

به نام خدا



دانشگاه تهران
دانشکدگان فنی
دانشکده مهندسی برق و کامپیوتر



درس بازیابی هوشمند اطلاعات

تمرین ۲

استاد درس: خانم دکتر آزاده شاکری

سرپرست دستیاران آموزشی: سمانه پیمانی راد

طرح تمرین: امیرحسین روشن دل

آبان ماه ۱۴۰۴

فهرست

| | | |
|----|-------|---|
| ۳ | | مقدمه |
| ۵ | | سوالات عملی (۷۰ نمره) |
| ۶ | | مجموعه داده |
| ۸ | | پیش نیازها - ایجاد شاخص |
| ۹ | | سوال ۱: هموارسازی (۴۰ نمره) |
| ۱۲ | | سوال ۲: تحلیل واژه‌های نادر و تأثیر گسترش پرس‌وجو (۳۰ نمره) |
| ۱۴ | | سوالات تئوری (۴۰ نمره) |
| ۱۵ | | سوال ۱: تحلیل و مقایسه روش‌های هموارسازی (۱۵ نمره) |
| ۱۷ | | سوال ۲: PRF و الگوریتم EM (۱۰ نمره) |
| ۱۸ | | سوال ۳: تحلیل رفتاری مدل‌های شباهت زبانی در رتبه‌بندی اسناد (۱۵ نمره) |
| ۱۹ | | ملاحظات (حتماً مطالعه شود) |
| ۲۱ | | استفاده مسئولانه از هوش مصنوعی |
| ۲۱ | | ۱. هدف و اصول کلی |
| ۲۱ | | ۲. استفاده مجاز از LLM‌ها |
| ۲۲ | | ۳. استفاده غیرمجاز از LLM‌ها |
| ۲۲ | | ۴. مستندسازی |
| ۲۲ | | ۵. آمادگی ارائه شفاهی |
| ۲۲ | | ۶. پیامدهای تخلفات |
| ۲۳ | | ۷. موارد تکمیلی |
| ۲۳ | | ۸. اظهارنامه |

با توجه به افزایش روزافزون حجم اطلاعات و تنوع بالای روش‌های بازیابی اطلاعات، استفاده از تکنیک‌های پیشرفته در این زمینه اهمیت فراوانی پیدا کرده است. این تکنیک‌ها به منظور بهبود کیفیت و دقت دسترسی به اطلاعات به کار می‌روند و نقش مهمی در بهینه‌سازی نتایج جستجو دارند. یکی از مؤثرترین ابزارهایی که در فرآیند بازیابی اطلاعات مورد استفاده قرار می‌گیرد، استفاده از مدل‌های وزن‌دهی Pseudo Relevance Feedback است. این روش‌ها با تحلیل نتایج اولیه جستجو و تطبیق آن‌ها با معیارهای ارتباط، به رتبه‌بندی و بهبود کیفیت نتایج بازیابی کمک می‌کنند.

از طرفی، تکنیک‌های هموارسازی نیز با هدف بهبود دقت و پایدارسازی مدل‌ها، در بهینه‌سازی فرآیند بازیابی اطلاعات نقش بسزایی ایفا می‌کنند. این روش‌ها از طریق کاهش نوسانات و بهینه‌سازی پارامترهای مدل، امکان ارائه نتایج دقیق‌تر و مرتبط‌تر را فراهم می‌سازند. به طور کلی، استفاده از این ابزارها و تکنیک‌ها منجر به بهبود کارایی سامانه‌های بازیابی اطلاعات و افزایش دقت در پاسخ‌گویی به نیازهای کاربران می‌شود.

ابزار مورد استفاده در این تمرین یک موتور جستجو و شاخص‌گذاری ساده به زبان پایتون است که توسط دستیاران آموزشی این درس به‌طور بخصوصی طراحی و توسعه داده شده است تا دانشجویان امکان انجام آزمایش‌ها را داشته باشند.

