

یا علیم

مبانی بازیابی اطلاعات و جستجوی وب

تمرین عملی اول

- پاسخ تمرین باید شامل بحث و بررسی نتایج باشد و صرف اشاره به کد یا نتایج کافی نیست.

۱- پیش پردازش متن و ذخیره سازی در الاستیک سرچ

در این تمرین، شما روی پردازش و ایندکس گذاری متون فارسی و انگلیسی در Elasticsearch کار خواهید کرد. این تمرین شامل دو بخش اصلی است:

۱. پیش پردازش متن: پردازش اولیه روی متون فارسی و انگلیسی با استفاده از کتابخانه های مناسب
۲. ایندکس کردن در Elasticsearch: ذخیره سازی متون پردازش شده و بررسی تغییرات آنها در Elasticsearch

مراحل انجام تمرین:

۱. جمع آوری داده:

هر گروه باید حداقل ۱۰ صفحه از ویکی پدیا (۵ صفحه فارسی و ۵ صفحه انگلیسی) را به صورت دستی انتخاب و ذخیره کند. دقت کنید که گروه ها نباید صفحات یکسانی را پردازش کنند بنابراین بدون مشورت با گروه های دیگر صفحات خود را انتخاب کنید.

۲. پیش پردازش متن:

متون باید پاک سازی و پردازش شوند (شامل حذف علائم نگارشی، تبدیل به حروف کوچک، حذف کلمات بی معنی (Stopwords) و ...)

می توانید از کتابخانه های پردازش متن مانند Hazm برای فارسی و NLTK برای انگلیسی استفاده کنید.

۳. ایندکس گذاری در: Elasticsearch

متون پردازش شده را در Elasticsearch ذخیره کنید. سپس با پرس و جوی مناسب در متن های ذخیره شده جستجو کنید. حداقل دو نوع پرس و جوی متفاوت را بررسی کنید

۴. بررسی داده های ذخیره شده در: Elasticsearch

بررسی کنید که آیا Elasticsearch هنگام ایندکس کردن، پردازش دیگری روی داده ها انجام می دهد یا خیر؟ حداقل دو مورد که می تواند انجام دهد را ذکر کنید.

کدهای مربوط به پیش پردازش متن (هم فارسی و هم انگلیسی)

فایل های متنی خام و پردازش شده

اسکرپت ایندکس گذاری در Elasticsearch

نتایج بررسی تغییرات در متن ذخیره شده در Elasticsearch