هوالحکیم

مبانی بازیابی اطلاعات و جستجو

وب

نیم سال دوم ۰۱-۰۲

دانشکده مهندسی کامپیوتر

تمرین سری اول | مهلت تحویل : ۱۸فروردین ۱۴۰۱

نکات:

۱- مهلت تمرین عملی از تئوری بیشتر بوده و در lms مشخص است. در تمرین عملی هم کدهای نوشته شده و هم توضیح کامل پروژه و تحلیل نتایج را در lms بارگذاری کنید.

۲- پاسخ تمرینات در قالب فایل پی دی اف صرفا در lms قرار گیرد.

۳- بخش تئوری تمرین توسط هر نفر جداگانه تحویل داده می شود، در lms دو تمرین جداگانه تئوری و عملی تعریف شده است.

۴- بخش عملی تمرین در گروه های حداکثر ۲ نفره انجام میشود. (گروه ها تا انتهای ترم یکسان باقی می ماند)

۵- تمرین عملی در زمانی که بعدا اعلام خواهد شد، توسط حل تمرین به صورت حضوری نیز تحویل گرفته می شود.

الف) بخش تئوری

۱- ایندکس معکوس (inverted index) و ماتریس رخداد (incidence matrix) را برای مستندات زیر رسم کنید.

سند اول : من در رشته مهندسی کامپیوتر تحصیل می‌کنم.

سند دوم : یکی از دروس رشته مهندسی کامپیوتر مبانی بازیابی اطلاعات است.

سند سوم: درس مبانی بازیابی اطلاعات درسی سه واحدی است.

سند چهارم: من این ترم فارغ التحصیل می‌شوم.

۲- با توجه به سوال یک به کوئری های زیر پاسخ دهید . به دو روش ماتریس رخداد و ایندکس معکوس

الف) بازیابی OR رشته

ب) رشته AND مبانی

ج) من AND (میشوم NOT)

د) مهندسی کامپیوتر AND (بازیابی AND اطلاعات)NOT

۳- یک سیستم بازیابی اطلاعات متنی که در دنیای واقعی مورد استفاده قرار می گیرد را مثال بزنید و اجزای اصلی معماری آن را بیان کرده و مفهوم کارایی Efficiency را در آن شرح دهید.

۴- با توجه به اندازه posting list ها ترتیب اجزای عملیات های پرس و جو زیر را مشخص کنید.

(computer OR lesson) AND (database OR me) AND (list OR city)

Term	Postings Size
Computer	316812
Lesson	46653
Database	107913
Me	271658
List	87009
City	213312

۵- برای کوئری های مشابع سوال سوم (conjunctive queries) آیا روش ترتیب پردازش وابسته به سایز Posting List ها همیشه کم هزینه ترین روش است؟ دلیل خود را توضیح دهید.

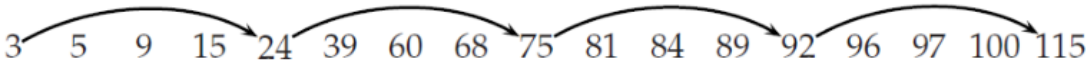
۶- درستی و نادرستی عبارات زیر را با توضیح مشخص کنید.

الف) ریشه‌یابی سبب افزایش اندازه واژگان (vocabulary) میشود.

ب) ریشه‌یابی باید درهنگام نمایه سازی (Indexing) فراخوانی شود نه درهنگام پردازش یک پرس وجو.

ج) در سیستم بازیابی بولی، ریشه‌یابی هرگزمقدار Recall را کاهش نمیدهد.

۷- قصد داریم بین لیست پست زیر (لیست اول با اشاره‌گر پرش) و نتایج عملیات گذشته (لیست دوم)، اشتراک بگیریم.



3 5 84 94 96 97 100 101 116

مطابق با الگوریتم intersectionPostings در صفحه 26 اسلاید (the term vocabulary and postings lists) به سوالات زیر پاسخ دهید.

الف) چند بارskip pointer طی می‌شود ؟ (یعنی p1 به (p2) skip پرش انجام می‌دهد)

ب) چه تعداد مقایسه postings توسط این الگوریتم درحالی که اشتراک دو لیست را انجام می‌دهد، اتفاق می‌افتد؟

ج) چه تعداد مقایسه postings باید انجام می پذیرفت، اگر posting list هابدون استفاده ازاشاره گرهای پرش،با هم اشتراک

گرفته شوند؟

۸- سند های زیر را در نظر بگیرید.

Doc1 : I am a student, and I currently attend the IR course.

Doc2 : I was a student; I have taken the IR course.

ایندکس مکانی (positional index) هر کدام از کلمات "student" و "I " را بدست آورید و به پرس وجوهای زیر پاسخ دهید.

کوئری اول : "I student"

کوئری دوم : "student I"

۹- چگونه یک سیستم IR میتواند استفاده از positional index واستفاده ازلیست واژه ها (stop words) را با هم ترکیب کند؟ مشکل احتمالی در این فرآیند چیست و چگونه می‌توان آن را مدیریت کرد؟

۱۰- برای کلمات زیر ساختار درختی را برای ذخیره دیکشنری ایجاد کنید:

able, ability, about, back, backup, bad, badly

۱۱- کلمات دیکشنری و نمایه permuterm را به صورت درختی برای کلمه paper بنویسید و بگویید برای پرس و جوهای "pa*r" و "*pe*"چه عبارتی جستجو می‌شود ؟

۱۲-

۱۳- فاصله کلمات "background" و "backboard" را توسط الگوریتم levenshtein بیابید.

ب) بخش عملی

۱- برنامه ای بنویسید که تعدادی از اسناد به دلخواه دریافت کند و پرس و جو هایی را از کاربر دریافت کرده (شامل AND و OR و NOT) و به روش ایندکس معکوس اسناد مرتبط با آن را بازگرداند. اسناد را می‌توانید از مجموعه داده‌های متنی از مخزن های متداول وب مثل Kaggle و UCI برداشت کنید. تعداد سندها نیازی نیست که زیاد باشد، و انجام کارهای پیش پردازش در این بخش نیازی نیست. نتایج پرس و جو را تحلیل کرده و درستی آنها را بررسی کنید.

۲- فایل‌های متنی تمرین قبل را (در تعداد بیشتر) به وسیله نرم‌افزار Lucene ایندکس کنید. همه حروف را کوچک کنید و کلمات ایست را حذف کنید. ریشه یابی را به دلخواه خود می توانید انجام دهید (این امکانات در Lucene موجود است). سپس حداقل پنج پرس و جو مطرح کنید که عملگرهای and، or، not و جستجوی عبارت و پرس و جوی wildcard را شامل شوند و نتایج را بررسی کرده و گزارش کنید که آیا نتایج و رتبه بندی آنها منطقی هستند یا خیر.