دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

درس:

بازیابی اطلاعات

تعریف پروژه

استاد درس
دکتر احمد نیک آبادی

نیم سال اول ۱۴۰۲

لطفاً در انجام پروژه به نکات زیر توجه فرمایید:

- پروژه انفرادی است.
- تنها در موارد ذکرشده مجاز به استفاده از کتابخانه‌های آماده هستید.
- مهلت تحویل پروژه، دوشنبه ۹ بهمن ماه است. ارسال با تاخیر برای پروژه با توجه مهلت ثبت قطعی نمرات امکان‌پذیر نیست. بنابراین لطفاً برای انجام پروژه برنامه‌ریزی لازم را داشته باشید.
- علاوه بر گزارش پروژه، در انتهای ترم یک امتحان نیز از این پروژه گرفته خواهد شد. تحویل تنهای کد یا گزارش کافی نخواهد بود و شامل نمره کامل نخواهد شد.
- پروژه درس شامل ۲ فاز است. انجام فاز اول پروژه الزامی بوده و فاز دوم امتیازی است.
- جلسات رفع اشکال و توجیهی پروژه متعاقبا اعلام خواهد شد.

راهنمایی: در صورت نیاز می توانید سوالات خود در خصوص پروژه را از تدریسیاران درس، کانال یا در گروه درس بپرسید.

id کانال: https://t.me/IR_Fall02

تدریسیاران درس: امیررضا رجبی، علی انصاری، مهدیه سادات بنیس

مقدمه

در این پروژه به صورت عملی از مفاهیم تدریس‌شده در کلاس درس استفاده می‌شود. در این پروژه هر دانشجو یک موتور جستجو برای بازیابی اسناد متنی ایجاد می کند به گونه‌ای که کاربر پرسمان خود را وارد و سامانه اسناد مرتبط را بازیابی می کند.

۱. ساخت شاخص مکانی

در مرحله اول از پروژه به منظور ایجاد یک مدل بازیابی اطلاعات ساده نیاز است تا اسناد شاخص‌گذاری شوند تا در زمان دریافت پرسمان از شاخص مکانی ایجاد شده برای بازیابی اسناد مرتبط استفاده شود.

۱.۱. مجموعه داده

مجموعه داده مورد استفاده در این پروژه مجموعه‌ای از خبرهای واکشی‌شده از چند وب‌سایت خبری فارسی است که در قالب یک فایل JSON در اختیار شما قرار خواهد گرفت. لازم است تنها محتوای "content" را به عنوان محتوای سند پردازش کنید. شماره‌ی هر خبر را به عنوان id آن سند (خبر) در نظر بگیرید و در زمان پاسخ به پرسمان، عنوان خبر و URL مربوط به سند بازیابی‌شده را نمایش دهید تا امکان بررسی صحت عملکرد سیستم وجود داشته باشد. شناسه خبرها یکتا است اما شماره ها الزاما پشت سر هم نخواهد بود.

۲.۱. پیش‌پردازش اسناد

از ساخت شاخص مکانی لازم است متون را پیش‌پردازش کنید. گام‌های لازم در این قسمت به صورت زیر می باشد.

- استخراج توکن

این اولین بخش از شروع کار است و تاثیر زیادی در نتیجه دارد. در این بخش باید به نکات زیر دقت شود:

- بین کلمات ممکن است از tab ، نیم فاصله ، فاصله و خط جدید (\n) استفاده شده باشد.
- عبارتی مانند "می‌تواند" یک کلمه است و لازم است که به عنوان یک توکن در نظر گرفته شود.
- باید ایمیل ها، آیدی ها و اعداد به درستی توکنایز شوند.
- برای قسمت اضافی در این بخش می توان به حل مشکل افعالی پرداخت که این افعال قرار است وارد ریشه یابی شوند. سوال اصلی این است که می‌خواهید "می‌خواهم بروم" را یک توکن بگیرید یا نه؟
- به انجام دادن قسمت های اضافه برای این بخش به نسبت ۲۰ درصد نمره این بخش امتیاز داده می‌شود (برای مثال مخفف ها رو بررسی کنید)

- نرمال‌سازی متون

نرمال‌سازی متن به فرآیند تبدیل داده های متنی به شکلی سازگار و استاندارد اشاره دارد. این امر به کاهش تغییرات در نمایش کلمات کمک می کند و تجزیه و تحلیل و مقایسه متن را آسان تر می کند.

