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 993614019**

**مهدی مالوردی 993644012**

فهرست مطالب

[مقدمه 3](#_Toc68821644)

[متد drop\_coulmns 3](#_Toc68821645)

[متد delete\_existing\_index 3](#_Toc68821646)

[متد delete\_stop\_words 3](#_Toc68821647)

[متد csv\_reader\_index 3](#_Toc68821648)

# 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](https://github.com/amirsartipi13/advanced-IR) 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 5 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 2.8 گیگابایت بود به 720 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. این کلمات توقفی در کتاب خانه 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 1 زمان اندازه گیری شده که در مشاهده می‌کنید عملیات ایندکس کرد تقریبا 2 دقیقه و 10 ثانیه به طول می‌انجامد.



شکل 1. مقدار زمان ایندکس کردن داده‌ها