به نام خدا

دانشگاه تهران
دانشکدگان فنی
دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین ۱

استاد درس: خانم دکتر آزاده شاکری

سرپرست دستیاران آموزشی: سمانه پیمانی‌راد

طراح تمرین: سینا کارگران

آبان ماه ۱۴۰۴

فهرست

مقدمه ۳

سوالات عملی (۶۰ نمره) ۴

مجموعه داده ۵

پیش‌نیازها – ایجاد شاخص (۵ نمره) ۷

سوال ۱: تابع بازیابی BM25 (۵۰ نمره) ۹

سوال ۲: تابع بازیابی Pivoted Length Normalization (۱۵ نمره) ۱۳

سوالات تئوری (۶۰ نمره) ۱۵

سوال ۱: آشنایی با وزن‌ها (۱۰ نمره) ۱۶

سوال ۲: تخمین اطلاعات (۱۵ نمره) ۱۷

سوال ۳: تاثیر پارامترها (۱۰ نمره) ۱۸

سوال ۴: مشکلات bm25 (۱۵ نمره) ۱۸

سوال ۵: رتبه‌بندی اسناد (۱۰ نمره) ۲۰

ملاحظات (حتما مطالعه شود) ۲۱

استفاده مسئولانه از هوش مصنوعی ۲۳

۱. هدف و اصول کلی ۲۳

۲. استفاده مجاز از LLMها ۲۳

۳. استفاده غیرمجاز از LLMها ۲۴

۴. مستندسازی ۲۴

۵. آمادگی ارائه شفاهی ۲۴

۶. پیامدهای تخلفات ۲۴

۷. موارد تکمیلی ۲۴

۸. اظهارنامه ۲۵

مقدمه

اهداف اصلی تمرین:

- پیاده‌سازی برخی از توابع بازیابی اطلاعات
- به کارگیری و انجام آزمایش‌ها بر روی مجموعه دادگان و بررسی کارکرد هرکدام از توابع بازیابی
- مقایسه عملکرد توابع مختلف و تفسیر نتایج به‌دست‌آمده

نکات قابل توجه در هنگام پاسخ به سؤالات:

- تمامی کدهای نوشته شده باید قابلیت اجرای مجدد داشته باشند. به تفسیرهایی که بدون آزمایش و صرفاً به‌صورت فرضی بیان‌گردند نمره‌ای تعلق نمی‌گیرد.
- نمره اصلی تمرین مربوط به کیفیت و درستی تفسیرهای ارائه شده‌است.
- بهتر است از نمودارها و کشف نمونه‌های مرتبط از اسناد و پرس‌وجوها برای افزایش کیفیت تفسیرها استفاده گردد.
- بدیهی است که حجم تمرین معیار نمره‌ی شما نیست، بلکه صحت انجام آزمایش‌ها و کیفیت تفسیرهای شما مهم است.
- توصیه می‌شود انجام تمرین را به روزهای آخر موکول نکنید! **اجرای آزمایش‌ها و پیدا کردن پارامتر بهینه زمان‌بر بوده** و تفسیر هرکدام نیازمند تحلیل و بررسی است.

سوالات عملی (۶۰ نمره)

در این تمرین، با مدل‌های مبتنی بر شباهت برداری آشنا خواهید شد و به انجام آزمایش‌های مختلف برای درک بهتر کارکرد آن‌ها خواهید پرداخت. هدف از این تمرین آشنایی با ابزارهای جستجوی متنی و همچنین آشنایی با معیارهای ارزیابی و توابع امتیازدهی به اسناد است. یک تابع امتیازدهی با توجه به میزان ارتباط یک سند با پرس‌وجو، امتیازی به سند تخصیص می‌دهد تا در نهایت اسناد براساس امتیازشان، رتبه‌بندی و نمایش داده‌شوند. رتبه‌بندی حاصل عموماً با رتبه‌بندی طلایی[۱] مقایسه شده و کارایی تابع بازیابی گزارش می‌گردد.

پاسخ سوالات این بخش را در همان نوت‌بوک تمرین خود *(IIR-CA1-Code.ipynb)* بنویسید.

باتوجه به حجم مجموعه داده اجرای این کد به طور تقریبی نیازمند ۵ گیگابایت رم در دسترس است. در صورتی که امکان اجرای کد روی سیستم خود را ندارید می‌توانید از Google Colab استفاده کنید (نیازی به اجرا بر روی GPU نیست و در حالت CPU می‌توانید کدهای خود را اجرا کنید.)

[۱] Golden Rankings

مجموعه داده

برای این تمرین، مجموعه داده‌ی **ANTIQUE**[۲] شامل ۲۶۲۶ سؤال به زبان طبیعی در حوزه‌ی باز است که از دسته‌بندی‌های متنوعی گردآوری شده‌اند. این سؤالات توسط کاربران واقعی در سرویس پرسش‌وپاسخ Yahoo! Answers مطرح شده‌اند. برای هر سؤال، تمام پاسخ‌های موجود از طریق crowdsourcing (جمع‌سپاری) مورد ارزیابی قرار گرفته و در مجموع ۳۴٬۰۱۱ برچسب ارتباطی به‌صورت دستی ثبت شده است.