۱) فاصله‌گذاری صحیح(correct spacing) در موارد زیر

ی - ای - ها - های - هایی - تر - تری - ترین - گر - گری - ام - ات - اش - اعداد - می - نمی

۲) تعویض یونیکد (unicode replacement)

تبدیل "ی"ها و "ک"ها و "آ" و همچنین تبدیل ده کلمه که به دو شکل نوشته می‌شوند و در نرمال‌سازی آن‌ها را به یک شکل تبدیل کنید مانند: بسم الله الرحمن الرحیم و بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

۳) حذف بعضی از کاراکترها

حذف فتحه، کسره، ضمه، تنوین، تشدید، الف کشیده، سکون، همزه و همچنین تمام علائم نشانه‌گذاری مانند !، >، ،، ...

۴) تبدیل اعداد انگلیسی به اعداد فارسی

۵) جدا کردن "می" و "نمی" با نیم فاصله از فعل‌ها

- حذف کلمات پر تکرار

در این بخش هدف حذف کلمات پرتکرار است. ابتدا فرکانس (تعداد تکرار) هر کلمه در مجموعه اسناد مشخص شود و کلمات بر اساس فرکانس به ترتیب از زیاد به کم مرتب شوند. سپس ۵۰ کلمه اول پر تکرار این کلمات باید حذف شوند. در داخل گزارش باید لیست کلمات حذف شده و تعداد تکرار هر کدام از آن‌ها آورده شود.

- ریشه‌یابی

برای انجام پیش‌پردازش‌های لازم در این بخش میتوانید با صلاحدید خود یکی از کتابخانه‌های آماده را انتخاب و از آن استفاده کنید (راهنمایی: کتابخانه۱ و کتابخانه۲) و یا پیاده‌سازی شخصی خود را داشته باشید. توجه: برای پیاده‌سازی شخصی بخشهای مربوط به پیش‌پردازش اسناد نمره‌ی ارفاقی لحاظ نمی‌شود. در صورت استفاده از هر یک از کتابخانه های یاد شده می بایست در گزارش خود به شکل دقیق بیان کنید که این کتابخانه ها چه پردازشی انجام می دهند.

۳.۱. ساخت شاخص مکانی

با استفاده از اسناد پیش‌پردازش‌شده در گام قبل، شاخص مکانی را بسازید. در شاخص مکانی ساخته شده علاوه بر جایگاه کلمات در اسناد، باید به ازای هر کلمه از دیکشنری مشخص باشد که تعداد تکرار آن کلمه در کل اسناد چقدر است. همچنین باید مشخص باشد که در هر سند تعداد تکرار یک کلمه‌ی مشخص چند بار است و کلمه در چه مکان هایی آمده است. جزئیات کامل این قسمت در بخش ۲.۴.۲ از کتاب مرجع درس قابل مشاهده است. برای پیاده‌سازی این قسمت میتوانید به اختیار خود یک ساختمان داده‌ی مناسب را انتخاب کنید. دقت کنید که ساختمان داده‌ی انتخابی به‌گونه‌ای نباشد که در زمان جستجو و دیگر عملیات، سرعت مدل را پایین آورد. در گزارش خود نحوه ساخت شاخص مکانی را با ذکر نمونه خروجی نمایش دهید.

۲. پاسخ‌دهی به پرسمان در فضای برداری

در این مرحله مدل بازیابی اطلاعات را گسترش و بازنمایی اسناد را به صورت برداری انجام دهید تا نتایج جستجو بر اساس ارتباط آنها با پرسمان کاربر رتبه‌بندی شود. به این صورت که برای هر سند یک بردار عددی استخراج می‌شود که بازنمایی آن سند در فضای برداری است و این بردارها ذخیره می‌شوند. در زمان دریافت پرسمان، ابتدا بردار متناظر با آن پرسمان در همان فضای برداری ساخته و سپس با استفاده از یک معیار شباهت مناسب، شباهت بردار عددی پرسمان با بردار تمام اسناد در فضای برداری محاسبه می‌شود و در نهایت نتایج خروجی بر اساس میزان شباهت مرتب‌سازی می‌شوند. برای افزایش سرعت پاسخگویی مدل بازیابی اطلاعات میتوان روش‌های مختلفی را به کار گرفت که به تفصیل در ادامه بیان می‌شود.

۱.۲. مدل‌سازی اسناد در فضای برداری

در مرحله قبل پس از استخراج توکن‌ها اطلاعات به صورت یک دیکشنری و شاخص مکانی ذخیره شدند. در این بخش هدف آن است که اسناد در فضای برداری بازنمایی شوند. با استفاده از روش وزن دهی tf-idf بردار عددی برای هر سند محاسبه خواهد شد و درنهایت هر سند به صورت یک بردار شامل وزن‌های تمام کلمات آن سند بازنمایی می‌شود. محاسبه‌ی وزن هر کلمه t در یک سند d با داشتن مجموعه‌ی تمام اسناد D با استفاده از است.

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

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 مرجع درس آمده است.