اهداف تمرین:

- پیاده‌سازی تکنیک‌های هموارسازی و تابع وزن‌دهی در بازیابی اطلاعات
- انجام آزمایش‌ها و ارزیابی عملکرد روش‌ها بر روی مجموعه داده‌های موجود
- مقایسه نتایج به‌دست‌آمده و تحلیل تأثیر هر روش بر بهبود دقت بازیابی اطلاعات

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 به صورت فرضی بیان گردند نمره‌ای تعلق نمی‌گیرد.
- نمره اصلی تمرین مربوط به کیفیت و درستی تفسیرهای ارائه شده است.
- بهتر است از **نمودارها** و کشف نمونه‌های مرتبط از اسناد و پرس‌و جوها 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 موقول نکنید! **اجرای آزمایش‌ها و پیدا کردن پارامتر** بهینه زمان بر بوده و تفسیر هر کدام نیازمند تحلیل و بررسی است.

سوالات عملی (۷۰ نمره)

در تمرین اول با معیارهای ارزیابی و توابع امتیازدهی به اسناد آشنا شدید دیدید که یک تابع امتیازدهی با توجه به میزان ارتباط یک سند با پرس وجو امتیازی به سند تخصیص می‌دهد تا در نهایت اسناد براساس امتیازشان رتبه بندی و نمایش داده شوند. در این تمرین قصد داریم روش‌های مختلف هموارسازی توابع بازیابی و Pseudo Relevance Feedback را مورد مطالعه قرار بدهیم.

پاسخ سوالات این بخش را در همان نوتبوک تمرین خود (*IIR-CA2-Code.ipynb*) بنویسید.

باتوجه به حجم مجموعه داده اجرای این کد به طور تقریبی حداقل نیازمند ۶ 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). برای بارگذاری فایل دادگان در گوگل کولب می‌توانید از لینک گوگل درایو قرارداده شده در سامانه ایلن استفاده کنید.

مجموعه داده

برای این تمرین، مجموعه داده‌ی **ANTIQUE**^۱ شامل ۲۶۲۶ سؤال به زبان طبیعی در حوزه‌ی باز است که از دسته‌بندی‌های متنوعی گردآوری شده‌اند. این سؤالات توسط کاربران واقعی در سرویس پرسش‌وپاسخ **crowdsourcing** **Yahoo! Answers** (جمع‌سپاری) مطرح شده‌اند. برای هر سؤال، تمام پاسخ‌های موجود از طریق **crowdsourcing** (جمع‌سپاری) مورد ارزیابی قرار گرفته و در مجموع ۱۱^{۳۴۰} برجسب ارتباطی به صورت دستی ثبت شده است.

پیکره متنی^۲ (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^۳:

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

دادگان طلایی^۴:

این فایل شامل ۶۵۸۹ قضاوت‌های مرتبط^۵ همراه با برجسب میزان ارتباط می‌باشد که در مرحله نهایی جهت ارزیابی کارایی توابع بازیابی مورد استفاده قرار می‌گیرد. بر حسب میزان ارتباط پرس‌وجو با سند، عدد برجسب

۱ ANTIQUE: A Non-factoid Question Answering Benchmark

۲ Corpus

۳ Queries

۴ Golden Dataset

۵ 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 به عنوان پاسخ به سؤال پذیرفت. — ۲.۴ هزار (۳۶.۷٪)
۳. می‌تواند پاسخی به سؤال باشد، اما به اندازه‌ی 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.

پیش نیازها – ایجاد شاخص^۶

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 شاخص‌گذاری 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 به کار 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، به جز فایل retrieval_models.py نیست. در سوال ۱ و ۲ درباره‌ی اینکه چه کدهایی باید به این بخش اضافه کنید توضیح داده شده است.

