به‌نام 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

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 کتاب خانه‌ای که مربوط به الستیک می‌باشد با API های آن ارتباط برقرار می‌شود. پورتی 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 پیش‌پردازشی 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.

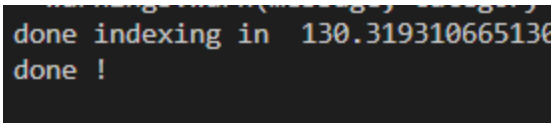
متد delete_stop_words

برای دقت بیشتر موتورجستجویی که می‌خواهیم بسازیم نیاز است تا کلمات توقفی را حذف کنیم. برای این کار از کتابخانه genism استفاده می‌کنیم که در داخل خود دارای لیستی از کلمات توقفی می‌باشد. ابتدا تکست را تماما به حروف کوچک تبدیل می‌کنیم. تابع remove_stopwords این کتابخانه یک متن را دریافت و کلمات توقفی را از آن حذف کرده و باز می‌گرداند. در نهایت اطلاعات پرداز شده در books_final.csv نوشته می‌شود. این عملیات نزدیک به ۳ دقیقه (۱۷۰) ثانیه طول کشید که ۱۶۴۵۱۸ رکورد را پردازش کرد. حجم فایل نهایی به نیز به ۴۰۰ مگ رسید.

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 تقریبا ۱دقیقه و ۲۰ ثانیه به طول می‌انجامد.



شکل ۱. مقدار زمان ایندکس کردن داده‌ها