دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

درس:

بازیابی اطلاعات

تعریف پروژه (فاز اول)

پاییز ۱۴۰۰

مقدمه

هدف از این پروژه ایجاد یک موتور جستجو برای بازیابی اسناد متنی است به گونه‌ای که کاربر پرسمان خود را وارد نموده و سامانه اسناد مرتبط را بازنمایی می‌کند. پروژه در سه مرحله تعریف شده است که عبارتند از:

مرحله‌ی اول: ایجاد یک مدل بازیابی اطلاعات ساده

مرحله‌ی دوم: تکمیل مدل بازیابی اطلاعات و ارائه‌ی قابلیت‌های کارکردی پیشرفته‌تر

مرحله‌ی سوم: پیاده‌سازی الگوریتم خوشه‌بندی و دسته‌بندی و بازیابی بر اساس خوشه/ دسته

در انجام پروژه به نکات زیر توجه فرمایید:

- 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، نمره‌ی تمام فازهای پروژه صفر خواهد شد.
- ملاک اصلی انجام فعالیت ارائه گزارش مربوطه است و ارسال کد بدون گزارش فاقد ارزش است. سعی کنید گزارش شما دقیقا در راستای موارد خواسته‌شده باشد و از طرح موارد اضافی خودداری کنید.
- مهلت ارسال فاز اول پروژه تا پایان روز ۲۸ آبان‌ماه، فاز دوم تا پایان روز ۱۹ آذرماه و فاز سوم تا پایان روز ۱ دی‌ماه می‌باشد.
- به ازای هر روز تاخیر در فاز اول و دوم ۵ درصد از نمره‌ی فاز مربوطه کسر می‌شود.
- ارسال فاز سوم با تاخیر امکان پذیر نخواهد بود.
- موعد تحویل 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.course1400@gmail.com

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. به طور خلاصه مواردی که در این فاز انجام شوند به شرح زیر می‌باشد.

- پیش‌پردازش داده‌ها
- ساخت شاخص مکانی
- 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 توکن
- نرمال‌سازی متون
- حذف کلمات پر تکرار[1]
- ریشه‌یابی

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.

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 کلمه‌ی مشخص چقدر است. جزئیات کامل این قسمت در بخش 2.4.2 از کتاب مرجع درس قابل مشاهده است. برای پیاده‌سازی 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 با دریافت پرسمان کاربر باید بتوانید اسناد مرتبط با آن را به صورت دودویی[2] بازیابی نمایید. پرسمان کاربر به دو صورت زیر می تواند باشد:

تک کلمه: تنها کافی است که لیست اسناد مربوط به آن را از روی دیکشنری بازیابی نمایید.

چند کلمه: در این بخش لیست فایل‌ها باید بر اساس میزان ارتباط مرتب شده باشد. مرتبط‌ترین سند، سندی است که تمام کلمات را به همان ترتیب موجود در پرسمان داشته باشد. (به طور مثال اگر پرسمان شامل ۳ کلمه بود، سندی مرتبط‌تر است که هر سه کلمه را داشته باشد، بعد از آن سندی مرتبط است که دو کلمه از کلمات پرسمان را در خود دارد.)

۱-۴ مجموعه داده

مجموعه داده مورد استفاده در این پروژه مجموعه‌ای از خبرهای واکشی‌شده از چند وب‌سایت خبری فارسی است که در قالب یک فایل اکسل در اختیار شما قرار خواهدگرفت. لازم است تنها ستون "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 مربوط به سند بازیابی‌شده 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 از حذف کلمات پرتکرار از واژه‌نامه و بعد از حذف کلمات پرتکرار بررسی کنید. 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؟

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.)

[2] Boolean

۵. پاسخ به پرسمان در حالت‌های زیر:

الف) یک پرسمان از کلمات ساده و متداول تک کلمه‌ای (بین‌الملل)

ب) یک پرسمان از عبارات ساده و متداول دو کلمه‌ای (دانشگاه امیرکبیر)

پ) یک پرسمان از عبارات ساده و متداول چند کلمه‌ای (دانشگاه صنعتی امیرکبیر، سازمان ملل متحد، جمهوری اسلامی ایران)

ت) یک پرسمان دشوار و کم تکرار تک کلمه‌ای (ژیمناستیک)

ث) یک پرسمان دشوار و کم تکرار دو کلمه‌ای (واکسن آسترازنکا)

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؟