^۶ indexing

سوال ۱: هموارسازی (۴۰ نمره)

یکی از مشکلات اساسی در مدل بازیابی Query Likelihood (QLM) ، که از رویکردهای مبتنی بر مدل سازی زبانی است، زمانی رخ می دهد که مدل زبانی اسناد $P(w|D)$ با استفاده از تخمین حداکثر درستنمایی (MLE) برآورد شود.

در این حالت، کلماتی از پرس و جو (Query) که در یک سند خاص مشاهده نشده اند، احتمال صفر می گیرند. این «مشکل احتمال صفر» باعث می شود که کل درستنمایی پرس و جو $P(Q|D) = \prod_{w \in Q} P(w|D)$ نیز صفر شده و رتبه بندی سند با شکست مواجه شود.

برای حل این مشکل، از روش های هموارسازی (Smoothing) استفاده می شود تا با اختصاص دادن مقداری احتمال به کلمات دیده نشده، رخداد آنها در اسناد به شکلی غیر صفر تخمین زده شود.

هدف از این سوال آشنایی با روش های هموارسازی می باشد. با استفاده از نمودار مناسب برای هر یک از روش های خواسته شده مقادیر مختلف λ و δ را بررسی کنید و مقدار بهینه را گزارش نمایید.

روشهایی که قصد داریم در این تمرین مورد بررسی قرار دهیم عبارتند از:

- روش JM با پارامتر λ
- روش Dirichlet Prior با پارامتر μ
- روش هموارساز Absolute Discounting با پارامتر δ

$$p(w|d) = \frac{\max(c(w,d) - \delta, 0) + \delta |d|_u p(w|C)}{|d|}$$

فرض کنید مجموعه‌ای از پاسخ‌ها از سایت Yahoo Answers (ANTIQUE) در اختیار دارد. پاسخ‌ها طول‌های متفاوتی دارند برخی بسیار کوتاه، برخی متوسط و برخی بلند هستند. ممکن است مدل‌های هموارسازی در مواجهه با این تفاوت طول رفتار متفاوتی از خود نشان دهند.

برای بررسی این موضوع باید دو آزمایش انجام دهید:

الف) ابتدا اسناد را بر اساس طولشان دسته‌بندی کنید به سه گروه:

- کوتاه (Short): کمتر از ۵۰ کلمه
- متوسط (Medium): بین ۵۰ تا ۱۵۰ کلمه
- بلند (Long): بیش از ۱۵۰ کلمه

سپس تعداد اسناد در هر گروه، درصد آن‌ها نسبت به کل، و میانگین طول هر گروه را محاسبه و چاپ کنید.

ب) در هر گروه، مدل‌های هموارسازی زیر را اجرا کنید:

- Jelinek–Mercer
- Dirichlet Prior
- Absolute Discounting

ج) در مرحله‌ی بعد، همین مدل‌ها را روی کل اسناد بدون دسته‌بندی اجرا کنید و نتایج را با میانگین گروه‌های سه‌گانه مقایسه نمایید. بررسی کنید که آیا تفکیک اسناد بر اساس طول باعث تغییر معنی‌دار در عملکرد مدل‌ها می‌شود یا خیر.

در گزارش خود به این پرسش‌ها پاسخ دهید:

- کدام گروه بیشترین تعداد سند را دارد؟
- در کدام گروه روش‌های هموارسازی عملکرد بهتری دارند؟
- چه ارتباطی میان طول سند و شدت تأثیر هموارسازی (smoothing bias) وجود دارد؟

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 به عنوان یک تابع جدید در ادامه فایل 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 P@5، nDCG@5، MAP، Recall@10 می‌باشند.

سوال ۲: تحلیل واژه‌های نادر و تأثیر گسترش پرس‌وجو (۴۰ نمره)

در بسیاری از سیستم‌های بازیابی اطلاعات، برخی پرس‌وجوها شامل واژه‌های نادر هستند؛ واژه‌هایی که در تعداد کمی از اسناد ظاهر می‌شوند.