پیکره متنی[۳] (اسناد):

این مجموعه شامل ۴۰۳۶۶۶ سند است که هر سند دارای فیلدهای زیر می‌باشد:

- doc_id شناسه هر سند.
- text متن پاسخ کوتاه سند از Yahoo! Answers.

پرس‌وجوها[۴]:

این فایل شامل مجموعه‌ای از ۲۰۰ پرس‌وجو به زبان طبیعی یا همان کوئری‌ها است.

دادگان طلایی[۵]:

این فایل شامل ۶۵۸۹ قضاوت‌های مرتبط[۶] همراه با برچسب میزان ارتباط می‌باشد در مرحله نهایی جهت ارزیابی کارایی توابع بازیابی مورد استفاده قرار می‌گیرد. بر حسب میزان ارتباط پرس‌وجو با سند، عدد برچسب

---

[۲] ANTIQUE: A Non-factoid Question Answering Benchmark
[۳] Corpus
[۴] Queries
[۵] Golden Dataset
[۶] Relevance Judgments

موجود می‌تواند از ۱ تا ۴ متغیر باشد. **برای محاسبه‌ی معیارهای ارزیابی، تنها برچسب‌های ۳ و ۴ به عنوان سند مرتبط در نظر گرفته می‌شوند.**

تفسیر برچسب‌ها:

۱. کاملاً خارج از موضوع است یا هیچ معنایی ندارد. — ۱.۶ هزار (٪۲۴.۹)

۲. به سؤال پاسخ نمی‌دهد، یا اگر پاسخ می‌دهد، پاسخی غیرمنطقی است. با این حال، از موضوع خارج نیست. بنابراین نمی‌توان آن را به‌عنوان پاسخ به سؤال پذیرفت. — ۲.۴ هزار (٪۳۶.۷)

۳. می‌تواند پاسخی به سؤال باشد، اما به‌اندازه‌ی کافی قانع‌کننده نیست. باید پاسخی با کیفیت بسیار بهتر برای این سؤال وجود داشته باشد. — ۱.۲ هزار (٪۱۸.۲)

۴. منطقی و قانع‌کننده به نظر می‌رسد. کیفیت آن هم‌سطح یا بهتر از «پاسخ احتمالاً درست یا برچسب ۳» است. توجه داشته باشید که لازم نیست دقیقاً همان پاسخ را ارائه دهد که در «پاسخ احتمالاً درست» آمده است.

پیش‌نیازها – ایجاد شاخص[۷] (۵ نمره)

همان‌طور که در مباحث درس بازیابی هوشمند اطلاعات اشاره شد، جهت استفاده از اسناد در توابع بازیابی، ابتدا باید اسناد پردازش و شاخص‌گذاری شوند. شاخص‌گذاری به منظور دسترسی سریع‌تر به اطلاعات و آمارهای مورد نیاز برای محاسبه‌ی مقادیر امتیازها صورت می‌گیرد.

برای شاخص گذاری و پردازش اسناد، نیازی به نوشتن کد جدید نیست. شما صرفاً باید کدهای موجود در فایل اصلی (فایل نوت‌بوک تمرین) را اجرا کنید. در فایل نوت‌بوک ارائه شده همراه تمرین، اجرای کدهای مربوط به شاخص‌گذاری و پردازش اسناد در ابتدای فایل قابل مشاهده است. این کدها شامل مراحل پیش‌پردازش و شاخص‌گذاری می‌باشند و نیازی به تغییرات در آن‌ها نیست.

موتور جستجو و شاخص‌گذاری به‌کار رفته در این تمرین شامل چند فایل جداگانه است که در فایل نوت‌بوک وارد (import) شده و استفاده می‌شوند:

- document_processing.py
- evaluation.py
- inverted_index.py
- retrieval_models.py

برای انجام سوال اول این تمرین نیازی به تغییر کدهای هیچ‌کدام از این فایل‌ها، به‌جز فایل retrieval_models.py نیست. در سوال ۱ و ۲ درباره‌ی اینکه چه کدهایی باید به این بخش اضافه کنید توضیح داده شده‌است.

با این حال، برای آشنایی بیشتر با موتور جستجو و فایل‌های مربوطه، توصیه می‌شود که کدها را مرور کرده و کارکرد کلی هر کدام را به‌طور خلاصه شرح دهید. (برای فهم کدهای نوشته شده می‌توانید یادداشت نوشته شده در کدها را مطالعه کنید. همچنین توصیه می‌شود از ابزارهایی مانند ChatGPT استفاده کنید.)

[۷] indexing

- پیش‌پردازش انجام شده روی اسناد و عبارات جستجو چیست و چرا انجام می‌شود؟ نقش هر پیش‌پردازش انجام شده در نتیجه نهایی چه می‌تواند باشد؟
- شاخص‌گذاری اسناد به چه شکلی انجام می‌شود؟
- به طور معمول برای شاخص‌گذاری اسناد از ساختارداده‌های مشخصی مانند:
  - Lexicon
  - PostingIndex یا Inverted index
  - DocumentIndex
  - MetaIndex

استفاده می‌شود، در مورد هرکدام از این موارد یک تحقیق و توضیح مختصر ارائه دهید.

