

یا علیم

پروژه عملی دوم درس بازیابی اطلاعات و جستجوی وب

ترم دوم ۱۴۰۴

بخش اول: طراحی موتور جستجو و خزنده وب

در این پروژه باید یک موتور جستجوی کوچک فارسی برای سایت‌های خبری/آموزشی فارسی طراحی کنید.

برای حداقل یکی از سایت‌های زیر خزنده (Crawler) پیاده‌سازی کنید:

- دیجیاتو <https://digiato.com> —
- زومیت <https://www.zoomit.ir> —
- فرادرس بلاگ <https://blog.faradars.org> —

خزنده باید:

۱. از یک URL اولیه شروع کند.
۲. فقط صفحات داخل همان دامنه را دنبال کند.
۳. لینک‌های داخلی را استخراج نماید.
۴. حداقل 100 صفحه معتبر جمع‌آوری کند.

برای هر صفحه، اطلاعات زیر استخراج شود:

- عنوان صفحه
- متن اصلی محتوا
- تاریخ انتشار (در صورت وجود)
- دسته‌بندی/موضوع
- URL صفحه

برای پردازش HTML می‌توان از کتابخانه‌هایی مانند:

- BeautifulSoup
- Scrapy
- Selenium در صورت نیاز

استفاده کرد.

بخش دوم — ساخت Index و موتور جستجو

پس از جمع‌آوری داده‌ها:

دانشجو باید داده‌ها را ایندکس‌گذاری کند.

امکان استفاده از:

- ElasticSearch
- Whoosh
- پیاده‌سازی شخصی با TF-IDF

مجاز است.

سیستم جستجو باید حداقل قابلیت‌های زیر را داشته باشد:

- جستجوی تک‌کلمه‌ای
- جستجوی چندکلمه‌ای
- رتبه‌بندی نتایج
- نمایش Top-K نتایج

خروجی هر نتیجه شامل موارد زیر باشد:

- عنوان سند
- بخشی از متن (Snippet)
- امتیاز شباهت
- لینک صفحه

بخش سوم — بهبود رتبه‌بندی

برای بهبود کیفیت بازایی، دانشجو می‌تواند وزن‌دهی بخش‌های مختلف صفحه را اعمال کند.

مثال:

- کلمات موجود در **Title** وزن بیشتر داشته باشند.

- کلمات موجود در **Heading** وزن متوسط داشته باشند.
- متن اصلی وزن پایه داشته باشد.

نتایج جستجو برای چند Query نمونه گزارش شود.

گزارش نهایی پروژه

گزارش باید شامل موارد زیر باشد:

۱. معماری خزنده
۲. روش استخراج داده
۳. نحوه ایندکس گذاری
۴. الگوریتم رتبه بندی
۵. چند نمونه Query و نتایج حاصل
۶. چالش های انجام پروژه