موردنظر برای هر قسمت تمرین شامل بخش‌هایی است که توضیحات آن به شرح زیر می‌باشد.
در بخش گزارش هر قسمت، اسامی اعضای گروه و مشارکت هر عضو که عددی بین ۱ تا ۱۰ هست را بنویسید.

گزارش

نوشتن فایل README که شامل توضیح کامل درباره نحوه انجام تمرین، کدهای نوشته شده، تصویر خروجی‌ها و همچنین تفسیر و تحلیل هر یک از بخش‌ها می‌باشد.

کد

فایلی با فرمت ipnby. که شامل کدهای شما می‌باشد.

در نهایت فایل مربوط به این تمرین را به صورت IR_1402-02_HW4_Group Number.zip که شامل موارد خواسته شده‌ی بخش گزارش و کد می‌باشد را آپلود کنید.

کافیست فقط یک نفر از اعضای تیم، فایل را آپلود کند.

موفق باشید.