نیازی به شرح فنی و دقیق کدها نیست و هدف از این مرحله، پیدا کردن یک دید کلی از نحوه کارکرد فایل‌های داده شده است. پس از این مرحله، به سراغ بخش اصلی تمرین یعنی پیاده‌سازی و اجرای آزمایش با استفاده از متدهای بازیابی می‌رویم.

سوال ۱: تابع بازیابی BM25 (۵۰ نمره)

هدف از این سؤال، آشنایی با مولفه‌های روش BM25 و تأثیر هر یک از پارامترهای آن بر روی کیفیت رتبه‌بندی است. در بخش الف، شما باید بازیابی اطلاعات را به روش BM25 برای پرس‌وجوها انجام دهید و تأثیر پارامترهای این روش، یعنی b و k را بررسی کنید. برای این کار، باید مقادیر مختلف را برای این پارامترها آزمایش کنید تا به مقدار بهینه برسید. هنگام بررسی مقادیر بهینه، به تأثیر هر یک از مولفه‌های تابع امتیازدهی دقت کنید. سپس در ادامه روش‌های پیشنهادی دیگر را مورد بررسی و آزمایش قرارخواهیم داد.

راهنمایی:

- کد مربوط به روش BM25 در فایل retrieval_models.py پیاده‌سازی شده است و همچنین یک اجرای نمونه از آن در ابتدای فایل نوت‌بوک داده شده است. شما می‌توانید از این نمونه برای فهمیدن چگونگی انجام جستجوها و دریافت نتیجه ارزیابی‌ها استفاده کنید.
- یکی از اهداف اصلی این تمرین توانایی اجرای آزمایش‌ها به نحوی صحیح و مهندسی شده برای یافتن مقدار پارامترهای مناسب برای هر روش است. ابتدا پارامترها را با گام‌های بلند و سپس با گام‌های کوچک آزمایش‌های خود را تکمیل کنید تا منابع محاسباتی تلف نشود.
- برای نمایش تأثیر هر یک از پارامترها، نمودارهایی مناسب رسم کرده و تفسیر خود را بر اساس آن‌ها بنویسید.
- برای پیاده‌سازی روش‌های پیشنهادی، می‌توانید هر روش را به‌عنوان یک تابع جدید در ادامه فایل retrieval_models.py و با ساختاری مشابه تابع BM25 پیاده‌سازی کنید. همچنین با مطالعه‌ی تابع BM25، متوجه بخش‌های مختلف مورد نیاز برای پیاده‌سازی دیگر روش‌ها خواهید شد. دقت داشته باشید که نیازی به تغییر تنظیمات import در ابتدای نوت‌بوک نیست. پس از تغییر و ذخیره فایل، می‌توانید از آن بدون import مجدد استفاده کنید.

- معیارهای ارزیابی MAP, MRR, P@3, P@10, R@3, R@10 و nDCG@10 می‌باشند. برای سادگی بهینه‌سازی و پیدا کردن بهترین پارامتر تنها مقدار MAP را در نظر بگیرید.
- **برای سادگی بیشتر بازه‌ی جستجوی پارامترها را نیز می‌توانید محدود به بازه‌ی زیر در نظر بگیرید:**
  - **K: [0, 3]**
  - **b: [0, 1]**
  - **δ: [0, 2]**

الف) در مرحله اول شما باید بازیابی را به روش BM25 برای پرس‌وجوها انجام دهید و تأثیر پارامترهای این روش (b, k) را بررسی کنید. شما باید مقادیر مختلف را برای پارامترها آزمایش کنید تا به مقدار بهینه برای این دو مقدار برسید. هنگام تفسیر مقادیر بهینه، به تأثیر هر یک از مولفه‌های تابع امتیازدهنده دقت کنید.

$$f(q,d) = \sum_{w \in q \cap d} IDF(w) \frac{c(w,d)(k+1)}{c(w,d) + k(1 - b + b\frac{|d|}{avdl})}$$

ب) در این قسمت ابتدا باید هرکدام از روش‌های پیشنهادی را پیاده‌سازی کنید و سپس به مانند بخش الف بازیابی را انجام و مورد آزمایش قرار دهید. در اینجا ۶ روش پیشنهادی آورده شده است، روش پیشنهادی هفتم را نیز خودتان باید پیشنهاد دهید! در نهایت با توجه به معیارهای ارزیابی، تمامی توابع را با یکدیگر مقایسه کنید و نتایج به‌دست‌آمده را تفسیر کنید.