چنین واژه‌هایی می‌توانند باعث شوند مدل زبانی تمرکز بیش از حد روی بخش کوچکی از اسناد داشته باشد یا در مدل‌های هموارسازی کم‌اثر شوند.

در این تمرین بررسی می‌کنید که:

- مدل‌های هموارسازی در مواجهه با پرس‌وجوهای دارای واژه‌های نادر چگونه عمل می‌کنند؟
- آیا استفاده از گسترش پرس‌جو (Query Expansion) می‌تواند عملکرد را بهبود دهد؟

سیستم پرسش‌وپاسخ شما بر پایه‌ی مجموعه‌داده‌ی ANTIQUE است. هدف، تحلیل نقش واژه‌های نادر و آزمودن راهکارهای زیر است:

۱. شناسایی واژه‌های نادر در کل مجموعه
۲. گروه‌بندی پرس‌وجوها به سه سطح مختلف از نظر واژه‌های نادر
۳. مقایسه مدل‌های هموارسازی در این گروه‌ها
۴. بهبود عملکرد با استفاده از روش‌های PRF

الف) گروه‌بندی پرس‌وجوها بر اساس واژه‌های نادر

کدی بنویسید که با استفاده از فراوانی واژه‌ها در کل مجموعه، واژه‌های نادر را شناسایی کند. (واژه‌هایی که در کمتر از ۱٪ از اسناد ظاهر می‌شوند).

- گروه `none`: شامل پرس‌وجوهایی است که هیچ واژه نادری ندارند.
- گروه `low`: شامل پرس‌وجوهایی است که دارای ۱ یا ۲ واژه نادر هستند.
- گروه `high`: شامل پرس‌وجوهایی است که بیش از ۲ واژه نادر دارند.

ب) اجرای مدل‌های هموارسازی (با پارامترهای بهینه‌ی سؤال ۱)

مدل‌های زیر را روی هر گروه از پرس‌وجوها اجرا کنید:

Jelinek–Mercer •

Dirichlet Prior •

Absolute Discounting •

مقادیر پارامترهای μ , λ و δ را مقادیر بهینه‌ی ای که در سوال ۱ پیدا کردید، در نظر بگیرید.

(ج) گسترش پرس‌وجو (Query Expansion)

اکنون بررسی کنید آیا گسترش پرس‌وجو می‌تواند عملکرد را بهبود دهد.

از سه روش زیر استفاده کنید:

Mixture Model PRF •

num_expansion_terms=۱۰, top_n=۱۰ با

Divergence Minimization PRF •

num_expansion_terms=۱۰, top_n=۱۰ با

Rm3 (امتیازی) •

num_expansion_terms=۱۰, top_n=۱۰ با

سوالات تئوری (۴۰ نمره)

باتوجه به مباحث تدریس شده سوالات زیر را حل کنید. پاسخ سوالات این بخش باید در قالب مربوط به سوالات تئوری (IIR-CA2-Theory) بنویسید.

نکات سوالات تئوری:

- صرفا به جواب نهایی نمره تعلق نمی‌گیرد و راه حل نوشته شده بخش اصلی نمره‌ی شماست.
- می‌توانید جهت اطمینان از محاسبات خود از کدنویسی استفاده کنید، اما باید راه حل به طور کلی به صورت ریاضی در گزارش تمرين آمده باشد و صرفا قراردادن کد راه حل باعث کسر نمره شما می‌شود.

