Quiz — 4
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

علی و رویا در حال طراحی یک موتور جستجو برای آرشیو پیام‌های یک شبکه اجتماعی مثل اینستاگرام هستند

در گزارش ویژگی داده‌ها گفتند توزیع فراوانی واژه‌ها از قانون Zipf پیروی می‌کند و رشد تعداد واژه‌های یکتا با تعداد اسناد از قانون Heaps تبعیت می‌کنه

$V(N)=kN^{\beta}$ ، $0.4 < \beta < 0.7$

علی ایده جدیدی برای کاهش فضای ایندکس ها پیشنهاد کرد که شامل این ویژگی هاست

Dictionary با Front Coding فشرده شده

Posting List ها به صورت Gap ذخیره شده

Gap ها با Gamma Encoding فشرده شده

اما پس از اضافه شدن حجم بزرگی از اسناد جدید، رفتارهای زیر مشاهده شد

1) اندازه Dictionary کمتر از انتظار رشد کرده

2) اندازه Posting File بسیار بیشتر از انتظار رشد کرده

3) برای برخی واژه‌های بسیار پرتکرار نسبت فشرده‌سازی Gamma Encoding بدتر از بعضی واژه‌های کم‌تکرار شده

رویا ادعا می‌کند که این رفتارها با یکدیگر ناسازگارند و حداقل یکی از قوانین Zipf یا Heaps در داده‌ها برقرار نیست

الف) تحلیل کنید آیا ادعای رویا درست است یا خیر؟؟

در صورت نادرست بودن توضیح دهید که چگونه ممکن است هر سه مشاهده بالا به صورت همزمان رخ دهد بدون آنکه هیچ‌کدام از قوانین Zipf و Heaps نقض شده باشد!؟

در پاسخ خود ارتباط بین مفاهیم زیر را به صورت کامل و دقیق تحلیل کنید.

قانون Zipf

قانون Heaps

Front Coding

Gap-based Posting Lists

Gamma Encoding

اثر ترتیب DocID ها بر فشرده‌سازی

ب) فرض کنید فقط «یک تغییر معماری» در سیستم مجاز است تنها یکی از گزینه‌های زیر قابل تغییر هست.

1) روش مرتب‌سازی و تخصیص DocID

2) روش فشرده‌سازی Dictionary

3) روش Encoding مربوط به Posting List ها

مشخص کنید کدام گزینه بیشترین کاهش فضای نهایی ایندکس را ایجاد می‌کند و دلیل انتخاب خود را کامل توضیح دهید

(دقت کنید وقتی میگم کامل تحلیل کنید یعنی این موارد داخل جوابتون باشه)

اثر قوانین Zipf و Heaps را در تحلیل دخیل کنید.

Trade-off های طراحی را بررسی کنید.

در اخر توضیح دهید چرا دو گزینه دیگر در این سناریو تأثیر کمتری دارد

سوال 2

یک شرکت بزرگ فناوری موتور جستجویی برای رتبه‌بندی رزومه متقاضیان استخدام پژوهشگر هوش مصنوعی طراحی کرده است این سیستم فقط از سه ناحیه زیر برای ایندکس‌گذاری و رتبه‌بندی استفاده می‌کند

Title <<< عنوان شغلی فرد

Skills <<< مهارت‌ها

experience <<< توضیحات سوابق کاری

تیم فنی متوجه شده که برخی داوطلبان با تکرار زیاد واژه‌های پرکاربرد مانند

a) هوش مصنوعی
b) یادگیری عمیق
c) پردازش تصویر

موفق می‌شوند رتبه غیرواقعی و بالایی کسب کنند در حالی که بعضی رزومه‌های تخصصی‌تر به دلیل استفاده از واژگان دقیق‌تر و کمتر تکرارشده، رتبه پایین‌تری دریافت می‌کنند و سیستم فعلی دارای ویژگی‌های زیر است:

مدل رتبه‌بندی: Vector Space Model با Cosine Similarity

وزن ناحیه‌ها:

title = 5 , skills = 3 , experience = 1

محاسبه tf-idf روی کل مجموعه اسناد انجام می‌شود و سیستم از tf خام (Raw Term Frequency) استفاده می‌کند.