۱) روش پیشنهادی اول

$$f(q,d) = \sum_{w \in q \cap d} IDF(w)$$

۲) روش پیشنهادی دوم

$$f(q,d) = \sum_{w \in q \cap d} \frac{c(w,d)(k+1)}{c(w,d)+k}$$

۳) روش پیشنهادی سوم (BM11)

$$f(q,d) = \sum_{w \in q \cap d} IDF(w)\left(\frac{c(w,d)(k+1)}{k\left(\frac{|d|}{avdl}\right)+c(w,d)}\right)$$

۴) روش پیشنهادی چهارم

$$f(q,d) = \sum_{w \in q \cap d} I(w,d)$$

$$I(w,d) = 1\ (if\ count(w) \neq 0\ ), 0(o.w.)$$

۵) روش پیشنهادی پنجم (وزن بیشتر عناصر کمیاب)

$$f(q,d) = \sum_{w \in q \cap d} IDF(w)^2 \frac{c(w,d)}{c(w,d)+k(1-b+b\frac{|d|}{avdl})}$$

۶) روش پیشنهادی ششم (BM25+[۸])

(مقادیر مختلف برای δ بررسی شود و بهترین مقدار گزارش شود)

$$f(q,d) = \sum_{w \in q \cap d} IDF(w)\left(\frac{c(w,d)(k+1)}{c(w,d)+k(1-b+b\frac{|d|}{avdl})} + \delta\right)$$

[۸] Lv, Y., C. Zhai, Lower-bounding term frequency normalization, CIKM 2011, p. 7-16

۷) روش پیشنهادی هفتم (روش پیشنهادی شما!) (سوال امتیازی)

در این قسمت یک روش پیشنهادی متفاوت از روش‌های بررسی شده در این تمرین ارائه دهید. این روش می‌تواند حاصل جستجو و تحقیق شما باشد یا می‌تواند حاصل خلاقیت خودتان باشد.

سوال ۲: تابع بازیابی PIVOTED LENGTH NORMALIZATION (۱۵ نمره)

هدف از این سؤال آشنایی با تأثیر تابع تبدیل استفاده‌شده برای مولفه TF در کیفیت رتبه‌بندی می‌باشد. این روش برای اولین بار در مقاله‌ای[۹] با عنوان Pivoted Document Length Normalization معرفی گردید.

الف) در این قسمت مانند بخش قبل ابتدا توابع را در فایل retrieval_models.py پیاده‌سازی کنید و سپس مورد استفاده و آزمایش قرار دهید. اجرای گام‌های بلند و کوچک آزمایش به طور همزمان برای هردو روش انجام شود (در هر قدم مقدار پارامترها برای هر دو مدل باید یکسان باشد تا بتوان مورد مقایسه قرار داد.) در نهایت با توجه به نتایج معیارهای ارزیابی، توابع را با یکدیگر مقایسه کنید و تفسیر خود را بیان نمایید.

۱) مدل اصلی:

$$f(q,d) = \sum_{w \in q \cap d} c(w,q) \frac{\ln\left(1+\ln\left(1+c(w,d)\right)\right)}{1-b+b\frac{|d|}{avdl}} \, log \frac{M+1}{df(w)}$$

۲) مدل بدون مولفه تودرتو

$$f(q,d) = \sum_{w \in q \cap d} c(w,q) \frac{\ln\left(1+c(w,d)\right)}{1-b+b\frac{|d|}{avdl}} \, log \frac{M+1}{df(w)}$$

ب) در درس با مفهوم آزمون‌های استنباط آماری برای مقایسه رتبه‌بندهای متفاوت آشنا شدید. حال نیاز است تا تمام توابع پیاده‌سازی شده‌ی خود در سوال ۱ و۲ را از با استفاده از معیار AP به ازای هر کوئری مورد مقایسه قرار دهید. روش BM25 را به عنوان روش پایه یا baseline در نظر بگیرید و با استفاده از t-test و مقایسه AP هر کوئری، تعداد کوئری‌ها با عملکرد بهتر نسبت به baseline و تعداد کوئری‌ها با عملکرد بدتر نسبت به baseline را گزارش کنید. مقدار p-value را نیز گزارش کنید. مقادیر p-value چه کمکی در تفسیر نتایج می‌کند؟

[۹] http://singhal.info/pivoted-dln.pdf

(برای تست هر تابع می‌توانید از پارامترهای بهینه‌شده هر روش استفاده کنید)

در نهایت خروجی این سوال باید جدولی مشابه به شکل زیر باشد(الزامی به یکی بودن اعداد و نتایج وجود ندارد و برای پیاده‌سازی می‌توانید از chatgpt استفاده کنید و مهم توضیح و تفسیر و درک شما است؛ در صورت امکان پیاده‌سازی خود را به صورت یک تابع جدید در ماژول evaluation انجام دهید و خروجی را در فایل پاسخ خود نمایش دهید.):

