« بسمه تعالی »

بازیابی اطلاعات پیشرفته

Modern Information Retrieval

مدرس: میثم روستائی

تمرین دوم | تاریخ تحویل: ۱۴۰۲/۰۳/۰۵

تمرین دوم درس بازیابی اطلاعات از دو بخش تشکیل شده است. بخش اول مربوط به پیاده‌سازی یک سیستم بازیابی اطلاعات متنی و بخش دوم به ارزیابی خروجی سیستم بازیابی اطلاعات اختصاص دارد.

بخش اول: پیاده‌سازی سیستم بازیابی اطلاعات

در این بخش، به پیاده‌سازی سیستم بازیابی اطلاعات بر اساس مجموعه اطلاعات و شاخص معکوس پیاده‌سازی شده در تمرین قبلی می‌پردازیم. روش کار سیستم به این صورت است که هر بار از کاربر یک عبارت را به عنوان ورودی (پرس‌وجو -Query) دریافت می‌کند و لیستی از اسناد مرتبط را بازیابی و بر اساس میزان شباهت با پرس‌وجو مطرح شده، رتبه‌بندی نموده و ده مورد از مرتبط‌ترین‌ها را ارائه می‌دهد. برای نمایش نتایج، نام فایلی که متن از آن استخراج شده، محتوای متن و میزان شباهت با پرس‌وجوی مطرح شده ارائه می‌شود.

- برای پاسخ‌دهی به این بخش می‌توانید از الگوریتم ارائه شده در Lecture 4- Page 33 استفاده کنید.

بخش دوم: ارزیابی عملکرد سیستم بازیابی اطلاعات

در این بخش به منظور ارزیابی کارایی سیستم، نیاز به مجموعه‌ای از پرس‌وجوهای آماده که اسناد مرتبط و غیرمرتبط با آنها مشخص است خواهیم داشت (مجموعه ارزیابی) که این پرس‌وجوها را می‌بایست با پردازش نام هر یک از فایل‌های موجود در دیتاست مسئله (در تمرین قبل) استخراج نمایید. بدیهی است نام یک فایل ممکن است در سال‌های مختلف تکرار شده باشد. به عنوان مثال، دو فایل 2007_volkswagen_touareg و 2008_volkswagen_touareg_2 در سال‌های ۲۰۰۷ و ۲۰۰۸ وجود دارد که پرس‌وجوی متناظر با این دو فایل، "volkswagen_touareg" می‌باشد. برای اینکه متون مرتبط با این پرس‌وجو را مشخص کنیم فرض می‌کنیم تمام داکیومنت‌هایی که در این دو فایل وجود دارند با این پرس‌وجو مرتبط هستند (و بقیه داکیومنت‌های موجود در فایل‌های دیگری که نام آنها volkswagen_touareg نیست با این پرس‌وجو غیرمرتبط هستند).

سیستم بازیابی اطلاعات پیاده سازی شده در بخش قبلی، تک تک پرس‌وجوهای مجموعه ارزیابی را خوانده و علاوه بر خروجی‌های بخش قبلی، برای هر پرس‌وجو، بر اساس رتبه اسناد بازیابی شده، میزان k-recall و k-precision را برای آن پرس‌وجو و همچنین میانگین k-recall و میانگین k-precision برای کل مجموعه ارزیابی محاسبه کند (k=1 و k=5 و k=10).

نکاتی در مورد انجام تمرین:

خروجی تمرین دوم ممکن است در گام‌های بعدی نیز استفاده شود. بنابراین بر درستی انجام تمرین اطمینان حاصل کنید. انجام تمرین در قالب گروه‌های حداکثر دو نفره امکان‌پذیر است. پاسخ تمرین را در قالب یک فایل فشرده که شامل (۱) کد برنامه، (۲) مجموعه ارزیابی تولید شده، (۳) گزارش و (۴) خروجی متنی سیستم بر روی مجموعه ارزیابی است را در مهلت تعیین شده در سامانه ارسال بفرمایید. ضمنا تاریخی برای تحویل حضوری تمرین مشخص خواهد شد که در آن هر گروه موظف به پاسخ به سوالات در مورد نحوه کد نویسی و عملکرد هر بخش می‌باشد. همه اعضای گروه باید بر کل هر بخش تسلط کامل داشته باشند.

فایل گزارش می‌بایست شامل بیان شیوه پیاده‌سازی هر یک از بخش‌ها و عملکرد و نتایج آن با توجه به مقادیر مختلف k (1, 5, 10) می‌باشد.

موفق باشید