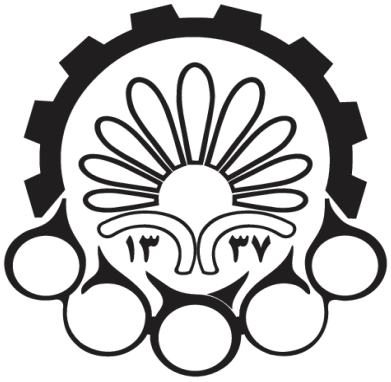

دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر

درس:

بازیابی اطلاعات

تعریف پروژه – مرحله دوم

نیمسال دوم

سال تحصیلی ۹۸–۹۹

۴- مرحله‌ی دوم پروژه

در این مرحله مدل بازیابی اطلاعات باید بتواند نتایج جستجو را بر اساس ارتباط رتبه‌بندی کند. مدل بازیابی اطلاعات این کار را با مدل‌سازی اسناد در فضای برداری انجام می‌دهد. به این صورت که برای هر سند یک بردار عددی استخراج می‌شود که بازنمایی آن سند در فضای برداری است. سپس با داشتن یک پرسمان از کاربر ابتدا آن را به فضای برداری برده و سپس با استفاده از یک معیار شباهت مناسب، فاصله‌ی بردار عددی پرسمان را با تمام اسناد در فضای برداری محاسبه کرده و در نهایت نتایج خروجی را بر اساس شباهت مرتب‌سازی می‌کنیم. همچنین برای افزایش سرعت پاسخگویی مدل بازیابی اطلاعات روش های مختلفی به کار گرفته خواهد شد. جزئیات هر بخش به تفصیل در ادامه بیان شده است.

۴-۱-مدل سازی اسناد در فضای برداری

در مرحله قبل پس از استخراج توکن ها اطلاعات به صورت یک دیکشنری ذخیره شدند. در این بخش هدف آن است که اسناد در فضای برداری بازنمایی شوند. با استفاده از روش وزن‌دهی $tf - idf$ بردار عددی برای هر سند محاسبه خواهد شد و درنهایت هر سند به صورت یک بردار شامل وزن های تمام کلمات آن سند بازنمایی می شود.

محاسبه‌ی وزن هر کلمه t در یک سند d با داشتن مجموعه‌ی تمام اسناد D با استفاده از معادله‌ی زیر محاسبه می‌شود:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) = \big(1 + \log(f_{t,d})\big) \times \log(\frac{N}{n_t})$$

که در آن $f_{t,d}$ تعداد تکرار کلمه‌ی t در سند d و n_t تعداد سندهایی است که کلمه‌ی t در آنها ظاهر شده است. توضیحات بیشتر این روش در فصل ۶ کتاب آمده است.

برای آنکه از به کار بردن فضای بیش از حد جلوگیری شود در بازنمایی اسناد به فضای برداری از تکنیک $Index\ elimination$ استفاده نمایید.

خروجی: یک سند را به دلخواه انتخاب کنید. سپس ۵ کلمه با بیشترین وزن و ۵ کلمه با کمترین وزن را در آن پیدا کرده و کلمات را به همراه عنوان خبر در گزارش خود بیاورید. آیا زیاد بودن وزن کلمه در آن سند نشانه‌ی اهمیت بالای آن کلمه است؟ در گزارش خود توضیح دهید.

۴-۲-پاسخگویی به پرسمان در فضای برداری

با داشتن پرسمان کاربر، بردار مخصوص پرسمان را استخراج کنید. سپس با استفاده از معیار شباهت سعی کنید اسنادی را که بیشترین شباهت (کمترین فاصله) را به پرسمان ورودی دارند پیدا کنید. سپس آنها را به ترتیب شباهت نمایش دهید. معیارهای فاصله‌ی مختلف می‌تواند برای این کار در نظر گرفته شود که ساده‌ترین آنها شباهت کسینوسی بین بردارها است که زاویه‌ی بین آنها را محاسبه می‌کند. این معیار به صورت زیر تعریف می‌شود:

$$similarity(a,b) = \cos(\theta) = \frac{a.b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}}$$

در انتهای کار برای نمایش یک صفحه از نتایج پرسمان فقط کافیست K سندی انتخاب شوند که بیشترین شباهت را به پرسمان داشتند. ساده‌ترین راه حل برای این کار مرتب‌سازی تمام اسناد براساس شباهت‌شان با پرسمان است که هزینه زمانی این کار از مرتبه‌ی $O(nlogn)$ است که با فرض زیاد بودن تعداد اسناد می‌تواند باعث زیاد شدن شدید زمان پاسخ موتور جستجو شود. برای حل این مسئله از پشته $(heap)$ استفاده کنید و برای نمایش هر صفحه تنها K سند با بیشترین شباهت را از آن بیرون بکشید. توجه کنید که ساختن پشته از مرتبه‌ی زمانی $O(2n)$ و استخراج K سند با بیشترین مقدار از مرتبه‌ی $O(logn)$ است و در مجموع این تکنیک می‌تواند حدوداً مشکل زیاد بودن زمان پاسخ را حل کند. توجه کنید که اسناد با امتیاز صفر نیازی نیست در پشته ریخته شوند. شناسایی این اسناد و حذف آنها با استفاده از تکنیک $Index\ elimination$ در مرحله اول انجام شده است.

