دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

درس:

بازیابی اطلاعات

تعریف پروژه

بهار ۱۴۰۲

لطفاً 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.
- فاز اول شامل ۳ زیر بخش: پیش‌پردازش داده‌ها، ساخت شاخص مکانی و پاسخ‌دهی به پرسمان کاربر است. ددلاین بخش پیش‌پردازش داده‌ها، ۱۷ فروردین ساعت ۲۳:۵۹، بخش ساخت شاخص مکانی ۲۴ فروردین ساعت ۲۳:۵۹ و بخش نهایی ۳۰ فروردین ساعت ۲۳:۵۹ است.
- برای هریک از بخش‌ها گزارش مورد نظر را ضمیمه کنید.
- 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 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.
- 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 نمره‌ی پروژه درس را به خود اختصاص می‌دهند. فاز نهایی امتیازی است.
- به ازای هر روز تاخیر ۵ درصد از نمره‌ی فاز مربوطه کسر می‌شود.
- موعد تحویل حضوری متعاقبا 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.course1402@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 و سامانه اسناد مرتبط را بازنمایی کند.

۱. 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 مکانی 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 خواهدگرفت. 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 بازیابی‌شده را نمایش دهید تا امکان بررسی صحت عملکرد سیستم وجود داشته باشد.

۱-۲ پیش‌پردازش اسناد

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

- استخراج توکن
- نرمال‌سازی متون
- حذف کلمات پر تکرار[1]
- ریشه‌یابی

[1] Stop Words

برای انجام پیش‌پردازش‌های لازم 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.

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

۱-۳ 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 چقدر 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.)

گزارش: نحوه ساخت شاخص مکانی را با ذکر نمونه خروجی نمایش دهید.

۱-۴ پاسخ‌دهی به پرسمان کاربر

در این بخش پرسمان کاربر در قالب یک متن آزاد دریافت می‌گردد. حداقل عملگرهای قابل استفاده در این بخش «!» بعنوان عملگرد NOT و "" برای تعین یک عبارت می‌باشد. با استفاده از این عملگرها می‌توانیم درخواست‌های کوئری را به صورت AND و AND NOT پردازش کنیم. 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.

گزارش: به پرسمان‌ها در حالات مورد نظر پاسخ دهید.

۱) یک پرسمان از کلمات ساده و متداول (مانند باشگاه‌های فوتسال آسیا، در نتایج بازیابی‌شده انتظار می‌رود اسنادی که کلمات باشگاه، فوتسال و آسیا را دارند در بالای لیست (حالت AND) و اسنادی که برخی از کلمات را ندارند در رتبه‌های پایین‌تر لیست قرار داشته باشند. (حالت OR))

۲) یک پرسمان با عملگر NOT (مانند باشگاه‌های فوتسال ! آسیا ، انتظار می‌رود اسنادی که شامل دو کلمه باشگاه و فوتسال هستند اما کلمه‌ی آسیا را ندارند (حالت AND NOT) در نتایج بازیابی‌شده وجود داشته باشند.)

۳) یک پرسمان با عملگر عبارت (مانند "سهمیه المپیک"، انتظار می‌رود اسنادی که شامل عبارت سهمیه المپیک در نتایج بازیابی‌شده وجود داشته باشند؛ بعبارت دیگر موقعیت مکانی کلمات در این حالت مهم است.)

۴) یک پرسمان پیچیده (مانند طلای "لیگ برتر" ! والیبال، انتظار می‌رود اسنادی که شامل عبارت لیگ برتر و کلمه‌ی طلا هستند اما کلمه‌ی والیبال را ندارند در نتایج بازیابی‌شده وجود داشته باشد.)

۵) یک پرسمان کلمات نادر (مانند مایکل ! جردن، خروجی مورد انتظار این قسمت مشابه با قسمت دوم 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 بوده‌اند، گزارش کنید. همچنین در هر مورد با ذکر جزئیات شرح دهید که آیا سند بازیابی شده به پرسمان کاربر مرتبط هست یا خیر؟

توجه۱: در مواردی که تعداد اسناد بازیابی‌شده زیاد است، تنها ۵ سند اول را در گزارش وارد کنید.

توجه۲: تیتر اخبار را با فرمت مناسب و خوانا در گزارش خود بنویسید.