

به نام خداوند بخشنده و مهربان



دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه نرم افزار

بازیابی پیشرفته اطلاعات

پروژه پایانی

فاز اول

استاد: دکتر الهام اسماعیلی

ارائه دهندگان:

امیر سرتیپی ۹۹۳۶۱۴۰۱۹

مهدی مالوردی ۹۹۳۶۴۴۰۱۲

فهرست مطالب

۳مقدمه
۳ drop_coulmns متد
۳delete_existing_index متد
۳ delete_stop_words متد
۳csv_reader_index متد

مقدمه

در این پروژه قصد داریم بر روی مجموعه‌ای از کتاب انگلیسی که اطلاعات در یک فایل csv ذخیره شده است را به کمک الستیک ایندکس کنیم. زبان استفاده شده برای کد نویسی زبان پایتون می‌باشد و به کمک کتابخانه‌ای که مربوط به الستیک می‌باشد با API های آن ارتباط برقرار می‌شود. پورتهی که الستیک بر روی آن اجرا می‌شود همان مقدار دیفالت (localhost:9200) می‌باشد. در ادامه به ترتیب روند کاری و متدهای داخل پروژه توضیح داده خواهند شد. پروژه و مستندات آن از طریق این [لینک](#) در بر روی گیت در دسترس می‌باشد.

متد drop_coulmns

با توجه به این که نیازی به ایندکس کردن صفحات HTML کتاب‌ها نیست این متد یک پیش پردازشی از داده‌ها را انجام می‌دهد و ستون ۵ ام دیتا ست که مربوط به متن HTML می‌باشد را از csv حذف کرده و در فایل جدید دیگری با نام books.csv می‌نویسد. پس از این عمل حجم قابل توجهی از فایل ورودی که ۲.۸ گیگابایت بود به ۷۲۰ مگ کاهش پیدا کرد.

ورودی‌های این تابع نام فایل csv که می‌خواهیم ویرایش کنیم و ورودی دوم نام فایل خروجی می‌باشد.

متد delete_existing_index

این متد در صورتی که ایندکسی با نام پارامتری که در ورودی دریافت می‌کند بر روی الستیک وجود داشته باشد، آن را حذف می‌کند.

متد delete_stop_words

برای دقت بیشتر موتور جستجویی که می‌خواهیم بسازیم نیاز است تا کلمات توقفی را حذف کنیم. این کلمات توقفی در کتابخانه nltk.corpus وجود دارند که با نصب آن می‌توان به آن دسترسی پیدا کرد. چون زمان حذف کلمات توقفی به کمک پایتون مقداری طولانی بود از اسپارک استفاده کرده و خروجی csv که از drop_coulmns دریافت کردیم را به آن دادیم تا کلمات توقفی از آن حذف بشود.

متد csv_reader_index

این متد با دریافت فایل ورودی و نام ایندکسی که قرار است ساخته شود در ابتدا یک شی از کلاس Elasticsearch ساخته و به آدرس localhost:9200 متصل می‌شود. سپس فایل csv را خوانده و با کمک تابع bulk به صورت دسته‌ای شروع به ایندکس کردن اطلاعات فایل ورودی می‌کند.

بر اساس شکل ۱ زمان اندازه گیری شده که در مشاهده می‌کنید عملیات ایندکس کرد تقریباً ۲ دقیقه و ۱۰ ثانیه به طول می‌انجامد.

```
done indexing in 130.319310665136  
done !
```

شکل ۱. مقدار زمان / ایندکس کردن داده‌ها