درس:

بازیابی اطلاعات

تعریف پروژه

فاز دوم

بهار ۱۴۰۲

لطفاً در انجام پروژه به نکات زیر توجه فرمایید:

- پروژه انفرادی است.
- تنها در موارد ذکرشده در تمرین مجاز به استفاده از کتابخانه‌های آماده هستید.
- مهلت تحویل فاز دوم، ۳۱ اردیبهشت ۱۴۰۲ است. به دلیل نیاز به تحویل گرفتن پروژه، ارسال با تأخیر برای این فاز از پروژه امکان‌پذیر نیست. بنابراین لطفاً برای انجام پروژه برنامه‌ریزی لازم را داشته باشید.
- کدهای خود را در کوئرا بارگذاری نمایید.
- کدهای شما (به همراه کدهای دانشجویان ترم‌های گذشته) توسط کوئرا بررسی می‌شود. در صورت وجود شباهت، نمره‌ی فرد صفر خواهد شد.
- ملاک اصلی انجام فعالیت ارائه گزارش مربوطه است و ارسال کد بدون گزارش نمره‌ای نخواهد داشت. سعی کنید گزارش شما دقیقا در راستای موارد خواسته‌شده باشد و از طرح موارد اضافی خودداری کنید.
- پروژه درس شامل ۳ فاز است. انجام دو فاز ابتدایی پروژه الزامی بوده و فاز اول ۴۰ درصد و فاز دوم ۶۰ درصد از کل نمره ی پروژه درس را به خود اختصاص می‌دهند. فاز نهایی امتیازی است.
- موعد تحویل حضوری متعاقباً از طریق سایت درس اعلام خواهد شد.

راهنمایی:

در صورت نیاز می‌توانید سوالات خود در خصوص پروژه را از تدریسیاران درس و یا از طریق ایمیل زیر بپرسید.

IR.course1402@gmail.com

تدریس‌یاران درس: آیلار صدائی روژینا کاشفی رها احمدی محمدجواد رجبی

مقدمه

در فاز اول، پس از پیش‌پردازش اسناد و ساخت شاخص مکانی، موتور بازیابی اطلاعات ساده‌ای طراحی کردیم که به پرسمان‌ها بر اساس وجود/عدم وجود کلمات در هر سند پاسخ می‌دهد. در فاز دوم قصد داریم این بازیابی را دقیق‌تر انجام دهیم؛ به گونه‌ای که موتور جستجو، اسناد مرتبط‌تر با پرسمان کاربر را تشخیص داده و در ابتدای لیست نتایج نمایش دهد.

۲- فاز دوم

در این مرحله می‌خواهیم مدل بازیابی اطلاعات را گسترش و بازنمایی اسناد را به صورت برداری انجام دهیم تا بتوانیم نتایج جستجو را بر اساس ارتباط آن‌ها با پرسمان کاربر رتبه‌بندی کنیم. 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.

۲-۱ 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 وزن‌دهی $tf-idf$ 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 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 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.

اول بردار کوئری رو می سازیم (جفت کلمه و tfidf)
بعد لیست های کلماتی که توی کوئری هستند رو میاریم
میریم روشون جلو و به ازای هر سند ضرب انجام می دیم و میریزیم داخل max heap (جفت docid و score)

۲-۲ پاسخ‌دهی به پرسمان در فضای برداری

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

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

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

- معیار شباهت ژاکارد: معیاری که نسبت تعداد اشتراک به تعداد اجتماع را می‌سنجد. این معیار به صورت زیر تعریف می‌شود:

$$similarity(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

دو روش بالا را پیاده‌سازی کنید و در بخش گزارش، بازیابی پرسمان‌ها را با هر دو روش فاصله‌یابی 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.

۲-۳ 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 $Champion\ lists$ استفاده کنید که قبل از آنکه پرسمانی مطرح شود و در مرحله پردازش اسناد، یک لیست از مرتبط‌ترین اسناد مربوط به هر $term$ در لیست جداگانه‌ای نگهداری شود. برای پیاده‌سازی این بخش پس از ساخت شاخص معکوس مکانی، $Champion\ list$ 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.

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

۲-۴ گزارش

۱. پاسخ به پرسمان در حالت‌های زیر:

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

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.

۲. موارد ب و ت را با روش مکانی فاز یک نیز تکرار کنید و نتایج دو حالت را با هم مقایسه و تحلیل کنید.

۳. نتایج دو روش معیارهای شباهت (شباهت کسینوسی و شباهت ژاکارد) را برای پرسمان‌های بالا با هم مقایسه و تحلیل کنید.

موفق و پیروز باشید.

پایان فاز دوم