سه رزومه زیر در سیستم وجود دارند.

R1

Title: Senior AI Researcher

Skills: machine learning, deep learning, computer vision

Experience: deep learning deep learning deep learning image analysis neural network

---

R2

Title: Medical Image Analyst

Skills: CT analysis, lung diagnosis, segmentation

Experience: lung cancer diagnosis using deep neural models for CT scan analysis

---

R3

Title: AI Engineer

Skills: AI, machine learning

Experience: artificial intelligence systems and AI products

کوئری کارفرما:

Q = “deep learning for lung cancer diagnosis”

بخش الف)

بدون انجام محاسبات عددی کامل، رفتار سیستم رتبه‌بندی را تحلیل کنید و به پرسش‌های زیر پاسخ دهید

1. کدام رزومه ممکن است به‌صورت گمراه‌کننده رتبه اول را کسب کند؟؟ دلیل خود را دقیق توضیح دهید.
2. کدام رزومه از نظر معنایی به کوئری نزدیک‌تر است اما ممکن است رتبه پایین‌تری دریافت کند؟ علت را تحلیل کنید.
3. توضیح دهید چگونه ترکیب همزمان عوامل زیر می‌تواند باعث ایجاد خطا در رتبه‌بندی شود:

1) tf خام

2) Cosine Normalization

3) وزن‌دهی ناحیه‌ای

4. آیا ممکن است افزایش وزن ناحیه title باعث کاهش کیفیت رتبه‌بندی شود، حتی اگر عنوان شغلی مهم‌ترین بخش رزومه باشد؟؟؟ پاسخ خود را با یک زنجیره استدلال کامل توضیح دهید.

در پاسخ باید تعامل میان مفاهیم زیر تحلیل شود.

a) طول سند

b) توزیع واژه‌ها

c) idf

d) وزن ناحیه‌ها

e) نرمال‌سازی برداری

بخش ب) تیم طراحی سیستم سه پیشنهاد برای بهبود رتبه‌بندی ارائه کرده است

پیشنهاد ۱: استفاده از تابع log(tf)

پیشنهاد ۲: استفاده از Binary tf در ناحیه skills و tf خام در سایر ناحیه‌ها

پیشنهاد ۳: حذف کامل Cosine Normalization با این استدلال که «رزومه‌های طولانی‌تر اطلاعات بیشتری دارند»

سوالات:

1. هر یک از پیشنهادهای بالا دقیقاً چه مشکلی را در سیستم هدف قرار می‌دهند؟؟

2. هر پیشنهاد ممکن است چه Bias یا خطای جدیدی ایجاد کند؟

3. آیا ممکن است پس از تغییر تابع tf، دو رزومه با شباهت معنایی متفاوت، امتیاز نهایی تقریباً یکسانی کسب کنند؟؟؟ توضیح دهید چگونه چنین اتفاقی ممکن است رخ دهد

4. آیا حذف Cosine Normalization می‌تواند باعث شود وزن ناحیه‌ها عملاً بی‌اثر یا بیش‌ازحد تقویت شوند؟؟ تحلیل کنید.

5. اگر فقط اجازه اعمال یک تغییر در سیستم وجود داشته باشد، کدام گزینه را انتخاب می‌کنید و چرا؟

سوال 3

یک خبرگزاری بین‌المللی سامانه‌ی جستجوی مقالات خود را بر اساس یک ایندکس چندلایه‌ی فهرستی طراحی کرده اما یه مشکل جدی دارد!

بسیاری از اسناد در این مجموعه بازنشر شده یا نزدیک به بازنشر (Near-Duplicate) هستند ، بازنشرها معمولاً تمام واژه‌های مهم پرس‌وجو را دارند، اما ارزش محتوایی یکسانی ندارند!

