به نام خدا

دانشگاه تهران
دانشکدگان فنی
دانشکده مهندسی برق و کامپیوتر

درس بازیابی هوشمند اطلاعات

تمرین ۵

استاد درس: سرکار خانم دکتر آزاده شاکری

سرپرست دستیاران آموزشی: سمانه پیمانی‌راد

طراح تمرین: محمد جواد رنجبر

دی ماه ۱۴۰۴

فهرست

مقدمه

اهداف اصلی تمرین:

- پیاده‌سازی و مقایسه روش‌های مختلف رتبه‌بندی اسناد
- آشنایی عملی با معماری‌های bi-encoder و cross-encoder و تفاوت‌های کاربردی آن‌ها
- تنظیم دقیق مدل‌های پیش‌آموزش‌دیده با روش‌های تنظیم‌دقیق
- طراحی و ارزیابی سیستم‌های Hybrid Retrieval و Reranking مبتنی بر LLM
- تحلیل عمیق نتایج با استفاده از معیارهای ارزیابی استاندارد و درک trade-off بین روش‌های

نکات قابل توجه در هنگام پاسخ به سؤالات:

- قابلیت بازتولید: تمامی کدها باید reproducible باشند. از random_seed برای نتایج قابل تکرار استفاده کنید.
- تفسیر مبتنی بر داده: پاسخ‌های نظری باید با نتایج آزمایش‌های شما پشتیبانی شوند. تفسیرهای صرفاً فرضی بدون تحلیل داده نمره نمی‌گیرند.
- تحلیل کیفی و کمی: از جداول مقایسه‌ای، نمودارها، و بررسی موردی برای تقویت تفسیرهای خود استفاده کنید.
- مدیریت هزینه: در بخش LLM reranking، ابتدا روی زیرمجموعه کوچک تست کنید تا از صحت promptخود اطمینان حاصل کنید.
- مدیریت زمان: تنظیم‌دقیق و ارزیابی زمان‌بر هستند. از Google Colab یا منابع GPU استفاده کرده و انجام تمرین را به آخر موکول نکنید.

سوال ۱: رنک کردن با استفاده از شبکه‌های عصبی

در این تمرین، به پیاده‌سازی و ارزیابی رویکردهای مختلف رتبه‌بندی اسناد، از بازیابی پراکنده[1] تا مدل‌های عصبی مدرن می‌پردازیم. تمرکز اصلی بر به‌کارگیری مدل‌های از پیش‌آموزش‌دیده از طریق تنظیم دقیق کامل[2] و روش‌های تطبیق کارآمد پارامتری[3] خواهد بود. علاوه بر این، نقش مدل‌های زبانی بزرگ در رتبه‌بندی مجدد و بهبود پرس‌وجوها مورد بررسی قرار می‌گیرد.

بخش ۱: مجموعه داده

برای این تمرین، قصد داریم از مجموعه‌داده BEIR/nfcorpus استفاده کنیم. این یک مجموعه داده تخصصی در حوزه بازیابی اطلاعات پزشکی/تغذیه است که پرس‌وجوهای مرتبط با موضوعات پزشکی و پیکره‌ای از اسناد علمی پزشکی را دربردارد

پیکره[4] متنی:

- تعداد اسناد: تقریباً ۳۶۰۰ سند پزشکی
- ساختار هر سند:
  - _id : شناسه یکتای سند
  - Title: عنوان مقاله یا سند
  - Text: متن کامل مقاله یا پاساژ پزشکی

پرس‌وجوها:

- تعداد کل پرس‌وجوها: ۳۲۳ پرسش به زبان طبیعی
- محتوا: سؤالات واقعی کاربران درباره موضوعات پزشکی و تغذیه
- ساختار هر پرس‌وجو:
  - _id : شناسه یکتای پرس‌وجو
  - Text: متپ پرسش

دادگان ارتباط[5]:

- فرمت: فایل tsv جدا
- ستون‌ها:

[1] Sparse Retreival
[2] Full fine-tuning
[3] Parameter-Efficient Fine-Tuning (PEFT)
[4] Corpus
[5] Qrels

- query-id: شناسه پرس‌وجو
- corpus-id: شناسه سند
- Score: امتیاز ارتباط

- متوسط تعداد اسناد مرتبط به هر پرس‌وجو:۳۸,۲

تقسیم‌بندی داده:

مجموعه داده شامل تقسیم‌بندی‌های آموزش، توسعه و آزمون در فایل qrels می‌باشد:

- از بخش آموزش برای تنظیم دقیق مدل استفاده کنید.
- از بخش آزمون برای ارزیابی نهایی عملکرد مدل استفاده کنید.

