هوالحکیم

مبانی بازیابی اطلاعات و جستجو وب

نیم سال دوم ۰۲-۰۱

دانشکده مهندسی کامپیوتر

تمرین سری دوم — مهلت تحویل : ۲۲ اردیبهشت ۱۴۰۱

نکات:

۱- پاسخ تمرینات در قالب فایل پی دی اف صرفا در lms قرار گیرد.
۲- بخش تئوری تمرین توسط هر نفر جداگانه تحویل داده می شود، در lms دو تمرین جداگانه تئوری و عملی تعریف شده است.
۳- بخش عملی تمرین در گروه های حداکثر ۲ نفره انجام میشود. (گروه ها تا انتهای ترم یکسان باقی می ماند)
۴- تمرین عملی در زمانی که بعدا اعلام خواهد شد ، توسط حل تمرین به صورت حضوری نیز تحویل گرفته می شود.
۵- مهلت تمرین عملی بیشتر از تمرین تئوری است.

الف) بخش تئوری

۱- کلمات دیکشنری نمایه permuterm را برای کلمه study بنویسید و بگویید برای پرس و جوی "s*dy" چه چیزی جستجو شود؟
۲- داده ساختارها اصلی برای جستجوی کلمات نمایه (دیکشنری) را نام ببرید و توضیح دهید چه ضوابطی را هنگام استفاده از آنها باید در نظر گرفت؟

۳- ضریب جاکارد را برای دو جمله ی زیر بدست آورید.

کوئری : دانشگاه صنعتی شاهرود

سند : من در دانشگاه صنعتی درس میخوانم

۴- Variable byte code را برای posting list زیر حساب کنید. در صورت امکان به جای DocID از gap ها استفاده کنید.

Posting list(777,17743,294068,3125136)

۵- یک مجموعه داده با یک میلیون کلمه که طول متوسط هر کلمه آن ۸ کاراکتر است و هر کلمه به طور متوسط در ۱۰۰۰۰ سند آمده است، را در نظر بگیرید. میزان فضای مورد نیاز ذخیره دیکشنری (در ایندکس معکوس) با در نظر گرفتن دیکشنری به صورت رشته با بلوک را حساب کنید (بلاک هایی با اندازه k=8 وk=16)..

۶- جدول تعداد تکرار کلمات را برای 3 سندDoc1،Doc2،Doc3،مطابق با مقادیر مفروض در نظر بگیرید.

کلمات	سند اول	سند دوم	سند سوم
مبانی	۰	۳۳	۲۹
دانشگاه	۱۴	۵۰	۱۷

۰	۳۳	۳	شاهرود
۲۴	۴	۲۷	بازیابی

در کل مجموعه ۸۰۰۰۰۰ سند موجود است، و df برای هر کلمه در جدول زیر داده شده است.

dft	کلمات
۱۹۲۴۱	مبانی
۲۵۲۳۵	دانشگاه
۶۷۲۳	شاهرود
۱۸۱۶۵	بازیابی

وزن های tf-idf برای کلمات مبانی، دانشگاه، شاهرود، بازیابی برای هر سند را محاسبه کنید.همچنین برای ترکیب های سه تایی چهار کلمه فوق به عنوان پرس و جو، ترتیب شباهت اسناد را با استفاده از رابطه کسینوسی بدست آورید.

ب) بخش عملی

در سایت

http://ir.dcs.gla.ac.uk/resources/test_collections/

تعدادی مجموعه داده مرجع برای بازیابی قرار دارد که توسط حل تمرین یکی از آنها به گروه شما تعلق می‌گیرد. در مجموعه داده، تعدادی سند، تعداد پرس و جو و همچنین ارتباط اسناد با پرس و جو ها 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 ۱۰ سند برتر مرتبط با پرس و جو را با روش شباهت کسینوسی محاسبه کنید و با مقایسه با مجموعه مرجع مقادیر صحت، یادآوری و معیار F1 را برای هر پرس و جو و میانگین برای کل مجموعه داده را محاسبه کنید.

تمرین عملی نسبتا ساده است، اما اگر در انجام تمرین عملی مشکل دارید حتما با حل تمرین و استاد درس مشورت کنید و تمرین را انجام دهید.

همچنین اگر قسمت عملی تمرین اول را انجام نداده اید، با حل تمرین و استاد درس مشورت کنید و با در نظر گرفتن تاخیر انجام دهید چون در پروژه نهایی درس به بخش Lucene از تمرین اول نیاز دارید.