برخی گزارش‌های اصلی، واژه‌ها را در موقعیت‌های نزدیک و مرتبط آورده‌اند ، در حالی که نسخه‌های بازنشر آن واژه‌ها را در بخش‌های مختلف سند پخش کرده‌اند.

برای افزایش سرعت و کاهش هزینه ها ساختار ایندکس به صورت زیر طراحی شده

1.لایه‌ی اول وجود یا عدم وجود واژه‌ها (Boolean layer)

2. لایه‌ی دوم فراوانی واژه‌ها (Term Frequency layer)

3. لایه‌ی سوم موقعیت واژه‌ها (Positional layer)

4. غربال اولیه بر اساس الگوریتم Fast Cosine Score

5. بازرتبه‌بندی نهایی بر اساس امتیاز نزدیکی واژه‌ها (Proximity)

نکته‌ی چالشی این است که فهرست‌های چندلایه باید همیشه از بالا به پایین پیمایش شوند اما هزینه‌ی دسترسی به لایه‌ی سوم بسیار بالا است.

مدیر سامانه همچنین تأکید دارد که بازرتبه‌بندی proximity نباید فقط روی سندهایی با cosine بالا اعمال شود چرا که ممکن است برخی اسناد با cosine پایین ولی نزدیکی بالای واژه‌ها از نظر معنایی ارزش بیشتری داشته باشند

پرس‌وجوی نمونه:

Q = “trade climate impact"

بخش اول طراحی جریان رتبه‌بندی (Pipeline Design)

با توجه به سناریو و محدودیت‌های موجود، طرح مفهومی خود را برای فرایند رتبه‌بندی بنویسید طراحی شما باید به‌گونه‌ای باشد که:

1. از ایندکس چندلایه استفاده کند بدون نقض محدودیت‌ها
2. مرحله‌ی غربال اولیه با Fast Cosine Score را طوری طراحی کند که سرعت و کارایی سیستم حفظ شود و اسناد بازنشرشده که cosine بالا ولی نزدیکی پایین دارند، رتبه‌ی اشتباهی نگیرند
3. فرایند بازرتبه‌بندی بر اساس Proximity را طوری تنظیم کند که فقط روی اسناد امیدبخش اجرا شود (برای کاهش هزینه)، اما الزام مدیر محصول را نیز رعایت کند تا سندهایی با cosine پایین اما proximity قوی حذف نشوند

4. توضیح دهید در صورت طراحی غلط pipeline، چگونه ممکن است اسناد بازنشرشده امتیاز ناعادلانه بالا بگیرند یا اسناد اصلی حذف زودهنگام شوند

بخش دوم طراحی توابع تحلیل و امتیازدهی (Analysis & Scoring Functions)

اکنون سامانه باید دارای دو تابع اصلی باشد

1. تابع تحلیل (Analysis Function)
2. تابع امتیازدهی نهایی (Final Scoring Function)

اما در این سناریو، تشخیص اسناد بازنشر بی‌ارزش تنها با cosine ممکن نیست

و proximity نیز به تنهایی کافی نیست زیرا سندهای طولانی ممکن است فاصله‌ی واژه‌ها را با نویز زیاد تغییر دهند

با توجه به این شرایط موارد زیر را توضیح دهید

1. تابع تحلیل باید چه اطلاعاتی از پرس‌وجو و سند استخراج کند، به‌طوری‌که هم در محاسبه‌ی Fast Cosine Score و هم در Proximity Scoring و همچنین درتشخیص near-duplicate بودن اسناد مفید باشد
2. تابع امتیازدهی نهایی چگونه باید وزن مؤلفه‌های مختلف را بین این سه عامل تقسیم کند
    1) شباهت برداری (Cosine Similarity)
    2) توزیع مکانی واژه‌ها در سند (Proximity Distribution)
    3) ساختار لایه‌ای ایندکس (Layer Impact)

تا در نهایت، اسناد بازنشرشده امتیاز بالا نگیرند و اسناد اصلی که از نظر معنایی مرتبط‌ترند رتبه‌ی دقیق‌تر بگیرند