کاربردها:

این مجموعه داده برای ارزیابی سیستم‌های بازیابی اطلاعات در موارد زیر مناسب است:

- جستجوی متون پزشکی
- سیستم‌های پرسش و پاسخ پزشکی
- موتورهای جستجوی تخصصی حوزه سلامت
- ارزیابی مدل‌های Zero-shot در حوزه تخصصی

بخش ۲: بازیابی با استفاده از مدل‌های پیش‌اموزش دیده

الف) بازیابی بدون آموزش مدل‌های

یک سیستم رتبه‌بندی بر اساس BM25 پیاده‌سازی کنید. پیاده‌سازی شما باید:

- تمام اسناد موجود در مجموعه داده را توکن‌بندی[6] و ایندکس‌گذاری[7] کند
- یک پرس‌وجو را دریافت کرده و اسناد top-k به همراه با نمراتشان را برگرداند
- از پارامترهای استاندارد BM25 استفاده کنید.(k1=1.5, b=0.75)

دقت کنید که با توجه به اینکه مفاهیم پایه BM25 در تمرین ۱ پوشش داده شده است، در این تمرین تمرکز اصلی بر روی مدل‌های عصبی است. لذا مجاز هستید برای بخش الف از کدهای قبلی خود یا کتابخانه‌های استاندارد استفاده کنید.

ب) بازیابی متراکم مبتنی بر bi-encoder

سیستم بازیابی bi-encoder را با استفاده از مدل از پیش آموزش‌دیده sentence-transformers/msmarco-distilbert-base-tas-b پیاده‌سازی کنید. سیستم شما باید:

---

[6] Tokenize
[7] Index

- تمام اسناد موجود در پیکره را به بردار متراکم تبدیل کند. (این کار یک بار در زمان indexing انجام شود)
- پرس‌وجو را در زمان جستجو رمزگذاری کند.
- شباهت کوسینوسی بین پرس‌وجو و document embeddings محاسبه کند.
- top-k نتایج را بازگرداند.

تفاوت بنیادی معماری بین bi-encoders و cross-encoders را توضیح دهید. چرا bi-encoders می‌توانند document embeddings را پیش‌محاسبه کنند اما cross-encoders نمی‌توانند؟ پیچیدگی محاسباتی بازیابی از بین N سند برای هر رویکرد چیست؟

پ) بازیابی متراکم مبتنی بر پایه cross-encoder

یک خط لوله بازیابی دو مرحله‌ای پیاده‌سازی کنید:

- مرحله ۱: از bi-encoder خود برای بازیابی ۱۰۰ کاندیدای برتر استفاده کنید
- مرحله ۲: این ۱۰۰ کاندیدا را با مدل cross-encoder/ms-marco-MiniLM-L-6-v2 دوباره رتبه‌بندی کنید

cross-encoder باید هر جفت پرس‌وجو–سند را مستقیماً امتیازدهی کند و ۱۰ نتیجه نهایی برتر را بازگرداند.

سوال نظری: چرا معمولاً خط لوله دو مرحله‌ای (bi-encoder + cross-encoder) استفاده می‌شود و از cross-encoder تنها استفاده نمی‌شود؟ از هردوجنبه‌ی دقت و کارایی بررسی کنید.

بخش ۴: بازیابی با استفاده از مدل‌های تنظیم دقیق شده

در این بخش، هدف تنظیم دقیق مدل sentence-transformers/all-MiniLM-L6-v2 بر روی مجموعه داده است. مدل پیاده‌سازی شده توسط شما باید شامل مراحل زیر باشد:

الف) آماده‌سازی داده‌ها

- ساخت Training Triplets: که باید به صورت (query, positive_document, negative_document) باشد.
- شناسایی مثبت‌ها: از qrels برای پیدا کردن اسناد مرتبط (relevance > 0) استفاده کنید.
- نمونه‌گیری منفی‌ها
  1. برای هر پرس‌وجو، ۱ مورد Hard Negative از بین ۵۰ نتیجه برتر BM25 که در واقع مرتبط نیستند، انتخاب کنید.
  2. تعداد ۲ مورد منفی به صورت تصادفی از کل پیکره انتخاب کنید.

ب) فرآیند آموزش

Component	Details
Loss Function	MultipleNegativesRankingLoss (Contrastive Loss)
Hyperparameters	Training for 3 epochs with Batch Size = 16
Optimizer	Learning Rate = 2×10^{-5} with Warmup Steps equal to 10% of total training steps
Checkpoints	Save model at the end of each epoch

مدل تنظیم‌دقیق شده را با نسخه Pre-trained اصلی بر روی مجموعه‌داده آزمون مقایسه و ارزیابی کنید.

