

Quiz — 1

MIR — Spring 2026

Instructor: Dr. Amin Eskandari

Head Teaching Assistants: Hamid Namjoo, Amir Hossein Hemati

Teaching Assistants: Ali Mojahed, Ali Nikvan, Amir Javvi

سوال 1

فرض کنید یک موتور جستجو از فهرست معکوس (Inverted Index) با لیست‌های مرتب‌شده استفاده می‌کند

Posting List های مربوط به واژگان به صورت زیر هستند :

$A \rightarrow (1, 3, 4, 8, 10)$

$B \rightarrow (1, 2, 4, 8)$

$C \rightarrow (2, 4, 6, 8, 10)$

$D \rightarrow (1, 4, 7, 8)$

کاربر کوئری زیر را وارد می‌کند :

$(A \text{ OR } B) \text{ AND } (C \text{ OR } D)$

بخش الف:

کوئری بالا را در مدل بولی (Boolean Model) ارزیابی کنید و شناسه اسناد (Document IDs) بازیابی‌شده را مشخص نمایید.

بخش ب:

یک دانشجو برای افزایش سرعت اجرا، کوئری بالا را به شکل زیر بازنویسی می‌کند :

$(A \text{ AND } C) \text{ OR } (A \text{ AND } D) \text{ OR } (B \text{ AND } C) \text{ OR } (B \text{ AND } D)$

او ادعا می‌کند :

این دو کوئری همیشه دقیقاً نتیجه یکسانی تولید می‌کنند آیا این ادعا درست است؟

-اگر درست است: دلیل خود را توضیح دهید.

-اگر نادرست است: یک مثال نقض ارائه دهید.

بخش ج:

فرض کنید اندازه Posting List ها (تعداد اسناد شامل هر واژه) به صورت زیر است :

$A : 1,000,000$

$B : 800,000$

C : 10,000

D : 5,000

یک دانشجو پیشنهاد می‌دهد :

برای اجرای سریع‌تر کوئری باید همیشه کوچک‌ترین لیست‌ها را ابتدا Intersect کنیم
آیا این استراتژی برای کوئری داده‌شده همیشه بهینه است؟
پاسخ خود را با تحلیل دقیق توضیح دهید.

بخش د:

فرض کنید موتور جستجو به جای Boolean Retrieval از Ranked Retrieval مانند TF-IDF استفاده کند.

اما ایندکس موجود فقط Boolean است؛ یعنی فقط حضور یا عدم حضور هر واژه در سند ذخیره شده و هیچ اطلاعات فرکانسی نگهداری نمی‌شود.

سؤال:

آیا در چنین سیستمی می‌توان نتایج رتبه‌بندی‌شده معنادار تولید کرد؟
اگر پاسخ مثبت است: روش انجام آن را توضیح دهید.

اگر پاسخ منفی است: مشخص کنید چه اطلاعاتی در ایندکس وجود ندارد و چرا بدون آن‌ها رتبه‌بندی امکان‌پذیر نیست.

سوال 2

فرض کنید مجموعه اسناد زیر در یک سیستم بازیابی اطلاعات ذخیره شده است :

D1: web search engines use ranking models

D2: boolean retrieval model for web search

D3: information retrieval models and ranking techniques

D4: web mining techniques for search systems

D5: boolean query processing in retrieval systems

D6: ranking models for information systems

فرض کنید:

همه واژه‌ها به حروف کوچک تبدیل شده‌اند.

stopword حذف نشده‌اند.

stemming انجام نشده است.

واژگان مورد بررسی:

web

search

retrieval

ranking

boolean

models

بخش الف:

برای واژگان داده شده:

1. ماتریس وقوع بولی (Boolean Incidence Matrix) را برای اسناد D1 تا D6 بسازید.

2. بر اساس همین داده‌ها فهرست معکوس (Inverted Index) را نیز ایجاد کنید.

بخش ب:

کوئری زیر را در نظر بگیرید:

(web AND search) OR (retrieval AND ranking) OR boolean

این کوئری را به صورت مرحله به مرحله با استفاده از فهرست معکوس ارزیابی کنید و اسناد نهایی بازیابی شده را مشخص نمایید.

بخش ج:

یک دانشجو پیشنهاد می‌دهد که برای اجرای سریع‌تر کوئری باید آن را به صورت زیر بازنویسی کرد:

(web OR retrieval OR boolean)

AND

(search OR ranking OR boolean)

1. آیا این دو پرسوجو از نظر منطق بولی معادل هستند؟
2. اگر معادل نیستند، یک مثال از اسنادی که در یکی از پرسوجوها بازیابی می‌شوند ولی در دیگری نه، ارائه دهید.
3. توضیح دهید چرا چنین بازنویسی‌هایی در سیستم‌های جستجوی واقعی می‌تواند خطرناک باشد.

بخش د

فرض کنید اندازه Posting List ها در مجموعه‌ای بسیار بزرگ از اسناد به صورت زیر است :

web → 40M

search → 30M

retrieval → 500K

ranking → 200K

boolean → 20K

models → 5M

پرسوجوی زیر را در نظر بگیرید :

(web AND search AND models)

OR

(retrieval AND ranking AND boolean)

1. یک طرح اجرای بهینه (Optimal Execution Plan) برای این کوئری پیشنهاد دهید.
2. توضیح دهید چرا ترتیب اجرای انتخاب‌شده کمترین هزینه محاسباتی را دارد.
3. نشان دهید اگر ترتیب اجرا اشتباه انتخاب شود، اندازه نتایج میانی چگونه می‌تواند چندین برابر بزرگ‌تر شود.