در اخر این موارد زیر را توضیح دهید:

1. توضیح دهید اگر نزدیکی واژه‌ها بدون توجه به الگوی توزیع مکانی درون سند لحاظ شود ، چگونه ممکن است سندهای طولانی و بازنشرشده به‌اشتباه امتیاز بالاتری از سند اصلی بگیرند؟؟
2. بیان کنید در طراحی عملی، چگونه باید از لایه‌ی موقعیتی (Positional Layer) فقط در بالاترین اسناد امیدبخش استفاده کرد تا هزینه‌ی سیستم قابل قبول باقی بماند؟

سوال 4

استارتاپ «DocFind» در حال توسعه یک موتور جستجوی تخصصی برای مقالات علمی فارسی در حوزه سلامت است هدف این موتور جستجو کمک به پژوهشگران و دانشجویان برای یافتن سریع و دقیق مقالات مرتبط با پرس‌وجوهای علمی است برای ارزیابی عملکرد سیستم تیم تحقیقاتی دو نسخه متفاوت از موتور جستجو را پیاده‌سازی کرده است و آن‌ها را برای یک پرس‌وجوی مشخص با یکدیگر مقایسه می‌کند.

پرس‌وجوی مورد بررسی به صورت زیر است:

«تأثیر مصرف قند بر دیابت نوع دو»

در مجموعه اسناد این سیستم بر اساس ارزیابی مرجع در مجموع ۲۰ سند واقعاً مرتبط با این پرس‌وجو وجود دارد ، برای ساخت مجموعه ارزیابی از روش Pooling استفاده شده در این روش داوران انسانی فقط اسنادی را بررسی کرده‌اند که در میان ۱۵ نتیجه اول هر سیستم ظاهر شده‌اند بنابراین توجه داشته باشید که ممکن است در مجموعه اسناد مدارک مرتبطی وجود داشته باشند که هنوز توسط داوران بررسی نشده‌اند

دو نسخه سیستم به صورت زیر هستند

سیستم A (بازیابی بدون رتبه)

این سیستم از یک مدل Boolean Retrieval استفاده می‌کند و اسناد را به صورت یک مجموعه بدون رتبه‌بندی برمی‌گرداند ، نتایج مشاهده‌شده برای این سیستم به شرح زیر است

1) تعداد کل اسناد بازیابی‌شده: ۳۰ سند
2) تعداد اسناد مرتبط در میان آن‌ها: ۱۵ سند
3) در میان ۱۰ نتیجه اولی که به کاربر نمایش داده می‌شود (با ترتیب تصادفی)، ۵ سند مرتبط هستند

سیستم B (بازیابی رتبه‌بندی‌شده با نمایش Snippet)

این سیستم اسناد را بر اساس میزان شباهت به پرس‌وجو رتبه‌بندی می‌کند و برای هر نتیجه یک snippet از متن مقاله نمایش می‌دهد تا کاربر بتواند سریع‌تر درباره مفید بودن نتیجه تصمیم بگیرد ، ۱۰ نتیجه اول این سیستم و قضاوت داوران درباره مرتبط بودن آن‌ها به صورت زیر است

رتبه نتایج >>> 1 2 3 4 5 6 7 8 9 10

مرتبط بودن >>> 0 1 1 0 1 0 0 1 0 1 <<< عدد ۱ نشان‌دهنده «مرتبط» و عدد ۰ نشان‌دهنده «نامرتبط»

همچنین گزارش شده است که سیستم B در مجموع توانسته ۱۸ سند مرتبط را در میان تمام نتایج خود بازیابی کند

در آزمایش‌های تجربه کاربری مشاهده شده است که در حدود ۷۰ درصد موارد کاربران روی نتیجه رتبه ۱ در سیستم B کلیک می‌کنند در حالی که این سند از نظر داوران انسانی نامرتبط ارزیابی شده است

