Quiz — 2
MIR — Spring 2026
Instructor: Dr. Amin Eskandari
Head Teaching Assistants: Hamid Namjoo, Amir Hossein Hemati
Teaching Assistants: Ali Mojahed, Ali Nikvan, Amir Javvi

سوال ۱

شما به‌عنوان مهندس ارشد یک کتابخانه دیجیتال ملی استخدام شده‌اید. این کتابخانه ۱۰ میلیون سند متنی دارد. بودجه شما محدود است و تنها یک سرور با **۲ گیگابایت RAM** در اختیار دارید. کاربران کتابخانه (که عمدتاً ریسرچر هستند) باید بتوانند پرسمان‌های بولی پیچیده شامل عملگرهای AND، OR و NOT همچنین پرسمان‌های عبارتی (با استفاده از " ")را جستجو کنند. برای مثال، یک پرسمان نمونه می‌تواند این باشد:

("هوش مصنوعی" AND "یادگیری ماشین") OR ("شبکه عصبی" AND NOT "پرسپترون")

تحلیل شما نشان می‌دهد که ماتریس وقوع بولی (Term-Document Incidence Matrix) برای این حجم داده به فضایی بیش از **۴۶۰ گیگابایت** نیاز دارد که عملاً غیرممکن است.

الف) شما باید یک فهرست معکوس (Inverted Index) برای این مجموعه بسازید. با توجه به محدودیت ۲ گیگابایت RAM ، آیا الگوریتم **BSBI (Blocked Sort-Based Indexing)** مناسب‌تر است یا **SPIMI (Single-Pass In-Memory Indexing)**؟ با ذکر دلیل، تحلیل پیچیدگی و حافظه مصرفی هر یک، روش برتر را انتخاب کرده ضمنا مراحل اجرای آن را با **جزئیات کامل** شرح دهید.

نکته: توضیح دهید که چگونه مشکل نگاشت term به termID در این دو الگوریتم بر مصرف حافظه تأثیر می‌گذارد.

ب) برای پشتیبانی از پرسمان‌های عبارتی مانند "هوش مصنوعی "در کنار پرسمان‌های بولی، چه نوع فهرستی باید به فهرست معکوس خود اضافه کنید؟ ساختار یک posting در این فهرست جدید را به‌دقت شرح دهید.

ج) یک پرسمان بولی ترکیبی مانند مثال بالا را در نظر بگیرید. شما دو گزینه برای پردازش دارید:

۱. ابتدا تمام زیرعبارات عبارتی را با استفاده از فهرست موقعیتی حل کنید، سپس نتایج را با عملگرهای بولی (AND , OR) ترکیب کنید.

۲. ابتدا پرسمان بولی کلی را با استفاده از فهرست غیرموقعیتی (ساده) حل کنید و سپس نتایج نهایی را با استفاده از فهرست موقعیتی برای اعمال قید عبارت فیلتر کنید.

MIR — Spring 2026

کدام استراتژی بهینه‌تر است و چرا؟ تحلیل کنید که اگر از استراتژی دوم استفاده کنید، چه خطای مفهومی بزرگی رخ خواهد داد؟ (هینت: به تفاوت معنایی "A B" AND C در برابر A AND B AND C فکر کنید..)

سوال ۲

شما در تیم جستجوی یک فروشگاه اینترنتی بزرگ (مثل آمازون) کار می‌کنید. گزارش‌ها نشان می‌دهد که ۱۵٪ از جستجوهای کاربران برای محصول "Samsung Galaxy" به صورت اشتباه تایپ می‌شود؛ مثلاً Samsang، Galaxi، یا Samsng Glaxy می‌باشد. مدیر محصول از شما می‌خواهد یک سیستم **پیشنهاد اصلاح املایی (Spelling Correction)** بسیار سریع و دقیق پیاده‌سازی کنید که بتواند در کسری از ثانیه، پیشنهاد درست را به کاربر نمایش دهد.

الف) شما دو روش پیشنهادی دارید:

۱) محاسبه **فاصله ویرایشی (Edit Distance)** با الگوریتم برنامه‌نویسی پویا بین کلمه اشتباه و *تمام* کلمات دیکشنری محصولات (که ۵ میلیون کلمه یکتا دارد).

۲) استفاده از (**ایندکس k-gram با k=3**) به همراه **ضریب Jaccard** برای فیلتر کردن اولیه و سپس محاسبه فاصله ویرایشی روی کاندیداهای محدود.