خروجی: برای ارزیابی این بخش از پروژه، شما باید یک واسط ساده به پروژه اضافه کنید که بتوان به استفاده از آن روی این موتور جستجو پرسمان اجرا کرد. ساده‌ترین حالت ممکن، دریافت پرسمان از طریق متن روی کنسول بعد از اجرای برنامه است. توجه کنید که بهتر است دو بخش ساختن شاخص و اجرای پرسمان را از هم جدا کنید. پس از خواندن اسناد و ساختن شاخص آن را در یک (یا چند) فایل ذخیره کنید. سپس یک برنامه‌ی دیگر به عنوان اجرا کننده‌ی پرسمان بنویسید که فقط فایل‌های مربوط به شاخص را خوانده و پرسمان‌های ورودی را روی آن اجرا می‌کند. این واسط برای اجرا کردن پرسمان‌ها استفاده می‌شود. خروجی این واسط، چاپ کردن عنوان و متن ۱۰ خبری است که بیشترین شباهت را با پرسمان ورودی داشته‌اند و در حقیقت نتایج موتور جستجو در پاسخ به پرسمان هستند. توجه کنید که پرسمان ورودی دنباله‌ای از کلمات خواهد بود که در کنسول برنامه وارد می‌شود و خروجی ۱۰ خبر مرتبط (نتایج جستجو) است که در کنسول چاپ می‌شوند.

۲-۴-افزایش سرعت پردازش پرسمان

با استفاده از تکنیک $Index\ elimination$ تاحدودی مشکل زیاد بودن زمان در مراحل قبل حل شد اما همچنان زمان پاسخگویی برای بسیاری از کاربردها قابل قبول نمی‌باشد. برای آنکه سرعت پردازش و پاسخگویی افزایش یابد روش های مختلفی وجود دارند که یکی از آنها روش $Champion\ lists$ می باشد که قبل از آنکه پرسمانی مطرح شود و در مرحله پردازش اسناد، یک لیست از مرتبط ترین اسناد مربوط به هر $term$ در لیست جداگانه‌ای نگهداری می شوند. برای پیاده سازی این بخش پس از ساخت شاخص معکوس، $Champion\ list$ را ایجاد کنید و تنها بردار پرسمان را با بردار اسنادی که از طریق جستجو در $Champion\ list$ به دست آورده اید مقایسه کنید و k سند مرتبط را به نمایش بگذارید. (توضیحات بیشتر این روش در فصل ۷ کتاب آمده است.)

خروجی: در این حالت نیز باید یک واسط ساده کاملا مشابه با واسط بخش قبل برای اجرای پرسمان وجود داشته باشد. همچنین برای ارزیابی زمانی، یک پرسمان دلخواه انتخاب کرده و آن را روی موتور جستجو بدون استفاده از تکنیک $Champion\ list$ اجرا کنید. سپس همان پرسمان را در موتور جستجو با استفاده از آن اجرا کرده و در دو حالت زمان

پاسخگویی و دقت ۱۰ سند موجود در نتایج جستجو را با هم مقایسه کنید. (به صورت شهودی بررسی کنید که آیا خروجی موتور جستجو برای پرسمان نتایج مناسبی هستند یا خیر)

به علاوه در گزارش خود برای کلمات زیر عنوان ۳ سند ابتدایی داخل $Champion\ list$ آنها را بیاورید. «ابریشم» «فرهنگیان» «آرامش»

نکات مهم

- پروژه به صورت فردی انجام می‌شود.
- می توانید وزن دهی $tf-idf$ و ایجاد $Champion\ lists$را با استفاده از شاخص معکوسی که در مرحله گذشته پیاده سازی کرده اید، انجام دهید و پس از آن بازنمایی اسناد به فضای برداری را انجام دهید.
- موتور جستجو بدون واسط کاربری قابل ارزیابی نیست. واسط کاربری ساده که از طریق کنسول پرسمان را گرفته و نتایج را در خروجی چاپ کند علی‌رغم سادگی، نقش اساسی در ارزیابی و نمره‌دهی پروژه‌ی شما دارد. پروژه‌ی شما در زمان تحویل پروژه با جستجوی پرسمان‌های مختلف ارزیابی می‌شود.
- علاوه بر کیفیت و دقت نتایج جستجو، زمان پاسخگویی به پرسمان‌ها نیز یکی از معیارهای ارزیابی پروژه است.
- برای تحویل پروژه می‌بایست برنامه اجرایی به همراه گزارش کتبی تحویل داده شود. گزارش کتبی می‌بایست نحوه پیاده‌سازی کلیه قسمت‌های مدل بازیابی اطلاعات را مشخص کند.