دانشکده مهندسی کامپیوتر

پروژه درس

مبانی بازیابی اطلاعات و جستجوی وب

استاد درس: دکتر رهائی

طراحان پروژه: صالح شیروانی - سید ماکان حاجی سید جوادی

نیم‌سال دوم

سال تحصیلی ۱۴۰۳-۱۴۰۴

اهداف

در این پروژه میخواهیم یک مجموعه داده از انواع اخبار در سراسر جهان در شبکه اجتماعی Reddit و سایتی دلخواه، به کمک دو تکنیک Crwaling و Scrapping جمع‌آوری کنیم و سپس آنها را به روش‌های پیشنهاد شده دسته بندی کنیم. همچنین قصد داریم پروژه را به نحوی پیاده‌سازی کنیم که تا حد امکان مقیاس‌پذیر باشد. استفاده از ابزار هوش مصنوعی برای کمک در کدنویسی کاملا آزاد است، اما باید بر مراحل انجام کار و کل کد مسلط باشید.

توضیحات پروژه

در این پروژه باید یک سامانه برای گردآوری خبر بسازید که داده‌ها را فقط از طریق اسکرپینگ HTML و crawling وب جمع‌آوری می‌کند. منبع اول پست‌های مرتبط با خبر در Reddit است (با رعایت کامل خط مشی‌ها و خواندن فایل robots.txt، جهت جلوگیری از مسدود شدن اکانت.) و منبع دوم وبسایت‌های خبری که باید با crawling به صفحات خبر برسید. خروجی خام باید یکتا و تمیز در قالب CSV تولید شود تا روی آن پیش‌پردازش (نرمال‌سازی، توکن‌سازی کلمه و stemming) انجام شود و در پایان برای هر رکورد برچسب موضوعی پیش‌بینی گردد. اسکرپر باید بصورت عمودی مقیاس‌پذیر باشد، مدیریت استثناها (کدهای متفاوت خطا) باید رعایت شود. (نکته حائز اهمیت این است که به هیچ عنوان با اکانت اصلی خود داده جمع‌آوری نکنید تا از مسدود شدن اکانت شخصی خودتان جلوگیری شود.)

۱. **اسکرپ Reddit:** صرفاً با خواندن HTML صفحه‌ها، پست‌های خبری را از چند ساب‌ردیت مرتبط استخراج کنید و برای هر پست شناسهٔ پایدار مبتنی بر ساختار/URL صفحه، عنوان، متن/خلاصه، پیوند، زمان انتشار، در صورت دسترس امتیاز و شمار نظرات و همچنین نام نویسنده را ذخیره نمایید. (میتوانید اطلاعات دلخواه را جمع‌آوری کنید.) استفاده از API ممنوع است؛ در عوض با User-Agent شفاف، تاخیرهای تصادفی، کنترل نرخ و دنبال‌کردن ایمن تغییرمسیرها کار کنید و با حذف تکراری‌ها و Canonicalization خروجی تمیز بسازید. رعایت robots.txt و خط‌مشی‌های Reddit الزامی است. **(۲۰ نمره)**

۲. **کرال وب خبری:** چند دامنه خبری عمومی یا تخصصی را به‌عنوان بذر انتخاب کنید و کرالری بنویسید که تا صفحات خبر پیش برود و متن اصلی را استخراج کند. پیوندهای کشف‌شده را در صف مرکزی نگه دارید، URL ها را Cannonical و موارد تکراری را حذف کنید، برای عمق کرال سقف بگذارید تا

از تله‌هایی مانند آرشیوهای بی‌پایان دوری کنید. احترام به robots.txt و هر crawl-delay اعلام‌شده ضروری است. **(۲۰ نمره)**

۳. **ذخیره‌سازی و الگوی CSV:** برای هر رکورد نوع و نام منبع (reddit/news و ساب‌ردیت/دامنه)، URL، شناسهٔ یکتا، عنوان، متن، نویسنده (در صورت وجود)، زمان انتشار و در صورت امکان سیگنال‌های کیفی مانند امتیاز و شمار نظرات را ثبت کنید. پس از پیش‌پردازش ستون‌های زبان و شمار توکن‌ها را بیفزایید و در پایان ستون برچسب(های) پیش‌بینی‌شده را اضافه نمایید. **(۵ نمره)**

