

درس:

بازیابی اطلاعا<mark>ت</mark>

تعريف پروژه





لطفاً در انجام پروژه به نكات زير توجه فرماييد:

- پروژه انفرادی است.
- تنها در موارد ذکرشده 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، نمره ی 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 خواسته شده باشد و از طرح موارد اضافی خودداری کنید.
- انجام هر دو فاز پروژه الزامی بوده و هر کدام ۵۰ درصد از کل نمرهی پروژه درس را به خود اختصاص میدهند.
 - به ازای هر روز تاخیر ۵ درصد از نمرهی فاز مربوطه کسر میشود.
- مهلت تحویل هر یک از فازهای پروژه و موعد تحویل حضوری متعاقبا از طریق سایت درس اعلام خواهد شد.

راهنمایی:

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

IR.course1400@gmail.com





مقدمه

در این پروژه میخواهیم بصورت عملی از مفاهیم تدریسشده در کلاس درس استفاده کنیم. پروژه در دو فاز تعریف میشود که انجام هر دو فاز الزامی میباشد. در این پروژه 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.

١- 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 میباشد.

- پیشپردازش دادهها
- ساخت شاخص مكاني
- پاسخدهی به پرسمان کاربر

در ادامه هر مورد به صورت کامل شرح داده میشود.

۱-۱ پیشپردازش اسناد

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

- استخراج توكن
- نرمالسازی متون
- حذف کلمات یر تکرار ۱
 - ریشهیابی

برای انجام پیشپردازشهای لازم میتوانید با صلاحدید خود یکی از کتابخانههای آماده را انتخاب و از آن استفاده کنید (راهنمایی: کتابخانه ۱ و کتابخانه از ۱ و کتابخانه از ایر از ای

توجه: برای پیادهسازی شخصی بخشهای مربوط به پیشپردازش اسناد نمرهی ارفاقی لحاظ نمی شود.

-

¹ Stop Words





۱-۲ 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 کلمه ی مشخص چقدر است. جزئیات کامل این قسمت در بخش ۲.۴.۲ از کتاب مرجع درس قابل مشاهده است. برای پیادهسازی این قسمت می توانید به اختیار خود یک ساختمان داده ی مناسب را انتخاب کنید. (دقت کنید که ساختمان داده ی انتخابی به گونه ای 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 می گردد. حداقل عملگرهای قابل استفاده در این بخش «ا» بعنوان عملگرد NOT و "" برای تعین 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.

۱-۲ مجموعه داده

مجموعه داده مورد استفاده در این پروژه مجموعهای از خبرهای واکشی 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. شماره ی 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 داشته باشد.

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 ا 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 اندازه ی واژه نامه و تعداد توکنها در ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول، اندازه ی واژه نامه مربوط به کل اسناد تخمین زده شود. در نهایت اندازه ی واقعی واژه نامه و اندازه ی 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.)

پروژه درس بازیابی اطلاعات





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

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 می- رود اسنادی که کلمات تحریم، آمریکا، علیه و ایران را دارند در بالای لیست و اسنادی که برخی از کلمات را ندارند در رتبههای پایین تر 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 کلمه ی 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؛ بعبارت دیگر موقعیت 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 کلمه و کلمه و کلمه 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 به میباشد با این تفاوت که کلمات استفاده شده در پرسمان از کلمات نادر هستند.)

در هر مورد، تیتر خبر بازیابی شده را به همراه جمله(هایی) از هر سند بازیابی شده، که حاوی عبارت پرسمان کاربر 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: تیتر 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 می- شوند. 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





می شود و در نهایت نتایج خروجی بر اساس میزان شباهت مرتبسازی می شوند. برای افزایش سرعت پاسخگویی مدل بازیابی اطلاعات می توان روشهای مختلفی را به کار گرفت که به تفصیل در ادامه بیان می شود.

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 بردار عددی برای هر سند محاسبه خواهد شد و درنهایت هر سند به صورت یک بردار شامل وزنهای تمام کلمات آن سند بازنمایی می شود. محاسبه ی وزن هر کلمه t در یک سند t با داشتن مجموعه ی تمام اسناد t با استفاده از معادله ی زیر محاسبه می شود:

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

که در آن تعداد تکرار کلمه t در سند t در سند t و t تعداد سندهایی است که کلمه 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 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 لیستهای پستها ذخیره کنید. در زمان پاسخ گویی به پرسمان کاربر که در ادامه توضیح داده می شود نیز همزمان با جستجوی کلمات در لیستهای پستها 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 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 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:

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

پروژه درس بازیابی اطلاعات





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 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 می شود اما همچنان زمان پاسخگویی برای بسیاری از کاربردها قابل قبول نمی باشد. برای آنکه سرعت پردازش و پاسخگویی افزایش یابد می توانید 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 پیاده سازی 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 term به دست آورده اید مقایسه کنید و term مرتبط را به نمایش بگذارید. توضیحات بیشتر این روش در فصل term کتاب آمده است.

توجه: می توانید وزن دهی tf—idf و ایجاد لیست Champion را با استفاده از شاخص مکانی که در مرحله قبل پیاده سازی کردید، انجام دهید.

۲-۲ 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.