مبانی بازیابی اطلاعات و جست و جوی وب

تمرین شماره 4 - پرس و جو

در ادامه تمرین قبل، باید به کمک ابزار الستیک سرچ(elasticsearch)، بر روی خروجی خزشگر خود چند پرس و جو(query) انجام دهید. ابزار الستیک سرچ امکان تعریف پرس و جو را به کمک Query DSL یا Domain Specific Language که مبتنی بر JSON است، فراهم نموده است.

- راه اندازی

شما ابتدا باید از اینجا، فایل پیکربندی مربوطه را دانلود و سرویس الستیک سرچ را راه اندازی نمایید.

- پیش پردازش

در فرایند پیش پردازش داده ها مواردی چون تبدیل نوع داده ها، یکسان سازی کاراکترها، مدیریت مقادیر گمشده و ... را در نظر بگیرید(برای دوره هایی که هزینه دوره یا گواهینامه آنها رایگان است، مقدار 0 و برای مقادیر نامعلوم، میانگین مقادیر موجود را جایگزین کنید).

- نمایه گذاری

برای نمایه گذاری داده های خود می توانید از ماژول های پایتون(یا زبان مورد استفاده خود)، برای انجام فرایند نمایه گذاری در الستیک سرچ استفاده کنید.

- پرس و جو

در این تمرین باید به کمک Query DSL، پرس و جوهای زیر را انجام داده و نتایج را به دست آورید:

1. تعداد تمام دوره ها
2. دوره هایی که توسط edx در دسته بندی science منتشر شده اند + تعداد این دوره ها
3. دوره با گرانترین گواهینامه که سطح مبتدی داشته و به زبان انگلیسی تدریس شده است.
4. میانگین هزینه گواهینامه ها به ازای هر دسته بندی
5. پرتکرارترین زبان زیرنویس
6. فعال ترین ارائه دهنده(فعال ترین به این معناست که ارائه دهنده در میان تمامی ارائه دهندگان، بیشترین تعداد دوره را منتشر کرده است)
7. فعال ترین موسسه(فعال ترین به این معناست که موسسه در میان تمامی موسسات، بیشترین تعداد دوره را تدریس کرده است)
8. تعداد دوره های با قیمت گواهینامه بین 60 تا 90 دلار
9. سه دوره با بیشترین تعداد زبان زیرنویس در هر دسته بندی به صورت نزولی

برای ارسال پرس و جوها به الستیک سرچ، می توانید از ابزار postman جهت ارسال درخواست به سرویس در حال اجرای الستیک سرچ استفاده نمایید. همچنین می توانید از امکانات development tools ابزار کیبانا(kibana) برای ارسال پرس و جو استفاده کنید. راه اندازی کیبانا مشابه الستیک سرچ بوده و فایل پیکربندی آن، از اینجا در دسترس است.

- موارد تحویلی
  1. گزارش کار پروژه(شامل توضیح مراحل مختلف کار)

2. کدهای پیش پردازش و نمایه گذاری
3. پرس و جوهای DSL: جهت سهولت، هر پرس و جو را در یک فایل جداگانه با پسوند json ذخیره نمایید.
4. خروجی پرس و جوهای DSL: جهت سهولت، هر پرس و جو را در یک فایل جداگانه با پسوند json ذخیره نمایید.
5. فایل خروجی فاز خزشگر