۴. **پیش‌پردازش متن:** متن‌ها را نرمال‌سازی کنید (حذف متن‌های تکراری، یکنواخت‌سازی فاصله‌ها و حروف، و برای انگلیسی *lowercasing* و حذف ایموجی/URL در صورت نیاز)، سپس در سطح کلمه توکن‌سازی و بعد *lemmatization/stemming* مناسب زبان را اعمال کنید. شمار توکن‌ها را به‌صورت یک ستون جداگانه ذخیره کنید و در صورت اختلاط زبان‌ها، مرحلهٔ تشخیص زبان را اضافه نمایید. (البته این مرحله ضروری نیست) **(۱۵ نمره)**

۵. **دسته‌بندی اخبار:** برای دسته‌بندی اخبار تعدادی دسته ثابت تعیین کنید و از دیتاست Reuters-۲۱۵۷۸ متن‌ها را استخراج کنید. همه خبرها را (پس از نرمال‌سازی، توکن‌سازی و stemming) با TF-IDF روی unigram و bigram بردارسازی کنید، سپس برای هر دسته یک لغت‌نامه وزندار بسازید: وزن n-gram ها را از TF-IDF خبرهای همان دسته جمع/میانگین بگیرید و top-K را نگه دارید. برای هر خبر، امتیاز هر دسته را با جمع وزن n-gram های مشترک محاسبه کنید و برچسب نهایی را با بیشینهٔ امتیاز بدهید؛ خروجی برچسب‌ها و امتیازها را در CSV ذخیره کنید. **(۲۵ نمره)**
لینک دسترسی به دیتاست مورد نظر:

```
https://www.daviddlewis.com/resources/testcollections/reuters21578/
```

۶. **مقیاس‌پذیری عمودی (فقط اسکرپر):** لایه اسکرپر را روی یک ماشین بصورت موازی مقیاس دهید تا توان عملیاتی افزایش یابد و زمان پردازش کاهش پیدا کند. صف باید امن و محدود باشد، سربار قفل‌ها حداقلی بماند و با back-pressure و throttling تعادل تولید و مصرف حفظ شود. با کلید یکتا و سیاست‌های بازاجرا (checkpoint) از تکرار داده در قطع/وصل جلوگیری کنید. **(۵ نمره)**

۷. **مدیریت استثنا و محدودیت نرخ:** خطاهای HTTP (۳xx/۴xx/۵xx)، timeouts و خطاهای کتابخانه‌ای نباید اجرای سامانه را متوقف کنند؛ تلاش مجدد با وقفهٔ نمایی، دنبال‌کردن تغییر مسیر، زمان‌سنجی‌های محافظه‌کارانه و ثبت لاگ ضروری است. در نزدیک‌شدن به سقف محدودیت‌ها سرعت را کاهش دهید و حداقل فاصلهٔ بین درخواست‌ها را رعایت کنید. **(۱۰ نمره)**

۸. **رعایت سیاست‌ها و نکات اجرایی:** گردآوری‌ها باید مطابق robots.txt و خط‌مشی‌های Reddit و سایت‌های خبری انجام شود؛ User-Agent واضح تنظیم کنید، نرخ درخواست‌ها را محدود نگه دارید و از مسیرهای ممنوع اجتناب کنید. از اکانت اصلی برای اسکرپ استفاده نکنید تا ریسک مسدود شدن را کاهش دهید. استفاده از ابزارهای هوش مصنوعی برای کمک در کدنویسی آزاد است، اما دانشجو باید بر کد مسلط باشد و درهنگام ارائه از جزئیات پیاده‌سازی دفاع کند.

نحوه تحویل

- **مهلت تحویل:** ۲۰ شهریور.
- **پیاده‌سازی:** حتماً با زبان پایتون انجام شود و استفاده از ابزارهای هوش‌مصنوعی آزاد است.
- **ترکیب تیم:** پروژه می‌تواند به‌صورت انفرادی یا در گروه‌های دو نفره انجام شود.
- **روش تحویل:** تمامی فایل‌های لازم را در قالب یک فایل zip با فرمت زیر روی سامانه LMS بارگذاری کنید.

`FirstName_LastName-StudentID-FinalProject.zip`

- **طریقه تحویل:** ارائهٔ نهایی به‌صورت مجازی و بر بستر Google Meet برگزار می‌شود.