



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

درس:

بازیابی اطلاعات

تعریف پروژه (فاز اول و دوم)

بهار ۱۴۰۱

مقدمه

در این پروژه می‌خواهیم بصورت عملی از مفاهیم تدریس شده در کلاس درس استفاده کنیم. پروژه در سه فاز تعریف می‌شود که انجام دو فاز اول الزامی و فاز سوم امتیازی می‌باشد. در دو فاز اول از شما می‌خواهیم یک موتور جستجو برای بازیابی اسناد متنی ایجاد کنید به گونه‌ای که کاربر پرسمان خود را وارد نموده و سامانه اسناد مرتبط را بازنمایی کند.

در انجام پروژه به نکات زیر توجه فرمایید:

- پروژه انفرادی است.
- تنها در موارد ذکر شده در تمرین مجاز به استفاده از کتابخانه‌های آماده هستید.
- کدهای خود را در کوئرا بارگذاری نمایید (آدرس مربوطه در سایت درس قرار داده می‌شود).
- کدهای شما (به همراه کدهای دانشجویان ترم‌های گذشته) توسط کوئرا بررسی می‌شود. در صورت وجود شباهت، نمره‌ی طرفین **صفر** خواهد شد.
- ملاک اصلی انجام فعالیت ارائه گزارش مربوطه است و **ارسال کد بدون گزارش فاقد ارزش است**. سعی کنید گزارش شما دقیقاً در راستای موارد خواسته شده باشد و از طرح موارد اضافی خودداری کنید.
- مهلت ارسال فاز اول پروژه تا پایان روز **۲۶ فروردین ماه** و فاز دوم تا پایان روز **۳۰ اردیبهشت ماه** می‌باشد.
- انجام فازهای یک و دو پروژه الزامی بوده و هر کدام ۵۰ درصد از کل نمره‌ی پروژه درس را به خود اختصاص می‌دهند.
- به ازای هر روز تاخیر در در فاز اول ۵ درصد از نمره‌ی فاز مربوطه کسر می‌شود.
- موعد تحویل متعاقباً از طریق سایت درس اعلام خواهد شد.

راهنمایی:

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

IR.course1400@gmail.com

۱- فاز اول

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

- پیش پردازش داده ها
- ساخت شاخص مکانی
- پاسخ دهی به پرسمان کاربر

در ادامه هر مورد به صورت کامل شرح داده می شود.

۱-۱ پیش پردازش اسناد

قبل از ساخت شاخص مکانی لازم است متون را پیش پردازش کنید. گام های لازم در این قسمت به صورت زیر می باشد.

- استخراج توکن
- نرمال سازی متون
- حذف کلمات پر تکرار^۱
- ریشه یابی

برای انجام پیش پردازش های لازم می توانید با صلاح دید خود یکی از کتابخانه های آماده را انتخاب و از آن استفاده کنید (راهنمایی: [کتابخانه ۱](#) و [کتابخانه ۲](#)) و یا پیاده سازی شخصی خود را داشته باشید. **توجه:** برای پیاده سازی شخصی بخش های مربوط به پیش پردازش اسناد نمره ی ارفاقی لحاظ نمی شود.

۱-۲ ساخت شاخص مکانی

با استفاده از اسناد پیش پردازش شده 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 کلمه ی مشخص چقدر است. جزئیات کامل این قسمت در بخش ۲.۴.۲ از کتاب مرجع درس قابل مشاهده است. برای پیاده سازی این قسمت

¹ Stop Words

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).

۱-۳ پاسخ‌دهی به پرسمان کاربر

در این بخش پرسمان کاربر در قالب یک متن آزاد دریافت می‌گردد. حداقل عملگرهای قابل استفاده در این بخش «!» بعنوان عملگر NOT و "" برای تعیین یک عبارت می‌باشد. پس از بازیابی، اسناد را بصورت رتبه‌بندی شده نمایش دهید. برای رتبه‌دهی به اسناد، سندی که تعداد بیشتری از کلمات پرسمان را در خود دارد مرتبط‌تر است.

۱-۴ مجموعه داده

مجموعه داده مورد استفاده در این پروژه مجموعه‌ای از خبرهای واکنشی شده از چند وب‌سایت خبری فارسی است که در قالب یک فایل JSON در اختیار شما قرار خواهد گرفت. لازم است تنها محتوای "content" را بعنوان 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 بازیابی شده را نمایش دهید تا امکان بررسی صحت عملکرد سیستم وجود داشته‌باشد.

۱-۵ گزارش

۱. با ذکر مثال شرح دهید که در گام پیش‌پردازش چه عملیاتی انجام داده‌اید. همچنین دلیل انجام هر پردازش را ذکر کنید.