بررسی بیشتر نشان داده است که snippet این نتیجه شامل عبارت «قند و دیابت نوع دو» به صورت پشت سر هم است، اما در متن کامل مقاله بحث درباره «مدل‌های حیوانی غیرانسانی» است که طبق دستورالعمل ارزیابی، مرتبط محسوب نشده است

برای ساخت مجموعه قضاوت‌ها دو داور مستقل برخی از اسناد pool شده را بررسی کرده‌اند نتایج توافق آن‌ها بر روی ۶۰ سند به صورت زیر گزارش شده است

1) هر دو داور سند را مرتبط دانسته‌اند ۲۲ مورد

2) هر دو داور سند را نامرتبط دانسته‌اند ۲۴ مورد

3) داور اول مرتبط و داور دوم نامرتبط ۸ مورد

4) داور اول نامرتبط و داور دوم مرتبط ۶ مورد

بخش الف

الف-۱) عملکرد سیستم A را با استفاده از معیارهای زیر محاسبه کنید

Precision

Recall

F1-score

Accuracy

در محاسبات خود فرض کنید ارزیابی فقط بر اساس اسناد قضاوت‌شده انجام شده است

الف-۲) عملکرد سیستم B را با استفاده از معیارهای زیر ارزیابی کنید

Precision@5

R-Precision

Mean Average Precision (MAP)

NDCG@10

تمام مراحل محاسبه باید به طور واضح نوشته شوند

الف-۳) بدون انجام محاسبه عددی جدید، توضیح دهید آیا ممکن است سیستمی مانند سیستم B در معیارهایی مانند MAP و NDCG عملکرد بهتری نسبت به سیستم A داشته باشد، اما در آزمایش‌های واقعی کاربران رضایت کمتری ایجاد کند؟؟

در پاسخ خود حتما باید به طور مشخص به موارد زیر اشاره کنید

1) نقش ترتیب نتایج در معیارهای رتبه‌بندی
2) تفاوت بین relevance تعریف‌شده در ارزیابی آفلاین و perceived relevance از دید کاربر
3) نقش snippet در ایجاد bias در رفتار کلیک کاربران

بخش ب

ب-۱) با استفاده از داده‌های ارائه‌شده، ضریب توافق Cohen’s Kappa را برای دو داور محاسبه کنید ، سپس توضیح دهید که آیا این میزان توافق برای اعتماد به مجموعه قضاوت‌های انسانی کافی است یا خیر؟؟

همچنین تحلیل کنید که اگر توافق داوران متوسط یا پایین باشد، کدام معیارهای ارزیابی (Precision، MAP، NDCG یا R-Precision) بیشتر تحت تأثیر قرار می‌گیرند و دلیل آن چیست؟؟؟

ب-۲) تیم فنی برای افزایش مقیاس‌پذیری سیستم تصمیم گرفته است تغییرات زیر را اعمال کند

1) ساده‌تر کردن ساختار ایندکس و حذف اطلاعات مربوط به موقعیت کلمات (در نتیجه امکان انجام phrase query حذف می‌شود)
2) محدود کردن زبان پرس‌وجو به عملگرهای ساده AND و OR
3) تولید snippet تنها بر اساس تطابق سطحی کلمات با پرس‌وجو و نه بر اساس بخش معنایی متن

با توجه به سناریوی مسئله، تحلیل کنید که

1) کدام معیارهای ارزیابی ممکن است به طور ظاهری بهبود پیدا کنند بدون آنکه کیفیت واقعی سیستم بهتر شده باشد
2) کدام معیارها ممکن است افت کنند حتی اگر تجربه واقعی کاربر بهتر شود
3) آیا ممکن است کاهش کیفیت قضاوت انسانی (مثلاً کاهش مقدار Kappa) باعث شود تیم مهندسی درباره کیفیت سیستم به نتیجه‌گیری نادرست برسد

در پاسخ خود باید ارتباط بین ارزیابی آفلاین، رفتار واقعی کاربران، طراحی سیستم و نقش snippet در تجربه کاربر را به طور تحلیلی توضیح دهید.

موفق و موید :(