	Method	MAP	MRR	Precision@3	Precision@10	Recall@3	Recall@10	nDCG@10	Better	Worse	p-val (MAP)
0	bm25_score	0.2427	0.6298	0.4167	0.2905	0.1095	0.2360	0.4388	-	-	-
1	method1_score	0.1499	0.4257	0.2517	0.1725	0.0745	0.1550	0.2528	21	170	0.0
2	method2_score	0.1259	0.4025	0.2383	0.1645	0.0621	0.1332	0.2240	19	172	0.0
3	method3_score	0.2109	0.5287	0.3350	0.2510	0.0955	0.2160	0.4009	71	120	0.0
4	method4_score	0.1053	0.3706	0.1900	0.1190	0.0556	0.1065	0.1820	14	177	0.0
5	method5_score	0.2385	0.6103	0.4100	0.2915	0.1080	0.2389	0.4384	90	98	0.0056
6	method6_score	0.0505	0.1834	0.0867	0.0585	0.0229	0.0494	0.1134	10	181	0.0
7	method7_score	0.2344	0.5962	0.3950	0.2850	0.1070	0.2334	0.4301	80	101	0.0448
8	method8_score	0.2343	0.6050	0.4050	0.2855	0.1062	0.2308	0.4316	84	104	0.0002
9	method9_score	0.2408	0.6171	0.4117	0.2960	0.1089	0.2418	0.4429	105	83	0.2297
10	pivoted_length_v1_score	0.0758	0.2683	0.1350	0.1070	0.0317	0.0849	0.1647	10	177	0.0
11	pivoted_length_v2_score	0.0643	0.2250	0.1150	0.0910	0.0286	0.0721	0.1380	9	178	0.0

شکل ۱ نمونه خروجی مورد نظر برای بخش دوم سوال دوم

سوالات تئوری (۶۰ نمره)

باتوجه به مباحث تدریس شـده سـوالات زیر را حل کنید. پاسـخ سـوالات این بخش باید در قالب مربوط به سوالات تئوری (*IIR-CA1-Theory*) بنویسید.

نکات سوالات تئوری:

- صرفا به جواب نهایی نمره تعلق نمی‌گیرد و راه‌حل نوشته شده بخش اصلی نمره‌ی شماست.
- می‌توانید جهت اطمینان از محاسـبات خود از کدنویسـی اسـتفاده کنید، اما باید راه‌حل به طور کلی به صـورت ریاضـی در گزارش تمرین آمده باشـد و صـرفا قراردادن کد راه‌حل باعث کسـر نمره شـما می‌شود.

سوال ۱: آشنایی با وزن‌ها (۱۰ نمره)

فرمول‌های زیر را برای تخمین امتیاز تشابه سند و پرس‌وجو در نظر بگیرید:

a. **PIV**:

$$\sum_{w \in q \cap d} \left(1 + \ln\left(1 + \ln\left(c(w,d)\right)\right)\right) \cdot \frac{N+1}{df(w) \cdot d} \cdot \frac{c(w,q) \cdot d}{(1-s) + s \cdot \frac{|d|}{avdl}}$$

b. **BM25**:

$$\sum_{w \in q \cap d} \ln\left(\frac{N - df(w) + 0.5}{df(w) + 0.5}\right) \cdot \frac{(k_1+1) \cdot c(w,d)}{k_1\left((1-b) + b \cdot \frac{|d|}{avdl}\right) + c(w,d)} \cdot \frac{(k_3+1) \cdot c(w,q)}{k_3 + c(w,q)}$$

c. **PL2**:

$$\sum_{c(w,q) \cdot \ln(1+\epsilon) \in q \cap d} \left(\frac{tfn_w^d}{tfn_w^d + 1} \cdot \log_2(\mathrm{tfn}_w^d \cdot \lambda_w^d) + \log_2 e \cdot \left(\frac{1}{\lambda_w} - tfn_w^d\right) + 0.5 \cdot \log_2(2\pi \cdot \mathrm{tfn}_w^d)\right)$$

Where:

$$\mathrm{tfn}_w^d = c(w,d) \cdot \log_2\left(1 + \boldsymbol{c} \cdot \frac{avdl}{|d|}\right)$$

$$\lambda_w^d = \frac{N}{c(w,C)}$$

هرکدام از فرمول‌های a تا c از سه بخش TF weighting، IDF weighting و Length Normalization تشکیل شده است. برای مثال در PIV سه‌بخش تشکیل دهنده را به صورت زیر می‌توان تفکیک کرد:

$$\boldsymbol{TF\ weighting} = \left(1 + \ln\left(1 + \ln\left(c(w,d)\right)\right)\right)$$

$$\boldsymbol{IDF\ weighting} = \frac{N+1}{df(w) \cdot d}$$

$$\boldsymbol{Length\ Normalization} = \frac{c(w,q) \cdot d}{(1-s) + s \cdot \frac{|d|}{avdl}}$$

۱. **برای فرمول‌های b و c نیز، سه بخش موثر را مشخص کنید.**

۲. **پارامتر قابل تنظیم c موجود در $PL2$ تاثیر بر روی کدام بخش می‌گذارد و تاثیر افزایش و کاهش آن را تفسیر کنید.**

سوال ۲: تخمین اطلاعات (۱۵ نمره) (سوال امتیازی)

مدل رابرتسون-اسپارک جونز (RSJ) یک مدل احتمالی برای رتبه‌بندی اسناد است.

$$\text{Rank} \propto \sum_{i=1, d_i=1}^{k} \log\left(\frac{p_i(1-q_i)}{q_i(1-p_i)}\right)$$

Two parameters for each term Ai:

