# دانشگاه اصفهان

## درس مبانی بازیابی اطلاعات و جستجوی وب

## فاز ۱ پروژه پایانی

**زمان تحویل: ۱۴۰۰/۱/۱۰** **اسفند ۱۳۹۹**

Wikibooks یک پروژه Wikimedia متعلق به بنیاد Wikimedia Foundation می‌باشد که با هدف ارائه کتاب‌های دیجیتالی و قابل ویرایش توسط کاربران توسعه داده شده است. در ابتدا (سال ۲۰۰۳) پروژه Wikibooks تنها با تمرکز بر روی کتاب‌های انگلیسی کار خود را شروع کرد و در ادامه و تا سال ۲۰۲۱ تعداد زبان‌های تحت پشتیبانی خود را به ۷۶ زبان مختلف افزایش داد.

در حال حاضر دادگان بسیار متعددی با اهداف مختلف مانند یادگیری ماشین[1]، پردازش زبان طبیعی[2]، معناشناسی[3] و غیره بر اساس محتوای موجود در Wikibooks ایجاد شده‌اند. یکی از این دادگان که به فرمت csv است داده‌های مربوط به کتاب‌های دیجیتالی به ۷ زبان مختلف شامل، انگلیسی، فرانسوی، آلمانی، پرتغالی، اسپانیایی، ایتالیایی و روسی می‌شود. هدف ما در این پروژه این است که بتوانیم با استفاده از مطالبی که در درس بازیابی اطاعات آموخته‌ایم یک موتور جستجو برای این دادگان توسعه دهیم. انجام این کار شامل چندین گام مختلف است که در هر گام پروژه ما قصد داریم خود را به این هدف نزدیک کنیم. قطعا نتایج جستجوی دقیق در این پروژه یک امتیاز ویژه به شمار می‌رود ولی هدف اول ما یادگیری نحوه استفاده از مفاهیم فراگرفته‌شده در کلاس در یک مسئله دنیای واقعی است. از این رو این پروژه بیشتر آموزش‌محور است تا نتیجه‌محور!

در ادامه شرح مختصری از دادگان مورد نظر ارائه شده است. ما در این پروژه تنها از بخشی از دادگان که به زبان انگلیسی است، استفاده می‌کنیم.

**فایل‌های دادگان (با فرمت csv):**

- en-books-dataset.csv

**اسکیما[4] این دادگان به صورت زیر است:**

**`title:`** عنوان کتاب

**`url:`** آدرس کتاب در سایت ویکی بوکز

[1] Machine Learning
[2] NLP: Natural Language Processing
[3] Semantic Detection
[4] Schema

**abstract:** چکیده کتاب در صفحه کتاب در ویکی بوکز

**body_text:** کتاب در قالب متن ساده

**body_html:** محتوای کتاب در قالب وب

این دادگان را می‌توانید از اینجا دانلود کنید.

## گام اول: شاخص‌گذاری کتاب‌ها

در این گام شما می‌باید هر کتاب را با استفاده از Elasticsearch شاخص‌گذاری کنید (به هنگام شاخص‌گذاری هر یک از کتاب‌ها باید stop words را حذف کنید که برای اینکار می‌توانید از Analyzerهای Elasticsearch استفاده کنید). Elasticsearch یک موتور جستجوی متن‌باز بر پایه Lucence است. این موتور جستجو یک Restful web service راه‌اندازی می‌کند که درخواست‌ها به این وب‌سرور ارسال می‌شوند. همچنین قابلیت توزیع گسترده و مقیاس‌پذیری سریع را نیز فراهم می‌کند. برای آشنایی بیشتر با این موتور جستجو، می‌توانید از این لینک استفاده کنید. شما برای ساخت شاخص، بعد از انتخاب نام شاخص، باید ابتدا یک شاخص خالی با نام books ایجاد کرده و سپس با ارسال درخواست‌های POST و یا PUT به وب‌سرور این موتور جستجو، کتاب‌ها را به شاخص اضافه کنید. همچنین می‌توانید از واسط‌هایی که در زبان‌های مختلف برای ارتباط با سرور Elasticsearch ایجاد شده است، برای ارتباط راحت‌تر با سرور از طریق فراخوانی توابع کتابخانه‌ای استفاده کنید. ذکر این نکته حائز اهمیت است که به هنگام شاخص‌گذاری نیازی به ذخیره فیلد `body_html` نیست و تنها باید از فیلد `body_text` استفاده کنید.

**تحویل دادنی‌ها:**

- سورس‌کد[5] برنامه نوشته شده که فایل دادگان به همراه آدرسی که سرور Elasticsearch در آن راه‌اندازی شده است (برای مثال localhost:9200) را گرفته و رکوردهای دادگان را را شاخص‌گذاری می‌کند.
- فایل مستندی که نحوه انجام کار در آن توضیح داده شده است.

**نکات:**

- زبان برنامه‌سازی برای پیاده‌سازی این پروژه می‌تواند هر زبانی باشد، اما اکیدا توصیه می‌شود که به دلیل وجود کتابخانه‌های آماده پردازش زبان طبیعی و منابع فراوان برای زبان‌های جاوا و پایتون، شما نیز از یکی از این دو زبان برای پیاده‌سازی این پروژه استفاده کنید.
- این پروژه می‌تواند در قالب تیم‌های دو نفره انجام شود. لذا همه دانشجویان باید در اسرع وقت اقدام به انتخاب هم‌گروهی‌های خود کنند و اطلاعات اعضای هر یک از گروه‌ها، توسط نماینده گروه برای بنده ایمیل شود.
- ارتباط با بنده از طریق آدرس ایمیل g.elhamesmaeeli@gmail.com امکان‌پذیر است.
- هرگونه تبادل نظر و همفکری با سایر گروه‌ها بلامانع است. اما تقلب ممنوع بوده و با گروه متقلب و تقلب‌شونده با کسر کامل نمره پروژه برخورد خواهد شد.

[5] Source code