با تحلیل پیچیدگی زمانی و محاسباتی، **ثابت کنید** چرا روش اول برای این مقیاس کاملاً غیرعملی و روش دوم یک راه‌حل صنعتی و استاندارد است. همچنین فرمول ضریب Jaccard را بر اساس k-gram ها بنویسید.

ب) فرض کنید کاربر "Samsang Galxy" را جستجو کرده است. با استفاده از ایندکس ۳ gram-که ساخته‌اید، مراحل دقیق پیدا کردن پیشنهاد "Samsung Galaxy"را توضیح دهید. به طور خاص، توضیح دهید که چرا کلمه "Samsang"می‌تواند پیشنهاد "Samsung"را دریافت کند، اما کلمه "Samsoon"ممکن است پیشنهاد دیگری بگیرد.

ج) در برخی موارد، پیشنهاد اصلاح املایی تک‌کلمه‌ای کافی نیست. برای عبارت "Samsng Glaxy"، ممکن است سیستم به اشتباه "Samson Glaze"را پیشنهاد دهد اگر اصلاحات را مستقل از هم انجام دهد. برای حل این مشکل در یک سیستم واقعی با میلیون‌ها عبارت جستجو، چه راهکاری بر

اساس **مدل بولی گسترش‌یافته و پرسمان‌های نزدیکی (Proximity Queries)** یا **آمار دوواژه‌ای (Biword Statistics)** ارائه می‌دهید؟ به طور مشخص، فرآیند پیشنهاد "Samsung Galaxy" به جای "Samson Glaze" را گام‌به‌گام شرح دهید.

سوال ۳

شما مدیریت سیستم جستجوی یک خبرگزاری بزرگ را بر عهده دارید که روزانه ۵ میلیون خبر جدید منتشر می‌کند. کاربران انتظار دارند اخبار جدید حداکثر **۵ دقیقه** پس از انتشار قابل جستجو باشند. شما یک فهرست معکوس اصلی (Main Index) دارید که ۱ میلیارد سند قدیمی را پوشش می‌دهد. ساخت مجدد کامل این فهرست اصلی از ابتدا، با توجه به حجم آن، **بیش از ۱۲ ساعت** زمان می‌برد.

الف) چرا راهکار «بازسازی کامل ایندکس هر شب» برای این سناریو غیرقابل قبول است؟ چه روش‌های **ایندکس‌سازی پویا (Dynamic Indexing)** برای حل این مشکل در کلاس معرفی شده‌اند؟ دو رویکرد اصلی و معایب هر یک را نام ببرید.

ب) شما تصمیم می‌گیرید از روش **ادغام لگاریتمی (Logarithmic Merging)** استفاده کنید. ساختار این ایندکس‌ها (شامل Z0 و ... ,I0, I1, I2) را با فرض n = 2 (==ظرفیت ایندکس کمکی درRAM) توضیح دهید. با یک مثال عددی، نشان دهید که چگونه با اضافه شدن تدریجی اسناد جدید، ایندکس‌ها ساخته و ادغام می‌شوند. (مثال برای ۸ سند اول بزنید)

ج) در این ساختار، برای پاسخ به یک پرسمان، باید آن را روی **تمام ایندکس‌ها** (چه ایندکس کمکی در RAM و چه ایندکس‌های روی دیسک ... ,I0, I1) اجرا کرد و سپس نتایج را با هم **ادغام** نمود. با توجه به اینکه اسناد خبری بسیار مهم هستند، شما می‌خواهید نتایج جستجو **بر اساس زمان (جدیدترین‌ها اول)** مرتب شوند. این استراتژی چه تأثیری بر پیاده‌سازی posting listها در هر یک از ایندکس‌های I0, I1, I2ادارد؟ آیا posting listها همچنان باید بر اساس docID مرتب باشند یا یک استراتژی دیگر (مثلاً بر اساس بردار تأثیر یا زمان) بهتر است؟ تحلیل کنید و بهترین رویکرد را شرح دهید.

سوال ٤

شما به عنوان مشاور ارشد یک استارتاپ جستجوی سازمانی استخدام شده‌اید. تیم فنی این استارتاپ، که عمدتاً مهندسان نرم‌افزار خبره اما کم‌تجربه در حوزه بازیابی اطلاعات هستند، یک طراحی اولیه برای موتور جستجوی خود ارائه داده‌اند که به شدت به آن افتخار می‌کنند. آنها ادعا می‌کنند این طراحی «فوق‌العاده بهینه و سریع» است. طراحی پیشنهادی آنها برای ایندکس اصلی شامل سه تصمیم کلیدی زیر است:

۱. **دیکشنری:** برای سرعت دسترسی O(1)، از یک **جدول هش (Hash Table)** بزرگ برای نگهداری تمام termهای دیکشنری استفاده شده است.

