دانشکده مهندسی کامپیوتر و فناوری اطلاعات

دانشگاه صنعتی امیرکبیر

پروژه درس ذخیره و بازیابی اطلاعات

فاز سوم:

خوشه‌بندی و جستجو در اسناد

ترم دوم ۱۳۹۴-۱۳۹۵

مقدمه

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.

کارهایی که در این فاز می‌بایست انجام شوند، به شرح زیر است:

1. **ایجاد شاخص سند-لغت**

در هنگام پیمایش هر سند شناسه لغات آن سند و تعداد تکرار هر کلمه در آن سند را بدست آورده و به صورت برداری برای آن سند ذخیره کنید. توجه داشته باشید که ممکن است نتوانید بردارهای حاصل را در حافظه اصلی نگه‌داری کنید.

2. **ایجاد شاخص وارون لغت-سند**

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

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

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

$$W_{ij} = \frac{n_{ij}}{n_{maxj}} * \log_2 \frac{N}{n_i}$$

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

n_{maxj}: 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.

۴. خوشه‌بندی اسناد

الگوریتم‌های متفاوتی برای خوشه بندی مقالات وجود دارد که مشهورترین آنها الگوریتم Kmeans می‌باشد. اجازه استفاده از Toolboxهای آماده پیاده سازی شده برای این الگوریتم را ندارید. 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. نحوه تعیین سرگروه هر گروه را نیز در گزارش خود توضیح دهید.

۵. 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، کلمات باقی مانده را 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.

۶. 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.

۷. مجموعه داده‌ها

از مجموعه داده‌هایی که در فاز دوم در اختیار شما قرار داده شده است استفاده کنید.

مواردی که باید تحویل داده شوند:

۱) کد کامل برنامه
۲) نسخه اجرایی به همراه فایل شرح نحوه اجرای بخش‌های مختلف برنامه
۳) گزارش پروژه که مهم‌ترین بخش پروژه بوده و در صورت ناقص بودن این بخش، از نمره‌ی شما کسر خواهد شد.

نکات مهم:

- پروژه خود را با زبان جاوا پیاده سازی کنید.
- پروژه بدون گزارش نمره ای نخواهد داشت.
- 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.
- 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.
- حداکثر مهلت آپلود این تمرین جمعه ۱۴ خرداد ۱۳۹۵ ساعت ۲۲ می باشد.
- به ازای هر روز تاخیر ۵ درصد از نمره شما کسر خواهد شد.
- در صورت مشاهده هرگونه تقلب در کدها و گزارش های ارسالی به تمامی افراد دخیل در تقلب، صفر داده خواهد شد.
- در صورت وجود هرگونه سوال یا ایراد، سوالات خود را باعنوان ISR-Project3 به آدرس ایمیل h.ramezany72@gmail.com ارسال کنید.

موفق باشید