سوال ۱: تحلیل و مقایسه روش‌های هموارسازی (۱۵ نمره)

فرض کنید مجموعه‌ای از اسناد با مشخصات زیر در اختیار دارید:

| شماره سند | محتوای سند | طول سند |
|----------------|---|---------|
| D ₁ | "بازیابی اطلاعات و یادگیری ماشین در علوم داده" | ۸ کلمه |
| D ₂ | "یادگیری عمیق برای بازیابی اطلاعات" | ۵ کلمه |
| D ₃ | "الگوریتم‌های پیشرفته یادگیری ماشین و کاربردهای آن در سیستم‌های توصیه‌گر و بازیابی اطلاعات" | ۱۳ کلمه |
| D ₄ | "روش‌های آماری در یادگیری" | ۴ کلمه |

$$P(\text{"یادگیری"} | C) = 0.003$$

$$P(\text{"بازیابی"} | C) = 0.002$$

$$|V| = 1000$$

الف) با استفاده از چهار روش هموارسازی زیر، احتمال کلمه "یادگیری" را در سند D₃ محاسبه کنید و مشخص کنید که کدام روش احتمال بالاتری می‌دهد و چرا؟

۱. Jelinek–Mercer smoothing ($\lambda=0.6$)

۲. Dirichlet Prior smoothing ($\mu=2000$)

۳. Absolute Discounting ($\delta=0.5$)

۴. Additive (Laplace) smoothing ($\alpha=1$)

ب) برای پرس و جوی

$q = "یادگیری بازیابی"$

با استفاده از روش ($\mu=2000$) Dirichlet Prior

• $\log P(q|D_3)$ و $\log P(q|D_1)$ را محاسبه کنید.

• توضیح دهید چرا با وجود تعداد بیشتر کلمات پرس و جو در D_3 ، امتیاز آن لزوماً از D_1 بالاتر نیست.

ج) فرض کنید مدل ترکیبی زیر تعریف شده است: (امتیازی)

$$P(w | d) = \alpha P_{JM}(w | d) + \beta P_{Dir}(w | d) + \gamma P_{AD}(w | d) + \theta P_{Add}(w | d)$$

به طوری که:

$$\alpha + \beta + \gamma + \theta = 1$$

• توضیح دهید در چه شرایطی هر یک از مؤلفه‌ها نقش مهم‌تری دارد (با توجه به طول سند و تعداد

واژه‌های منحصر به فرد).

• بیان کنید چگونه می‌توان وزن‌های $\alpha, \beta, \gamma, \theta$ را به صورت خودکار تعیین کرد.

سوال ۲: PRF و الگوریتم EM (۱۰ نمره)

سیستم بازیابی با پرس‌وجوه "الگوریتم یادگیری عمیق" اجرا شده و ۵ سند برتر به ترتیب زیر بازیابی شده‌اند:

| امتیاز اولیه | محظوظ | سند |
|--------------|---|----------------|
| ۱۲.۵ | "الگوریتم‌های یادگیری عمیق و شبکه‌های عصبی" | D ₁ |
| ۱۱.۸ | "یادگیری عمیق در بیتایی ماشین" | D ₂ |
| ۸.۳ | "الگوریتم‌های ژنتیک و بهینه‌سازی" | D ₃ |
| ۱۰.۲ | "یادگیری تقویتی عمیق" | D ₄ |
| ۹.۷ | "معماری‌های عمیق در پردازش زبان" | D ₅ |

فرض کنید از PRF برای Mixture Model استفاده می‌شود با $\lambda = 0.3$.

(الف)

در EM iteration اول الگوریتم

:E-step •

محاسبه $P(z_i=1|w_i)$ برای کلمات {الگوریتم، یادگیری، عمیق}

○ با فرض:

○ $P(w|\theta_F^{(0)})$ یکنواخت

○ $P(w|C)$ الگوریتم=۱۵، یادگیری=۰۰۰۲، عمیق=۰۰۰۱

:M-step •

محاسبه $\theta_F^{(1)}$ برای کلمات {الگوریتم، یادگیری، عمیق}

(ب)

• کدام کلمات وزن بیشتری می‌گیرند و چرا؟

• اگر λ افزایش یابد، تأثیر آن چیست؟

• آیا D₃ باعث نویز می‌شود؟ چگونه می‌توان آن را شناسایی کرد؟

سوال ۳: تحلیل رفتاری مدل‌های شباهت زبانی در رتبه‌بندی اسناد (۱۵ نمره)

در مدل‌های مبتنی بر شباهت توزیعی (مانند Query Likelihood و KL-Divergence)، هدف مقایسه‌ی توزیع واژه‌ها در پرس‌وجو و اسناد است تا اسنادی که از نظر زبانی به پرس‌وجو نزدیک‌ترند رتبه‌ی بالاتری بگیرند.

