به نام خدا

دانشگاه صنعتی شاهرود

دانشکده مهندسی کامپیوتر

مبانی بازیابی اطلاعات و جستجوی وب

تمرین سری اول

مهلت تحویل: ۱۰ آبان ۱۴۰۰

- لطفا تمرین‌ها را در قالب یک فایل پی دی اف صرفا در بخش ارسال تمرین‌های سیستم مدیریت آموزش مجازی دانشگاه آپلود کنید.
- هر سری از تمرین‌ها دارای مهلت تحویل بوده و باید در زمان معین بارگذاری شوند ،به ازای هر روز تاخیر در ارسال تمرین‌ها، مقداری از نمره کسر خواهد شد.
- تمامی تمرین‌ها به صورت تایپ شده یا به صورت اسکن شده از دست نویس (خوانا و خوش خط) تحویل گرفته خواهد شد. برای هر سری از تمرین‌ها، حتما صفحه اولی که در آن نام، نام خانوادگی، شماره دانشجویی و عنوان تمرین وجود دارد، در نظر گرفته شود.
- متن تمرین را پس از مطالعه درس به زبان خودتان بنویسید (لطفا کپی نکنید!)

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

سند ۱ : من درس بازیابی اطلاعات را در دی ماه پاس می‌کنم.

سند ۲: یکی از دروس مهندسی کامپیوتر بازیابی اطلاعات است.

سند ۳ : من این ترم درس بازیابی اطلاعات دارم.

سند ۴ : این ترم بیست واحد دارم.

۲- با توجه به سوال ۱ به کوئری های زیر پاسخ دهید. به دو روش ماتریس رخداد و ایندکس معکوس

الف) بازیابی AND اطلاعات

ب) درس OR ترم

ج) واحد AND (دارم NOT)

د) مهندسی کامپیوتر AND (بازیابی AND اطلاعات) NOT

Computer Engineering AND NOT (Information AND Retrieval)

۳- یک سیستم بازیابی اطلاعات متنی که در دنیای واقعی مورد استفاده قرار می‌گیرد 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 اجرای عملیات‌های پرس وجوی زیر را مشخص کنید

(tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes)

Term	Postings size
eyes	316812
kaleidoscope	46653
marmalade	107913
skies	271658
tangerine	87009
trees	213312

۵- برای کوئری های مشابه سوال ۳ (conjunctive queries) آیا روش ترتیب پردازش وابسته به سایز Posting List ها همیشه کم هزینه ترین روش است؟ در صورتی که موافق هستید، دلیل خود را توضیح دهید.

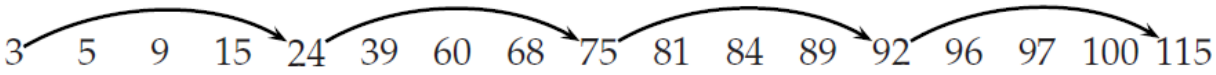
۶- درستی و نادرستی عبارات زیر را مشخص کنید:

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

ب) Stemming باید در هنگام نمایه‌سازی (Indexing) فراخوانی شود نه در هنگام پردازش یک پرسوجو.

ج) در سیستم بازیابی بولین، stemming هرگز مقدار Recall را کاهش نمی‌دهد.

۷- یک Postings intersection بین postings list با skip pointers را به صورت زیر در نظر بگیرید:



و میانگین نتایج حاصله‌ی postings list مطابق زیر است (از این رو هیچ اشاره گر پرشی ندارد):

3 5 89 95 97 99 100 101

مطابق با الگوریتم Postings intersection در صفحه ۲۶ اسلاید (the term vocabulary and postings lists) به سوالات زیر پاسخ دهید.

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

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

ج) چه تعداد مقایسه postings باید انجام می‌پذیرفت، اگر posting list ها بدون استفاده از اشاره گرهای پرش، با هم اشتراک گرفته شوند؟

۸- سندهای زیر را در نظر بگیرید.

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

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

positional index هر کدام از کلمات "student" و " I " را بدست آورید و به پرس‌وجوهای زیر پاسخ دهید

Query1: "I student"
Query2:"student I"

۹- چگونه یک سیستم IR می تواند استفاده از positional index و استفاده از ایست واژه‌ها (stop words) را با هم ترکیب کند؟ مشکل احتمالی در این فرآیند چیست و چگونه می توان آن را مدیریت کرد؟

۱۰- برای پرس‌وجوهای زیر آیا می‌توانیم عملیات اشتراک را در زمان O(x+y) با در نظر گرفتن اینکه x و y طول postings lists برای Brutus وCaesar اجرا کنیم؟ اگر اینطور نیست پس چگونه ما میتوانیم آن را محاسبه کنیم ؟

الف) Brutus **AND NOT** Caesar

ب) Brutus **OR NOT** Caesar