۲. صحت قانون Zipf را در دو حالت قبل و بعد از حذف کلمات پرتکرار از واژه‌نامه بررسی کنید (رسم نمودار برای هر حالت الزامی است). در صورت برقراری/عدم برقراری این قانون در هر حالت، علت را شرح دهید.

۳. صحت قانون heaps را در دو حالت قبل و بعد از ریشه‌یابی بررسی کنید. برای بررسی این قانون لازم است با استفاده از اندازه‌ی واژه‌نامه و تعداد توکن‌ها در ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول، اندازه‌ی واژه‌نامه مربوط به کل اسناد تخمین زده شود. در نهایت اندازه‌ی واقعی واژه‌نامه و اندازه‌ی تخمینی در هر دو حالت مقایسه و تحلیل شود. آیا در هر دو حالت قانون برقرار است؟ چرا؟ (رسم نمودار برای هر حالت الزامی است).

۴. حداقل سه مورد از مواردی که در ریشه‌یابی با چالش روبرو بودید را ذکر کنید. (بطور مثال کلماتی که نیازی به ریشه‌یابی ندارند اما طبق روند ریشه‌یابی از دست می‌روند).

۵. پاسخ به پرسمان در حالت‌های زیر:

الف) یک پرسمان از کلمات ساده و متداول (مانند تحریم‌های آمریکا علیه ایران، در نتایج بازیابی شده انتظار می‌رود اسنادی که کلمات تحریم، آمریکا، علیه و ایران را دارند در بالای لیست و اسنادی که برخی از کلمات را ندارند در رتبه‌های پایین‌تر لیست قرار داشته باشند).

ب) یک پرسمان با عملگر NOT (مانند تحریم‌های آمریکا ! ایران، انتظار می‌رود اسنادی که شامل دو کلمه تحریم و آمریکا هستند اما کلمه‌ی ایران را ندارند در نتایج بازیابی شده وجود داشته باشند).

پ) یک پرسمان با عملگر عبارت (مانند "کنگره ضدتروریست"، انتظار می‌رود اسنادی که شامل عبارت کنگره ضدتروریست در نتایج بازیابی شده وجود داشته باشند؛ عبارت دیگر موقیت مکانی کلمات در این حالت مهم است).

ت) یک پرسمان پیچیده (مانند "تحریم هسته‌ای" آمریکا ! ایران، انتظار می‌رود اسنادی که شامل عبارت تحریم هسته‌ای و کلمه‌ی آمریکا هستند اما کلمه‌ی ایران را ندارند در نتایج بازیابی شده وجود داشته باشد).

ث) یک پرسمان کلمات نادر (مانند اورشلیم ! صهیونیست، خروجی مورد انتظار این قسمت مشابه با قسمت ب می‌باشد با این تفاوت که کلمات استفاده شده در پرسمان از کلمات نادر هستند).

در هر مورد، تیتتر خبر بازیابی شده را به همراه جمله(هایی) از هر سند بازیابی شده، 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؟

توجه ۱: در مواردی که تعداد اسناد بازیابی شده زیاد است، تنها ۵ سند اول را در گزارش وارد کنید.

توجه ۲: تیتتر اخبار را با فرمت مناسب و خوانا در گزارش خود بنویسید.

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 مدل سازی اسناد در فضای برداری

در مرحله قبل پس از استخراج توکن ها 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 در نهایت هر سند به صورت یک بردار شامل وزن های تمام کلمات آن سند بازنمایی می شود. محاسبه ی وزن هر کلمه t در یک سند d با داشتن مجموعه ی تمام اسناد D با استفاده از معادله ی زیر محاسبه می شود:

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

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

در نمایش برداری فوق برای کلمه ای که در یک سند وجود نداشته باشد وزن صفر در نظر گرفته می شود 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 می توانید وزن کلمات در اسناد مختلف را در همان لیست های پست ها ذخیره کنید. در زمان پاسخ گویی به پرسمان کاربر که در ادامه توضیح داده می شود نیز همزمان با جستجوی کلمات در لیست های پست ها می توانید وزن کلمات در اسناد مختلف را نیز واکشی کنید و به این شکل تنها عناصر غیر صفر بردارهای اسناد ذخیره و پردازش می شوند.

۲-۲ پاسخ دهی 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 فاصله ی مختلفی می تواند برای این کار در نظر گرفته شود که ساده ترین آنها شباهت کسینوسی بین بردارها است که زاویه ی بین دو بردار را محاسبه می کند. این معیار به صورت زیر تعریف می شود:

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

توجه کنید که برای افزایش سرعت می توانید با استفاده از تکنیک *Index elimination* شباهت کسینوسی را با اسنادی که امتیاز صفر خواهند گرفت محاسبه نکنید. در انتهای کار برای نمایش یک صفحه از نتایج پرسمان تنها کفایست 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، تیترا 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.