$pi = P(A_i = 1|Q, R = 1)$: prob. that term Ai occurs in a relevant doc

$qi = P(A_i = 1|Q, R = 0)$: prob. that term Ai occurs in a non-relevant doc

۱. برای محاسبه‌ی p_i و q_i نیاز به چه اطلاعاتی داریم؟ برای تخمین پارامترهای p_i و q_i در مدل RSJ، در صورتی که relevance judgments در دسترس نباشد، چه فرضیات ساده‌سازانه‌ای اتخاذ می‌شود؟ p_i را چه در نظر می‌گیرید؟ q_i را چه در نظر می‌گیرید؟(جهت راهنمایی به اسلایدهای درس مراجعه کنید)

۲. مقادیر p_i و q_i را با تخمین‌های خود جایگزین کنید و فرمول ذکر شده در بالا را بر اساس n_i که تعداد اسناد نامرتبط دارای واژه i است و N که تعداد کل اسناد نامرتبط است باز نویسی کنید. (مجددا برای راهنمایی به اسلایدهای درس مراجعه کنید)

۳. این رابطه‌ی ساده شده را در نظر بگیرید، فرض کنید کنید N ثابت باشد با افزایش n_i (تعداد اسناد حاوی واژه A_i)، میزان ارتباط (وزن) واژه A_i برای پرس‌وجو بیشتر می‌شود یا کمتر؟ این رفتار مشابه کدام قسمت از فرمول BM25 است؟

سوال ۳: تاثیر پارامترها (۱۰ نمره)

در درس مشاهده کردید که سه پارمتر k1، b و k3 در BM25 موجود است.
در این سوال از شما خواسته می‌شود که تاثیر هرکدام از این پارامتر ها را بر روی امتیاز نهایی تشابه بررسی کنید. برای مشاهده‌ی چشمی تاثیر تغییرات هر پارامتر برای روی امتیاز نهایی می‌توانید از این لینک استفاده کنید؛ اما در نهایت نتیجه‌گیری و جواب نهایی باید بر اساس فرمول BM25 باشد. جواب یک خط یا دو خطی برای هر کدام از پرسش‌ها کافی است.

۱. در صورتی که بخواهیم تغییرات avgdl و |d| هیچ تاثیری بر روی امتیاز نهایی نداشته باشد برای چه پارامتری، چه مقداری باید تنظیم کرد؟
۲. اگر avgdl و |d| با هم برابر باشند آیا تاثیر پارامتر b همچنان پابرجاست؟ در این حالت فرمول BM25 به چه صورتی ساده می‌شود؟
۳. در نمودار **b Parameter Effect** موجود در لینک ضمیمه‌شده، در صورتی که مقدار avgdl بزرگتر از |d| باشد، شیب خط مثبت است یا منفی؟ اگر avgdl کوچکتر از |d| باشد چطور شیب خط مثبت است یا منفی؟
۴. مجددا در نمودار **b Parameter Effect** اگر پارامتر k۱ را از ۰.۱ تا ۳ افزایش یا کاهش دهید تاثیر تغییرات b بر روی امتیاز نهایی شدیدتر می‌شود یا جزئی‌تر؟ با چه مقداری از k۱ تاثیر تغییر b شدیدتر می‌شود؟

سوال ۴: مشکلات BM25 (۱۵ نمره)

در داخل این لینک صفحه‌ی ساده‌ای برای مصورسازی محاسبه امتیاز BM25 یک پرس‌وجو با تنها دو سند آماده شده است:
با استفاده از selector بالای صفحه می‌توانید میان ۸ نمونه سند و پرس‌وجوی از پیش تعریف شده انتخاب داشته باشید. در صورت نیاز خودتان نیز می‌توانید متن اسناد و پرس‌وجو را نیز تغییر دهید و رتبه‌بندی را مشاهده‌کنید. سپس در نهایت یک تا دو خط به هرکدام از پرسش‌های زیر پاسخ کوتاه دهید.

۱. "sample 1" را انتخاب کنید. با توجه به پرس‌وجو، آیا رتبه‌بندی دو سند به درستی انجام شده؟ متوسط طول اسناد و طول هر سند چقدر است؟ آیا می‌توانید با تغییر پارامترها ترتیب رتبه‌بندی اسناد را عوض کنید؟ پارامتر تاثیرگذار و مقدار آن را گزارش کنید.

