به نام خالق زیبایی‌ها

مرحله سوم پروژه درس ذخیره و بازیابی اطلاعات

نیم سال دوم سال تحصیلی ۹۴ -۹۳

تعریف پروژه:

در این مرحله از پروژه در ادامه‌ی فازهای قبلی قرار است عملیاتی بر روی مقالات صورت پذیرد تا بتوان با سرعت و دقت بالا عملیات جست و جو را در بین مقالات انجام داد.

در این مرحله شما باید در ابتدا مقالات را به صورت یک بردار نمایش دهید که نحوه دقیق این کار در ادامه توضیح داده شده است. سپس عملیات خوشه بندی را بر روی مقالات انجام داده و درخواست‌های کاربر را در بین مقالات جست و جو کرده و نتایج مناسب را بازگردانید.

هدف از خوشه بندی آن است که درخواست‌های کاربر را با تعداد کمتری از مقالات مقایسه کرده و در نتیجه سرعت جستجوی خود را افزایش دهید. الگوریتم پیشنهادی برای خوشه بندی اسناد الگوریتم kmeans است.

همانند مراحل قبلی پروژه در بخش‌های مختلف پروژه می‌توانید ایده‌های جدیدی به کار ببرید و نتایج به دست آمده را ارزیابی و گزارش نمایید.

۱. **ایجاد شاخص وارون لغت-سند**

علاوه بر خروجی‌های فاز دوم پروژه باید برای هر کلمه نیز برداری ایجاد کنید که این بردار وقوع آن کلمه در اسناد مختلف را نمایش می‌دهد. نحوه ذخیره‌سازی postingها در حافظه اصلی و جانبی در پروژه خود را شرح دهید. حداقل، حداکثر و متوسط طول postingها را محاسبه و گزارش نمایید. حجم فایل ذخیره‌سازی postingها را با حجم فایل اولیه اسناد مقایسه کنید.

۲. **به دست آوردن وزن هر کلمه در هر مقاله و نمایش آن مقاله به صورت یک بردار:**

برای بدست آوردن وزن کلمه iام در مقاله j ام از رابطه زیر استفاده کنید:

$$W_{i,j} = \frac{n_{i,j}}{n_{max,j}} * \log_2 \frac{N}{n_i}$$

$n_{i,j}$: تعداد تکرار کلمه i ام در مقاله j ام

$n_{max,j}$: تعداد بیشترین تکرار یک کلمه در مقاله j ام

N: تعداد کل اسناد

n_i تعداد اسنادی که کلمه i در آنها آمده است.

برای بدست آوردن پارامترهای بالا از شاخص‌های سند-لغت و لغت سند پیاده سازی شده در قسمت‌های قبلی استفاده کنید و برداری از وزن لغات برای هر کدام از مقاله‌ها طبق فرمول ارائه شده بدست آورید. بهتر است این بردارها را بر روی حافظه جانبی به صورتی که دسترسی مستقیم به آنها داشته باشید ذخیره کنید.

۳. **خوشه‌بندی اسناد:**

الگوریتم‌های متفاوتی برای خوشه بندی مقالات وجود دارد که مشهورترین آنها الگوریتم Kmeans می‌باشد. با استفاده از این الگوریتم اسناد را به $k=\sqrt{N}$ خوشه تقسیم کرده و سر گروه هر یک از اسناد را مشخص کنید. تعداد تکرار این الگوریتم را با توجه به معیارهای مشخص خودتان تعیین کرده و گزارش کنید.

استفاده از خوشه بندی سلسله مراتبی در این بخش نمره اضافه دارد. در صورتی که از خوشه بندی سلسله مراتبی استفاده کردید باید مقایسه‌ی دقیقی بین این حالت و خوشه بندی معمولی از نظر حافظه مصرفی، سرعت جست‌وجو، میانگین تعداد مقایسه اسناد و همچنین شباهت اسناد یافته شده، در گزارش خود ارائه کنید.

۴. **وارد کردن درخواست کاربر:**

با استفاده از یک رابط کاربری ساده درخواست کاربر را گرفته و آن را پردازش کرده و Stop word های آن را حذف، ریشه یابی، و بردار مربوط به درخواست کاربر را آماده کنید. (لغاتی که توسط کاربر وارد شده و در فرهنگ لغات وجود ندارد را دور بریزید.)

جست‌وجوی کلمات مشابه نیز نمره اضافی دارد. یعنی وقتی کاربر درخواست خود را وارد کرد کلمات هم معنی کلمات وارد شده نیز به درخواست کاربر اضافه گردد و این کلمات نیز در مقالات جست‌وجو گردد. به طور مثال در صورت وارد کردن کلمه خودرو کلمه ماشین نیز به کلمات مورد جست‌وجو اضافه شده و به کمک آن سرچ انجام شود.

۵. **باز گرداندن نتیجه:**

شباهت بردار حاصل از درخواست کاربر را با استفاده از فاصله‌ی کسینوسی بدست آورده و ۱۰ مقاله دارای بیشترین شباهت با درخواست کاربر را نمایش دهید.

نمایش مقالات بازیابی شده در رابط کاربری و برجسته کرده لغات جست و جو شده در متن مقالات نمره اضافی دارد.

مقادیر پارامترهای b1 , b2 (lecture7-vectorspace slide 36) را به گونه‌ای برای خود تعیین کنید که بهترین نتایج از نظر شباهت و سرعت جست و جو را بگیرید. نحوه‌ی تعیین این پارامترها را در گزارش پروژه خود بیاورید.

نکات مهم در پیاده‌سازی:

- در تمامی بخش‌های پروژه باید میزان استفاده از حافظه اصلی استفاده شده کمتر از ۱۲۸ مگابایت باشد. هر چه میزان بیشتری از این مقدار استفاده کنید نمره کمتری خواهید گرفت و استفاده کمتر از حافظه نمره اضافی دارد.
- هر چه سرعت اجرای پروژه شما بیشتر باشد بهتر است و نمره بیشتری خواهید گرفت.
- قسمت‌هایی که به صورت سبز رنگ مشخص شده‌اند نمره اضافی دارد.
- دانشجویانی که به Moodle دسترسی ندارند پروژه خود را با Subject، IR_PROJECT_3 به ایمیل h.ramezany72@gmail.com ارسال کنند.
- امکان ارائه با تاخیر این فاز از پروژه وجود ندارد.

مهلت ارسال:

- مهلت ارسال این مرحله از پروژه روز ۵ تیر خواهد بود.
- تحویل حضوری فاز دوم و سوم پروژه در بازه‌ی ۶ تا ۷ تیر خواهد بود که تاریخ دقیق تحویل حضوری این دو فاز متعاقباً از طریق سایت درس اعلام خواهد شد. در تحویل حضوری هر دو فاز را باید به صورت جداگانه اجرا کرده و دقیقاً نحوه کار خود را توضیح دهید. در صورتی که در تاریخ‌های ذکر شده تحویل حضوری پروژه شرکت نکنید بخش زیادی از نمره خود را از دست خواهید داد.
- تاریخ ذکر شده به هیچ عنوان قابل تمدید نخواهد بود.

موفق باشید