بخش ه

فرض کنید سیستم فقط از Boolean Index استفاده می‌کند و هیچ اطلاعاتی درباره:

-تعداد تکرار واژه‌ها

-طول سند
-وزن واژه‌ها
ذخیره نشده است.

یک کاربر کوئری زیر را وارد می‌کند:

information retrieval models ranking

1. در مدل بولی ساده چه اسنادی بازیابی می‌شوند؟
2. توضیح دهید چرا این مدل نمی‌تواند تفاوت معنی‌داری بین اسناد مرتبط ایجاد کند.
3. توضیح دهید چگونه Extended Boolean Model می‌تواند مفهوم partial matching را معرفی کند.
4. توضیح دهید چرا Ranked Retrieval مانند TF-IDF یا BM25 معمولاً نتایج بسیار بهتری تولید می‌کند.

بخش و:

فرض کنید مجبور هستید یک موتور جستجو بسازید که:

-فقط از فهرست معکوس بولی استفاده کند.

-هیچ اطلاعات فرکانسی ذخیره نکند.

اما همچنان باید نوعی رتبه‌بندی نتایج ارائه دهد.

یک روش طراحی کنید که:

1. فقط با اطلاعات موجود در Boolean Index کار کند.
2. بتواند نتایج را تا حدی مرتب کند.
3. توضیح دهید محدودیت‌های اساسی این روش چیست.

سوال 3

نبرد در دیتابیس «کونوها» (چالش بازیابی پیشرفته)

فرض کنید شما مسئول بخش فناوری دهکده کونوها هستید و وظیفه دارید سیستمی طراحی کنید تا در میان میلیون‌ها گزارش مأموریت، نینجاهای مناسب را پیدا کنید. دیتابیس شما شامل دو واژه کلیدی با شرایط خاص است:

واژه‌ی «شارینگان» (Sharingan): واژه‌ای بسیار رایج که در ۵۰۰,۰۰۰ گزارش آمده است (لیست ارسال بسیار طولانی).

واژه‌ی «آماتراسو» (Amaterasu): تکنیکی نادر که فقط در ۵۰ گزارش آمده است (لیست ارسال بسیار کوتاه).

الف) چالش سرعت (Skip Pointers):

اگر بخواهید نینجاهایی را پیدا کنید که هر دو تکنیک را دارند (Sharingan AND Amaterasu)، استفاده از الگوریتم دو نشانگر معمولی (Linear Scan) باعث می‌شود سیستم برای هر ۵۰ مورد «آماتراسو»، تمام ۵۰۰,۰۰۰ مورد «شارینگان» را چک کند.

توضیح دهید ساختار Skip Pointers (اشاره‌گرهای پرشی) چگونه می‌تواند در این نبرد نابرابر، زمان جستجو را به شدت کاهش دهد؟ (نحوه ایجاد پرش‌ها در لیست «شارینگان» را تحلیل کنید)

ب) چالش ذخیره‌سازی (Data Structures):

در آرشیو نینجاها، اطلاعات برخی تکنیک‌ها مدام به‌روزرسانی می‌شود (مثل تکنیک‌های جدید) اما برخی تکنیک‌های قدیمی ثابت هستند. برای ذخیره لیست مأموریت‌های مربوط به یک تکنیک قدیمی و ثابت (با لیست طولانی) و یک تکنیک جدید و در حال تغییر (با لیست کوتاه)، کدام ساختار را انتخاب می‌کنید؟ (Variable-length Array یا Linked List)؟ تضاد بین «سرعت جستجو» و «سهولت در درج (Insert)» را در پاسخ خود لحاظ کنید.

پ) چالش مکان‌مندی (Positional Indexing):

فرض کنید می‌خواهید مأموریت‌هایی را پیدا کنید که در آن‌ها عبارت «ناروتو» و «ساسکه» دقیقاً کنار هم یا با فاصله حداکثر ۲ کلمه از هم آمده باشند (نشان‌دهنده همکاری نزدیک).
چرا یک Inverted Index معمولی که فقط شناسه اسناد (DocID) را ذخیره می‌کند، برای پاسخ به این پرسش ناتوان است؟

سوال 4

جاوید قصد دارد بازی world of warcraft شروع کند و یک سری از دانجن هارو بزنه برای این کار نیاز به تیم (اسکواد) دارد. اولین چیزی که تیمش نیاز دارد یک تانک که بتونه ضربات تحمل کنه دومین یاری که نیاز به هیلر یا ساپورتر که بتونه هیل تانک و تیم نگه دارد، کلاس (هیرو) خود جاوید DPS میلی (دمیج زن از راه نزدیک) کلاسش سرعت زیادی هم ندارد پس کل تیم باید سرعتشون بالا باشه. پس تیم نیاز به دوتا (DPS Range) دارد. ضمناً دیتابیس بازی شامل یک میلیون پلیئر (بازیکن) است!

الف) یک پرسمان بولی بنویسید که بازیکنانی که "تانک" یا "هیلر" هستند را پیدا کند، به شرطی که "سرعت پایین" نباشند.

ب) اگر لیست "تانک‌ها" ۵۰۰۰ نفر و لیست "هیلرها" ۱۰۰ نفر باشد، طبق استراتژی Query Optimization، سیستم باید ابتدا کدام لیست را بررسی کند؟ چرا؟

پ) الگوریتم Intersection (اشتراک) را برای پیدا کردن بازیکنی که هم "تانک" است و هم "دمیج‌زن راه دور" با رسم دو لیست فرضی و نشانگرها (Pointers) نمایش دهید.

ج) اگر هزینه پردازش عملگر NOT بیشتر از OR باشد، آیا باز هم ترتیب شروع از کوچکترین لیست صادق است؟