گزارش کار نمایه گذاری

در این پروژه هدف نمایه گذاری تعداد زیادی داده مربوط به کورس‌های آموزشی میباشد.

نمایه گذاره

برای نمایه گذاری داده‌ها از اسم پلتفرم آن کورس استفاده میکنیم یعنی هر پلتفرم یک ایندکس در الاستیک سرچ دارد،
همچنین برای بهینه سازی سرعت در دسترسی به ارائه دهندگان از نمایه گذاری معکوس به ارائه دهنده استفاده میکنیم زیرا هر کورس میتواند چندین ارائه دهنده داشته باشد یعنی هر پرزنتر یک ایندکس از کورس‌ها در الاستیک سرچ دارد،
برای دسته بندی نیز از نمایه گذاره معکوس استفاده می کنیم زیرا هر کورس می تواند تعداد زیادی دسته بندی داشته باشد بنابراین برای هر دسته بندی یک ایندکس در الاستیک سرچ میسازیم،
و در نهایت برای زبان زیرنویس‌ها هم یک نمایه گذاری انجام میدهیم یعنی هر زبان زیرنویس یک ایندکس از کورس‌ها در الاستیک سرچ دارد.
به نظر میرسد که مقداری داپلیکیت دیتا در ایندکس‌ها وجود داشته باشد، به عنوان مثال یک کورس میتواند در ۱۰ ایندکس باشد، اما مهم‌ترین نکته این است که با این کار پرفورمنس بهتری برای دسترسی به نتایج به دست می‌آوریم.

پاکسازی داده

برای پاکسازی‌داده‌ها بعد از لود کردن دیتای اولیه با استفاده از پانداز ابتدا قیمت سرتیفیکیت‌هایی که مشخص نیستند را برابر میانگین قرار میدهیم،
همچنین کاراکترهایی که برای ذخیره‌سازی در الاستیک‌سرچ قابل قبول نیستند را به آندرلاین تغییر میدهیم.

درنهایت با استفاده از کتابخانه الاستیک‌سرچ پایتون کورس‌هارا در الاستیک‌سرچ اینسرت میکنیم که حدود یک و نیم ساعت طول میکشد.

کوعری‌ها

کوعری‌های خواسته شده را با استفاده از پایتون انجام داده و نتایج را ذخیره میکنیم.

ساختار پروژه

- Indexer.py

  برنامه ایندکس کننده دیتای تولید شده توسط خزشگر
- Query.py

  برنامه اجرا کننده کوعری‌های خواسته شده با نام گذاری از q1 تا q9، میتوانید خود کوعری‌ها و ایندکس‌هایی روی آن اجرا میشوند را در این فایل مشاهده کنید.
- Req.txt

  ریکوایرمنت‌های پروژه برای اجرا شدن.
- Config.py

  کانفیگ‌های مربوط به پروژه شامل آدرس الاستیک‌سرچ.
- Data_new.py

  آخرین دیتای ماین شده از خزشگر.
- crawler/courses_spyder.py

  خزشگری که دیتای اولیه کورس‌هارا استخراج میکند.
- results/
  - ?.json

    ریسپانس مربوط به هر کوعری با شماره ?
  - ?.query.json

    کوعری مربوط به هر سوال با شماره ?

افراد گروه

یوسف سلملیان
برنا کلهر
فرناز بنائیان
بهار عالمی
اقاقیا محمدی