۲. **پرسمان‌های عبارتی:** برای پشتیبانی از پرسمان‌های عبارتی (مانند "fast car"، به جای ایندکس موقعیتی که «حجیم و کند» است، از یک **ایندکس دوواژه‌ای (Biword Index)** استفاده کرده‌اند. آنها معتقدند با این کار، هر عبارت حداکثر ۲-کلمه‌ای به یک جستجوی ساده در دیکشنری تبدیل می‌شود.

۳. **بهینه‌سازی پرسمان‌های ترکیبی:** برای پرسمانی مانند

("artificial intelligence" OR "machine learning") AND NOT "deep learning"

الگوریتم آنها به این صورت است: ابتدا تمام posting listهای مربوط به تک‌کلمات (artificial, intelligence, machine, learning, deep) را از دیسک می‌خواند، سپس یک عملیات بولی سنگین روی این لیست‌ها انجام می‌دهد و **در انتها** بررسی می‌کند که آیا کلمات artificial و intelligence به صورت یک عبارت در کنار هم آمده‌اند یا خیر.

حالا شما با توجه به دانشتون تا اینجای درس:

الف) تیم فنی می‌گوید: «دسترسی به دیکشنری در O(1) عالی است!» شما بلافاصله دو قابلیت حیاتی که در نیازمندی‌های سیستم ذکر شده را به آنها گوشزد می‌کنید:

---

۱. کاربران باید بتوانند از پرسمان‌های wildcard مانند econ* برای جستجوی economy, economics, economical استفاده کنند.
۲. سیستم باید یک پیشنهاد دهنده (Auto-Complete) داشته باشد که با تایپ comp، کلماتی مانند computer, company, compare را پیشنهاد دهد.

MIR Spring 2026

سوال :با یک استدلال فنی دقیق توضیح دهید که چرا ساختار هش برای پشتیبانی از این دو قابلیت **کاملاً نامناسب** است. چه ساختار داده جایگزینی (هینت: تدریس شده:)) معرفی شده که این قابلیت‌ها را فراهم می‌کند و پیچیدگی جستجوی آن چقدر است؟ **تله‌ای** که مهندسان در آن افتاده‌اند چیست؟

ب) تیم فنی افتخار می‌کند که با ایندکس دوواژه‌ای، پرسمان "fast car" خیلی سریع حل می‌شود. شما یک پرسمان دیگر از نیازمندی‌ها را مطرح می‌کنید: «کاربران ما وکیل هستند و اغلب پرسمان‌های نزدیکی (Proximity Queries) جستجو می‌کنند، مثلاً "breach" /3 "contract" یعنی دو کلمه contract و breach حداکثر با فاصله ۳ کلمه از هم.

سوال :ثابت کنید که ایندکس دوواژه‌ای **هیچ راه‌حل کارآمدی** برای این نوع پرسمان ندارد. سپس فرمول یا روش استاندارد برای پاسخ به این پرسمان با استفاده از **ایندکس موقعیتی** را شرح دهید. چرا این روش از نظر ریاضی و الگوریتمی برای پرسمان‌های نزدیکی مناسب است؟

ج) فرض کنید در مجموعه اسناد، واژه "artificial" در ۱۰ میلیون سند، "intelligence" در ۸ میلیون سند، و عبارت دقیق "artificial intelligence" فقط در ۵۰ هزار سند وجود دارد. استراتژی پردازش تیم فنی (خواندن همه لیست‌های بزرگ و فیلتر در انتها) را با استراتژی بهینه (پردازش عبارت اول با ایندکس موقعیتی و سپس ترکیب) مقایسه کنید.

سوال :با استفاده از مفهوم **فراوانی سند (Document Frequency)** و الگوریتم INTERSECT تعداد *تقریبی* مقایسه‌های docID که در هر دو روش انجام می‌شود را محاسبه کنید. نشان دهید که روش تیم فنی چندین مرتبه بزرگتر (orders of magnitude) کندتر است؟ و **تله‌ای** که در تفکر الگوریتمی آنها بوده، چی بوده؟

موفق و موید :(

MIR — Spring 2026