۲. ”sample 2“ را انتخاب کنید مجددا آیا ترتیب رتبه‌بندی اسناد درست است؟ این بار نیز می‌توانید با تغییر پارامتری رتبه‌بندی را بهم بزنید؟ با تغییر چه پارامتری و به چه دلیل رتبه‌بندی تغییر میکند؟

۳. حال هر کدام از نمونه‌های ”sample 3“ تا ”sample 6“ را انتخاب کنید؛ هرکدام از این جفت پرس‌وجو و اسناد یک ناتوانی اساسی در رتبه‌بندی BM25 را نشان می‌دهد، برای هرکدام از نمونه‌ها بیان کنید که مشکل چیست و چرا BM25 نتوانسته رتبه‌بندی مناسبی داشته باشد؟ (برای هر کدام در یک خط مشکل را بیان کنید)

۴. حال ”sample 7“ را انتخاب کنید؛ همان‌طور که مشاهده می‌کنید این نمونه شامل دو متن جدا از هم title و Document برای هر سند است. در صورتی که اسناد ما چندین قسمت متفاوت داشته باشد، چه ایده‌ای برای تجمع امتیاز BM25 دارید؟ آیا در صورتی که امتیاز BM25 هر کدام از قسمت‌های title و document به صورت جدا حساب شود و تجمیع نهایی امتیاز با جمع امتیازها باشد، نتیجه‌ی نهایی رتبه‌بندی صحیح است؟ برای تجمیع امتیاز چه ایده‌ی دیگری می‌توان داد؟؟ (در صورت علاقه می‌توانید در مورد BM25F جستجو کنید)

۵. حال نمونه نهایی و نمونه ۸ را انتخاب کنید. آیا رتبه‌بندی انجام شده به نظر شما درست است؟ به نظر شما در کوئری داده‌شده اهمیت توکن “پرواز” بیشتر است یا اهمیت توکن “مالزی”؟ آیا الزاما وجود IDF می‌تواند تعیین کننده‌ی میزان اهمیت هر توکن در پرس‌وجو باشد؟

سوال ۵: رتبه‌بندی اسناد (۱۰ نمره)

برای مصورسازی معیارهای ارزیابی رتبه‌بندی اسناد و مقایسه دو مدل رتبه‌بند متفاوت، یک برنامه ساده در این لینک در دسترس شما قرار گرفته است.
حال به پرسش‌های زیر پاسخ دهید(در یک یا نهایت دو خط با ارائه یک نمونه توضیح دهید):

۱. با استفاده از لینک داده شده دو نمونه رتبه‌بندی ایجاد کنید، با این شرایط که MRR رتبه‌بند دوم بهتر از رتبه‌بند اول باشد اما در تمام معیارهای P@۵، R@۵، AP و NDCG@۵ رتبه‌بندی اول بهتر عمل کرده باشد. (دقت کنید که عدد rel آیتم‌ها و تعداد آیتم‌های مرتب و غیرمرتب هر دو نمونه یکی باشد و صرفا ترتیب اسناد متفاوت باشد)
۲. در چه صورت و برای چه تسکی MRR یک معیار مفید برای ارزیابی مدل جستجو می‌تواند باشد؟
۳. آیا امکان دارد که P@۵ مدل رتبه‌بند اول بهتر از رتبه‌بند دوم باشد اما NDCG@۵ مدل دوم بهتر از مدل اول باشد؟
۴. در کدام یک از معیارها حساسیت به رتبه‌بندی اسناد نیز وجود دارد؟
۵. در مورد AP، آیا امکان دارد معیار P@۵ مدل اول با مدل دوم برابر باشد اما معیار AP یکی بهتری از دیگری باشد؟ آیا AP به recall حساس است؟ چرا؟ با بیان فرمول محاسبه AP دلیل حساسیت و استدلال خود را بیان کنید.

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IIR-CA1-StudentID تحویل داده شود.

- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد. دقت کنید که حتما گزارشات خود را در قالب ارائه شده برای تحویل تکالیف که در سامانه برای شما بارگذاری شده است ارسال بفرمایید.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آن‌ها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید. **دقت کنید که تمامی کدها باید توسط شما اجرا شده باشند و نتایج اجرا در فایل کدهای ارسالی مشخص باشد. به کدهایی که نتایج اجرای آن‌ها در فایل ارسالی مشخص نباشد نمره‌ای تعلق نمی‌گیرد.**
- تمرین تا یک هفته بعد از مهلت تعیین شده با تاخیر تحویل گرفته می شود. دقت کنید که شما جمعا برای تمام تکالیف، ۱۴ روز زمان تحویل بدون جریمه دارید که تنها از ۷ روز آن برای هر تمرین می‌توانید استفاده کنید، در صورتی که این ۱۴ روز به اتمام رسیده باشد، به ازای هر روز تأخیر در ارسال تمرین، ده درصد جریمه می‌شوید.
- **توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تشابه به همه افراد مشارکت کننده، نمره ۵۰- تعلق می‌گیرد و به استاد نیز گزارش می‌گردد.**
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:kargaran.sina@gmail.com

تاریخ آپلود تمرین	۴ آبان ۱۴۰۴

مهلت تحویل بدون جریمه	۱۴ آبان ۱۴۰۴
مهلت تحویل با تأخیر، با جریمه ۱۰ درصد	۲۱ آبان ۱۴۰۴

استفاده مسئولانه از هوش مصنوعی

۱. هدف و اصول کلی

هدف

- ترویج استفاده **اخلاقی** و **مسئولانه** از LLMها (مانند ChatGPT، Deepseek) به عنوان ابزار کمکی
- اطمینان از **مشارکت فعال** دانشجویان در تکالیف و درک راه‌حل‌های آن‌ها
- حفظ **صداقت علمی** در عین بهره‌گیری از ابزارهای مدرن هوش مصنوعی

