پروژه درس بازیابی اطلاعات

پاییز ۱۳۹۹

مقدمه

هدف این پروژه ایجاد یک موتور جستجو برای بازیابی اسناد است که کاربر پرسمان خود را وارد نموده و سامانه اسناد مرتبط را بازیابی و به کاربر ارائه می‌کند. پروژه در سه مرحله تعریف شده است که عبارتند از:

- **مرحله اول:** ایجاد یک مدل بازیابی اطلاعات ساده
- **مرحله دوم:** تکمیل مدل بازیابی اطلاعات و ارائه قابلیت‌های کارکردی پیشرفته‌تر
- **مرحله سوم(اختیاری):** استفاده از خوشه بندی و جستجو بر روی صفحات ویکی پدیا

فاز اول

در این فاز از پروژه به منظور ایجاد یک مدل بازیابی اطلاعات ساده نیاز است تا اسناد شاخص‌گذاری شوند تا در زمان دریافت پرسمان از شاخص معکوس برای بازیابی اسناد مرتبط استفاده شود. به طور خلاصه مراحل انجام این فاز از پروژه به شرح زیر می باشد.

۱) استخراج توکن
۲) ساخت شاخص معکوس
۳) پیاده سازی ۵ قاعده همسان سازی
۴) اعمال یک ایده برای جلوگیری از تغییر داده در بخش همسان سازی
۵) حذف کلمات پرتکرار
۶) پاسخ دهی به پرسمان کاربر

در ادامه به شرح نحوه انجام هریک از مراحل می پردازیم.

شاخص گذاری اسناد

برای شاخص‌گذاری اسناد لازم است بخش‌های زیر پیاده‌سازی شوند:

- واکشی اسناد
- استخراج توکن
- همسان‌ساز کلمات
- حذف کلمات پرتکرار

پس از آنکه محتوای تمامی اسناد را به صورت توکن استخراج کردید، توکن ها را به صورت یک شاخص معکوس ذخیره کنید. توجه داشته باشید که این شاخص گذاری نباید در زمان دریافت پرسمان کاربر انجام شود بلکه باید شاخص از قبل ذخیره شده باشد و در زمان پاسخ گویی به کاربر تنها از آن استفاده نمود. برای آنکه بتوانید درستی عملکرد خود را نشان دهید یک تابع آزمون برای این بخش تعریف نمایید تا با دریافت یک کلمه لیستی از شماره اسناد (شماره فایل) که شامل این کلمه بودند را به صورت مرتب نمایش دهد. دقت داشته باشید که در شاخص معکوس هم کلمات و هم شماره اسناد باید به صورت مرتب شده باشند.

همان طور که می دانید همسان‌سازی کلمات و حذف کلمات پرتکرار برای بهبود عملکرد موتور جستجو الزامی است. برای بخش همسان سازی قواعدی را در نظر بگیرید و بر روی توکن‌ها اعمال نمایید.لازم به ذکر است که وجود حداقل ۵ قاعده برای همسان‌سازی الزامی است و سعی نمایید تنوع را در این قواعد داشته باشید. به طور مثال می‌توانید بر روی فعل‌ها، حروف جمع، پیشوندها و پسوندها، شناسه‌ها و حروف بیانگر برتری مانند "تر" و "ترین" همسان‌سازی را اعمال نمایید.

گاهی اوقات همسان سازی ممکن است اطلاعاتی را از بین ببرد. به طور مثال اگر دو قاعده حذف "می" از ابتدای فعل ها و حذف شناسه مربوط به فعل ها را داشته باشیم دو کلمه زیر معادل می شوند.

میدانم ← دان ، میدان← دان

درنتیجه حداقل یک ایده برای جلوگیری از چنین اشتباهاتی در بخش همسان سازی خود به کار ببرید.

علاوه بر اقدامات بالا نیاز است تا کلمات پرتکرار را حذف نمایید. برای این مرحله می توانید لیستی از پرتکرارترین کلمات را استخراج کنید و از آن لیست استفاده نمایید و یا در زمان ساخت شاخص معکوس زمانی پرتکرارترین کلمات را بیابید.

پاسخ دهی به پرسمان کاربر

در این بخش با دریافت پرسمان کاربر باید بتوانید اسناد مرتبط با آن را به صورت دودویی بازیابی نمایید. پرسمان کاربر به دو صورت زیر می تواند باشد:

تک کلمه: تنها کافی است که لیست مربوط به آن را از روی دیکشنری بازیابی نمایید.

چند کلمه: در این بخش لیست فایل‌ها باید بر اساس میزان ارتباط مرتب شده باشد. مرتبط‌ترین سند، سندی است که تمام کلمات را داشته باشد.

فاز دوم

در این مرحله مدل بازیابی اطلاعات باید بتواند نتایج جستجو را بر اساس ارتباط آنها با پرسمان کاربر 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 روش‌های مختلفی به کار گرفته خواهد شد. جزئیات هر بخش به تفصیل در ادامه بیان شده است.

