یاعلیم

مبانی بازیابی اطلاعات و جستجوی وب

تمرین دوم

نیم‌سال دوم ۰۳-۰۲

مهلت تحویل بخش تئوری ۱۴۰۳/۰۲/۲۲

مهلت تحویل بخش عملی ۱۴۰۳/۰۲/۲۹

نکات:

۱. پاسخ تمرینات درقالب فایل پی دی اف (PDF) و صرفا در سامانه LMS قرار گیرد.
۲. بخش تئوری تمرین توسط هرنفر جداگانه تحویل داده می‌شود، در LMS دو تمرین جداگانه تئوری و عملی تعریف شده است.
۳. بخش عملی تمرین درگروه‌های حداکثر ۲ نفره انجام می‌شود. (گروه‌ها تا انتهای ترم یکسان باقی می‌مانند.)
۴. تمرین عملی درزمانی که بعدا اعلام خواهد شد، توسط حل تمرین به صورت حضوری نیز تحویل گرفته می‌شود.
۵. در تمرین عملی هم کدهای نوشته شده و هم توضیح کامل پروژه و تحلیل نتایج را در LMS بارگذاری نمایید.

بخش تئوری

۱. ضریب جاکارد را برای دو جمله ی زیر محاسبه کنید.

کوئری: دانشجویان دانشگاه صنعتی شاهرود

سند: یکی از دانشجویان دانشگاه صنعتی شاهرود هستم.

۲. کلمات دیکشنری نمایه permuterm را برای کلمه survey بنویسید و بگویید برای پرس‌وجوی "s*rvey" چه چیزی جست‌وجو می‌شود؟

۳. داده ساختارها اصلی برای جستجوی کلمات نمایه(دیکشنری) را نام ببرید و توضیح دهید چه ضوابطی را هنگام استفاده از انها باید درنظر گرفت؟

۴. فرض کنید شما دارای مجموعه‌ای از سه سند هستید که با نام‌های Doc1، Doc2 و Doc3 مشخص شده‌اند. برای هر یک از این سندها، تعداد تکرار برخی کلمات مشخص شده است. در جدول زیر، تعداد تکرار این کلمات برای هر سند آمده است:

کلمات	سند اول	سند دوم	سند سوم
تاریخ	۱۱	۲۲	۱۴
فناوری	۲۵	۶۰	۱۷
پژوهش	۸	۱۶	۵
داده	۱۸	۲۰	۱۰

همچنین فرض کنید که در مجموعه شما ۸۰۰۰۰۰۰ سند وجود دارد و برای هر کلمه موجود در جدول زیر، تعداد سندهایی که این کلمه در آنها دیده شده است (df)، ذکر شده است:

کلمات	Df

۱۲۰۰۰	تاریخ
۲۰۰۰۰	فناوری
۸۰۰۰	پژوهش
۲۵۰۰۰	داده

وزن های tf-idf برای کلمات تاریخ، فناوری، پژوهش، داده برای هر سند را محاسبه کنید. همچنین ترکیب سه‌تایی چهارکلمه فوق به عنوان پرس‌وجو، ترتیب شباهت اسناد را با استفاده از رابطه شباهت کسینوسی به دست آورید.

۵. شیوه محاسبه term-at-a-time و document-at-a-time را برای محاسبه لیست مرتب اسناد مرتبط با پرس و جو را به صورت مختصر توضیح دهید. در محاسبه شباهت شباهت کسینوسی، اندازه یک سند را چه زمانی و چطور محاسبه می‌کنیم؟
۶. نمودار ROC برای ارزیابی یک سیستم بازیابی اطلاعات چیست و چگونه رسم می‌شود؟
۷. Variable byte code را برای posting list زیر حساب کنید. درصورت امکان به جای DocID از Gap ها استفاده کنید.

Posting list: (19875, 398750, 5976250, 79651250)

بخش عملی

در سایت(https://ir.dcs.gla.ac.uk/resources/test_collections/) تعدادی مجموعه داده برای بازاریابی قرار دارد که توسط حل تمرین یکی از اونها به گروه شما تعلق می گیرد. در مجموعه داده، تعدادی سند، تعداد پرسوجو و همجنین ارتباط اسناد با پرس و جوها مشخص شده است. برنامه ای بنویسید که بردار هر سند و هر پرس و جو را در فضای برداری با روش وزن دهی tf-idf به دست آورد. سپس برای هر پرس و جو 10سند برتر مرتبط با پرس و جو را با روش شباهت کسینوسی محاسبه کنید و با مقایسه با مجموعه مرجع مقادیر صحت، یادآوری و معیار F1را برای هر پرس و جو و میانگین برای کل مجموعه داده را محاسبه کنید.