با این حال، در عمل رفتار این مدل‌ها همیشه با شهود انسانی یکسان نیست و ممکن است در برخی شرایط منجر به رتبه‌بندی‌های غیرمنتظره شود.

(الف)

توضیح دهید منظور از «شباهت توزیعی» میان پرس‌وجو و سند چیست و چرا ممکن است این معیار گاهی به نفع اسناد بلندتر یا کلی‌تر عمل کند، حتی اگر اسناد کوتاه‌تر از نظر معنایی دقیق‌تر باشند.

(ب)

نقش هموارسازی (Smoothing) را در این مدل‌ها بررسی کنید. سپس تحلیل کنید در چه شرایطی هموارسازی می‌تواند باعث افت عملکرد برای اسناد کوتاه یا خاص شود.

(ج)

پدیده‌ی Query Drift را در این چارچوب تعریف کنید و توضیح دهید چگونه اضافه‌شدن چند واژه‌ی عمومی به پرس‌وجو می‌تواند باعث تغییر توزیع زبانی و در نتیجه جابه‌جایی رتبه‌ی اسناد شود. دو روش مفهومی پیشنهاد دهید که بتوان با استفاده از آن‌ها از بروز Query Drift جلوگیری کرد یا اثر آن را کاهش داد.

(د)

مفهوم Distribution Shift را در زمینه‌ی بازیابی اطلاعات توضیح دهید و تحلیل کنید چرا مدل‌های زبانی ممکن است در مواجهه با اصطلاحات یا واژه‌های جدید عملکرد ضعیف‌تری داشته باشند. دو راهکار مهندسی یا طراحی پیشنهاد دهید که بتوان با استفاده از آن‌ها پایداری سیستم را در برابر چنین تغییراتی افزایش داد. (امتیازی)

ملاحظات (حتماً مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IIR-CA2-StudentID تحویل داده شود.

- خوانایی و دقیقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرين‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد. **دقیقت** کنید که حتماً 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. **دقیقت** 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.
- تمرين تا یک هفته بعد از مهلت تعیین شده با تأخیر تحویل گرفته می‌شود. دقیقت کنید که شما جماعت برای تمام تکالیف، ۱۴ روز زمان تحویل بدون جریمه دارید که تنها از ۷ روز آن برای هر تمرين می‌توانید استفاده کنید، در صورتی که این ۱۴ روز به اتمام رسیده باشد، به ازای هر روز تأخیر در ارسال تمرين، ۵ درصد جریمه می‌شود.
- توجه کنید این تمرين باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتمن تمرين 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:roshandel2004@gmail.com>

| | |
|--------------|---------------------------------------|
| ۱۸ آبان ۱۴۰۴ | تاریخ آپلود تمرین |
| ۲۸ آبان ۱۴۰۴ | مهلت تحويل بدون جريمہ |
| ۵ آذر ۱۴۰۴ | مهلت تحويل با تأخير، با جريمہ ۱۰ درصد |

استفاده مسئولانه از هوش مصنوعی

۱. هدف و اصول کلی

هدف

- ترویج استفاده اخلاقی و مسئولانه از LLM‌ها (مانند Deepseek, ChatGPT) به عنوان ابزار کمکی
- اطمینان از مشارکت فعال دانشجویان در تکالیف و درک راه حل‌های 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 راه حل: 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 پaramپت‌ها 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 تمرين‌ها: 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 ساعت 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."