اصول کلی

- تمرین باید نتیجه تلاش و زحمت شخصی شما باشد.
- باید به تمام بخش‌های تمرین، اعم از پیاده‌سازی و تحلیل نتایج مسلط باشید.
- تمامی کدها باید توسط خود شما اجرا شده و نتایج قابل مشاهده باشند.
- تمام مراحل انجام تمرین باید مستند و قابل پیگیری باشد.
- هرگونه نتیجه‌گیری و تحلیل باید بر اساس درک شخصی شما باشد.
- **LLMها ممکن است پاسخ‌های نادرست یا قدیمی تولید کنند، اولویت با مطالب و کارگاه‌های درس است.**

موارد ذکر شده در ادامه این سند، به عنوان **راهنمایی بیشتر** برای انجام تمرین آورده شده‌اند. با این حال، **مسئولیت تطبیق کار با اصول کلی فوق بر عهده شماست**. توجه داشته باشید که ممکن است مواردی در ادامه ذکر نشده باشند که با اصول کلی ذکر شده در تضاد باشند. در چنین مواردی به تشخیص دستیار آموزشی و دستیار مسئول، شما موظف به پاسخ‌گویی در قبال تمرین خود هستید. عدم رعایت هر یک از اصول فوق می‌تواند منجر به کسر نمره یا عدم پذیرش تمرین شود.

۲. استفاده مجاز از LLMها

شما می‌توانید از LLMها برای موارد زیر استفاده کنید:

- روشن‌سازی مفاهیم (مثال: "خوشه‌بندی DBSCAN چگونه کار می‌کند؟")
- **کمک** در اشکال‌زدایی (مثال: شناسایی خطاهای گرامری یا منطقی در کد)
- ایده‌پردازی رویکردها (مثال: "روش‌های مدیریت داده‌های missing را پیشنهاد دهید")

الزامات استفاده مجاز:

- ثبت تعاملات اصلی: (به بخش ۴ مراجعه کنید.)
- درک راه‌حل: باید قادر به توضیح هر خط کد یا منطق استفاده شده باشید.

۳. استفاده غیرمجاز از LLMها

اقدامات ممنوع شامل:

- کپی-پیست مستقیم خروجی‌های LLM بدون تغییر
- استفاده از LLMها برای حل اصلی مسائل (مثال: "این سؤال تکلیف را برای من حل کن")
- گرفتن کد از سایر دانشجویان به هر شکل غیر مجاز است، تغییر و پارافریز کردن کد دیگران توسط LLM نیز قابل قبول نیست.
- هرگونه استفاده که منجر به عدم احاطه شما به موضوع تمرین شود.

۴. مستندسازی

ارجاع به مشارکت‌های LLM: افزودن پانویس یا توضیح (مثال: کد با رعایت قوانین به کمک ChatGPT نوشته شده است.)

- نیازی به اشتراک گذاری پرامپت‌ها و سابقه چت نیست.
- مستندسازی تک تک تعاملات با هوش مصنوعی هدف این بخش نیست. اشاره کوتاه و کلی در بخش‌های مورد استفاده کافی است. در نظر داشته باشید که مستندسازی به معنای رفع مسئولیت نبوده و باید **اصول کلی** را رعایت کنید.

۵. آمادگی ارائه شفاهی

آماده دفاع از کار خود باشید: در صورت درخواست دستیار تمرین در بازه زمانی اعلام شده برای ارائه شفاهی، باید:

- رویکرد، کد یا نتایج خود را توضیح دهید.
- درک مفاهیم کلیدی را نشان دهید (مثلاً چرا یک الگوریتم خاص انتخاب شده است)

عدم توضیح کافی کار شما ممکن است منجر به جریمه شود (بخش ۶)

۶. پیامدهای تخلفات

- **تخلفات جزئی (مثل مستندسازی ناقص):** کاهش نمره
- **تخلفات عمده (مثل کپی-پیست بدون تغییر):** نمره ۵۰- در تکلیف
- **تخلفات مکرر:** نمره ۵۰- در تکلیف و گزارش به استاد

۷. موارد تکمیلی

- **از LLMها به عنوان معلم استفاده کنید، نه پاسخ‌نامه تمرین‌ها:** اولویت را به مهارت‌های حل مسئله خود بدهید.

- **خروجی‌ها را متقابلاً تأیید کنید:** پیشنهادات LLM را با کتاب مرجع درس، اسلایدها و کارگاه‌ها مقایسه کنید.
- **از دستیاران آموزشی کمک بگیرید:** اگر پاسخ LLM یا نحوه استفاده شما را گیج می‌کند، در ساعات متعارف از دستیاران آموزشی کمک بگیرید.

۸. اظهارنامه

این عبارت را در تکلیف ارسالی خود قرار دهید:
"تأیید می‌کنم که از LLMها مطابق با دستورالعمل‌های بارگذاری شده در سامانه Elearn درس به طور مسئولانه استفاده کرده‌ام. تمام اجزای کار خود را درک می‌کنم و آماده بحث شفاهی درباره آنها هستم."