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 بر 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:

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

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.

پاسخ دهی 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.

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 نمی باشد. برای آنکه سرعت پردازش و پاسخگویی افزایش یابد روش های مختلفی وجود دارند که یکی از آن روش ها 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 جداگانه ای نگه داری می شوند. برای پیاده سازی این بخش پس از ساخت شاخص معکوس زمانی، $Champion\ list$ را ایجاد کنید و تنها بردار پرسمان را با بردار اسنادی که از طریق جستجو در $Champion\ list$ به دست آورده‌اید مقایسه کنید و K سند مرتبط را به نمایش بگذارید. توضیحات بیشتر این روش در فصل ۷ کتاب آمده است.

نکته: می توانید وزن دهی $tf-idf$ و ایجاد لیست $Champion$ را با استفاده از شاخص معکوس که در مرحله گذشته پیاده‌سازی کرده‌اید، انجام دهید و پس از آن بازنمایی اسناد به فضای برداری را انجام دهید.

مجموعه داده

مجموعه داده مورد استفاده برای ارزیابی و تست دو فاز اول به صورت فایل های متنی در اختیار شما قرار می گیرد. هر فایل متنی شامل متن یک خبر است که شما به بررسی آنها می پردازید. نام هر فایل متنی یک عدد است که شما باید از آن به عنوان id فایل استفاده نمایید.

فاز سوم

در این مرحله مدل بازیابی اطلاعات باید بتواند از خوشه‌بندی اسناد استفاده نماید. همچنین داده های مورد استفاده این بخش از صفحات ویکی پدیا استخراج می شود. در این فاز از پروژه، پس از استخراج داده‌های خوشه‌بندی شده از سایت ویکی پدیا، محتوای به دست آمده را باید به صورت بردار درآورید تا بتوانید با استفاده از بخش‌های پیاده‌سازی شده در فاز دوم پاسخگو پرسمان کاربر باشید.

استخراج مجموعه داده

مطالب موجود در صفحات ویکی پدیا به طور دسته‌بندی شده و به زبان های مختلف در دسترس می باشند. درحال حاضر کتابخانه‌هایی وجود دارند که می‌توانند داده‌های این صفحات را به زبان‌های مختلف و در دسته‌بندی های مختلف استخراج نمایند. برای آنکه بتوانیم از خوشه‌بندی جهت بهبود عملکرد و سرعت موتور جستجو استفاده کنیم نیاز است تا تمامی اسناد را به فضای برداری ببریم و پس از آن الگوریتم‌های لازم را پیاده‌سازی نماییم. در این پروژه نیازی به اجرای الگوریتم‌های خوشه‌بندی اسناد نیست چرا که می‌توان اسناد را بر اساس دسته‌بندی‌های جداگانه آنها از ویکی‌پدیا به دست آورد و از همین دسته‌بندی به عنوان خوشه‌بندی اسناد استفاده کرد.
برای شروع شما باید محتوای متن اصلی ۵۰ صفحه از هرکدام از ۵ دسته‌بندی زیر را به زبان فارسی استخراج نمایید:

۱. فیزیک
۲. ریاضیات
۳. سلامتی
۴. تاریخ
۵. تکنولوژی

حال که شما ۵ خوشه را به طور آماده در دسترس دارید باید به محاسبه مراکز هر دسته بپردازید. برای انجام این کار نیاز است تا داده‌های مرحله قبل را به صورت بردار درآورید.(از کدهای فازهای قبل خود برای ساخت شاخص معکوس و بردار اسناد استفاده نمایید.)
با انجام این مراحل شما باید اطلاعات زیر را داشته باشید:

- بازنمایی هر سند در فضای برداری + خوشه‌ای که سند به آن تعلق دارد.
- مرکز هر خوشه به صورت برداری

پاسخ دهی پرسمان با استفاده از خوشه ها

در این مرحله برای پاسخ دهی به پرسمان کاربر تنها باید پرسمان را با مراکز هر خوشه مقایسه کنید و نزدیک ترین مرکز خوشه را پیدا کنید. پس از آن پرسمان را تنها با اسناد همان خوشه مقایسه نمایید و مرتبط ترین ها را به ترتیب بازگردانید. با اینکار علاوه بر افزایش سرعت، اسناد بازگردانده شده شباهت مفهومی بیشتری نیز دارند.

نکات مهم

۱. مهلت تحویل فاز اول تا پایان روز ۲۰ آذر، فاز دوم تا پایان روز ۱۵ دی و فاز سوم تا پایان روز ۱ بهمن می باشد.

۲. تحویل تمامی فازها در تاریخ ۲ بهمن در اسکایپ انجام خواهد شد.

۳. کدهای خود را در کوئرا بارگذاری نمایید.(لینک آن در سایت درس قرار داده می شود)

۴. کدهای شما توسط کوئرا بررسی می شود. درصورت شباهت نمره تمام فازهای پروژه صفر خواهد شد.

"موفق باشید"