در نمایش برداری فوق برای کلمه‌ای که در یک سند وجود نداشته باشد وزن صفر در نظر گفته میشود و از این جهت بسیاری از عناصر بردارهای محاسبه شده صفر خواهد بود. برای صرفه جویی در مصرف حافظه به جای آن که برای هر سند یک بردار عددی کامل در نظر بگیرید که بسیاری از عناصر آن صفر هستند می‌توانید وزن کلمات در اسناد مختلف را در همان لیستهای پستها ذخیره کنید. در زمان پاسخگویی به پرسمان کاربر که در ادامه توضیح داده میشود نیز همزمان با جستجوی کلمات در لیستهای پست‌ها می‌توانید وزن کلمات در اسناد مختلف را نیز واکشی کنید و به این شکل تنها عناصر غیر صفر بردارهای اسناد ذخیره و پردازش می‌شوند.

۲.۲. پاسخ‌دهی به پرسمان در فضای برداری

با داشتن پرسمان کاربر، بردار مخصوص پرسمان را استخراج کرده (وزن کلمات موجود در پرسمان را محاسبه کنید.) سپس با استفاده از معیار شباهت سعی شود اسنادی را که بیشترین شباهت (کمترین فاصله) را به پرسمان ورودی دارند پیدا شود. سپس نتایج را به ترتیب شباهت نمایش دهید. معیارهای فاصله‌ی مختلفی میتواند برای این کار در نظر گرفته شود که در این پروژه، دو مورد از این معیارها با هم مقایسه می‌شوند.

- شباهت کسینوسی بین بردارها: معیاری که زاویه‌ی بین دو بردار را محاسبه می‌کند. این معیار به صورت زیر تعریف می‌شود:

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

روش بالا را پیاده‌سازی کرده و در بخش گزارش، بازیابی پرسمان‌ها را به روش فوق انجام دهید. توجه کنید که برای افزایش سرعت میتوانید با استفاده از تکنیک Index elimination، معیار فاصله را با اسنادی که امتیاز صفر خواهند گرفت محاسبه نکنید. در انتهای کار برای نمایش یک صفحه از نتایج پرسمان تنها کافیست K سندی انتخاب شوند که بیشترین شباهت را به پرسمان دارند.

۳.۲. افزایش سرعت پردازش پرسمان

با استفاده از تکنیک Index elimination تا حدودی مشکل زیاد بودن زمان در مرحله قبل حل میشود اما همچنان زمان پاسخگویی برای بسیاری از کاربردها قابل قبول نمی‌باشد. برای آنکه سرعت پردازش و پاسخگویی افزایش یابد میتوانید از Champions lists استفاده کنید که قبل از آنکه پرسمانی مطرح شود و در مرحله پردازش اسناد، یک لیست از مرتبط‌ترین اسناد مربوط به هر کلمه در لیست جداگانه‌ای نگهداری شود. برای پیاده‌سازی این بخش پس از ساخت شاخص معکوس مکانی، Champions lists را ایجاد کنید و تنها بردار پرسمان را با بردار اسنادی که از طریق جستجو در Champions lists به دست آورده اید مقایسه کنید و K سند مرتبط را به نمایش بگذارید. توضیحات بیشتر این روش در فصل ۷ کتاب آمده است.

توجه: میتوانید وزن دهی tf-idf و ایجاد لیست Champions lists را با استفاده از شاخص مکانی که در مرحله قبل پیاده‌سازی کردید، انجام دهید.

۴.۲. گزارش

۱. جزئیات روش پیاده سازی شده خود در هر بخش پروژه و مواردی که در بالا به آنها اشاره شد را در گزارش ذکر کنید.
۲. پاسخ به پرسمان در حالت‌های زیر را در گزارش بیان کنید.
الف) یک پرسمان از کلمات ساده و متداول تک کلمه‌ای
ب) یک پرسمان از عبارات ساده و متداول چند کلمه‌ای
پ) یک پرسمان دشوار و کم تکرار تک کلمه‌ای
ت) یک پرسمان دشوار و کم تکرار چند کلمه‌ای

منظور از پرسمان ساده پرسمانی شامل کلمات متداول در مجموعه‌داده خبری مانند ایران، فارس، اخبار و ... است و منظور از پرسمان دشوار پرسمانی شامل کلمات غیرمتداول مانند نام اشخاص یا اسم مناسبتی خاص و کلمات کمتر استفاده شده در مجموعه‌داده خبری می‌باشد. مانند کریسمس، کمیسیون، ... است.

در هر مورد، تیتر خبر بازیابی شده را به همراه جمله(هایی) که حاوی عبارت پرسمان بوده‌اند، گزارش کنید. همچنین در هر مورد با ذکر جزئیات شرح دهید که آیا سند بازیابی شده به پرسمان کاربر مرتبط هست یا خیر؟ تحلیل هر مورد الزامی است.