د) **سوال نظری**: مفاهیم زیر را توضیح دهید:

1. مفهوم Hard Negative Mining را توضیح دهید.
2. چرا Hard Negatives برای آموزش مدل‌های بازیابی اطلاعات مفیدتر از Random Negatives هستند؟
3. اگر در فرآیند آموزش تنها از Hard Negatives استفاده کنید، چه مشکلاتی ممکن است برای مدل به وجود آید؟
4. مفهوم فراموشی فاجعه‌بار[8] چیست و راه‌حل‌های مقابله با آن چیست؟

بخش ۴: آموزش مدل‌ها با استفاده از تنظیم‌دقیق کارآمد پارامتری

در این بخش، به جای آموزش کامل، مدل پایه را با استفاده از تکنیک LoRA تنظیم دقیق می‌کنیم.

الف) پیاده‌سازی LoRA

Component	Details
PEFT Library	Using PEFT to add LoRA adapters to the base model
LoRA Settings	Rank (r) = 8, Alpha = 16
Target Matrices	Apply LoRA to the Query and Key projection matrices

ب) آموزش و ذخیره‌سازی

- از همان داده‌ها و پیکربندی ذکر شده در بخش قبل استفاده کنید.

ج) ارزیابی

- مدل LoRA-adapted را روی مجموعه‌داده آزمون ارزیابی کرده و عملکرد آن را گزارش کنید.
- تعداد پارامترهای LoRA را با حالت تنظیم دقیق کامل مقایسه کنید.

بخش ۵: ارزیابی جامع مدل‌ها

تمام سیستم‌های پیاده‌سازی شده خود را روی مجموعه‌داده آزمون با استفاده از معیارهای زیر ارزیابی کنید:

- MRR@10 (Mean Reciprocal Rank)

[8] Catastrophic Forgetting

- NDCG@10 (Normalized Discounted Cumulative Gain)
- Recall@10, Recall@100
- Precision@10
- Mean query latency

دو جدول مقایسه‌ای ایجاد کنید:

1. تمام روش‌های baseline: (BM25, pre-trained bi-encoder, cross-encoder pipeline)
2. تمام روش‌های قابل آموزش: (full fine-tuning, LoRA)

سوال نظری: با توجه به نتایج خود، در چه شرایطی تنظیم‌دقیق کامل و چه زمانی peft را در توصیه می‌کنید؟

بخش ۶: رنک کردن HYBRID

الف) یک hybrid ranker پیاده‌سازی کنید که ترکیبی از موارد زیر باشد:

- BM25 scores (normalized to [0,1] using min-max normalization across results)
- Your best-performing neural bi-encoder scores (cosine similarity, already in [0,1])

از ترکیب خطی وزنی استفاده کنید:

$$\text{final\_score} = \lambda \times \text{neural\_score} + (1 - \lambda) \times \text{BM25\_score}$$

- مقادیر $\lambda = [0.1, 0.3, 0.5, 0.7, 0.9]$ را تست کنید λ که بهترین NDCG@10 را می‌دهد گزارش کنید.

ب) **سوال نظری**: به سوالات زیر پاسخ دهید.

- توضیح دهید چرا hybrid retrieval systems که روش‌های sparse و dense را ترکیب می‌کنند، معمولاً بهتر از هر کدام به تنهایی عمل می‌کنند
- مشکل vocabulary mismatch چیست و چگونه dense retrievers آن را حل می‌کنند؟

بخش ۷: پیاده‌سازی یک RERANKER مبتنی بر LLM با استفاده از GPT-۵.۱-MINI API

الف) در این بخش یک reranker با استفاده از مدل‌های زبانی بزرگ طراحی می‌کنیم که به صورت زیر عمل می‌کند:

- از bi-encoder خود برای بازیابی top-20 candidates برای هر پرس‌وجو استفاده کند
- هر جفت پرس‌وجو–سند را به GPT-5.1-mini با یک prompt طراحی‌شده دقیق ارسال کند
- relevance judgment LLM را پارس کند (با مقیاس 0-10 یا دودویی relevant/not-relevant)
- ۲۰ کاندیدا را بر اساس LLM scores دوباره رتبه‌بندی کند
- top-10 results را بازگرداند.

برای پرامت دادن از هر دو روش زیر استفاده کنید:

- Zero-shot Direct Scoring: از مدل زبانی بخواهید میزان ارتباط را روی مقیاس 0-10 امتیاز دهد.
- Chain-of-Thought Reasoning: از مدل زبانی بخواهید ابتدا توضیح دهد که چرا سند مرتبط است و سپس امتیاز دهد.

دقت کنید برای مدیریت هزینه، ابتدا روی یک زیرمجموعه ۱۰ تایی از مجموعه‌داده خود این فرایند را انجام دهید. همچنین در نهایت نیز در صورت محدودیت منابع لازم نیست روی کل مجموعه‌داده پیاده‌سازی شود ولی حداقل ۲۰۰ نمونه را ارزیابی کنید.

ب) **سوال‌های نظری**:

- لطفاً LLM-based reranking و cross-encoder reranking را با هم مقایسه کنید، با تمرکز بر معیارهای: دقت، هزینه، تأخیر، قابلیت تفسیر، و سازگاری با حوزه‌های مختلف و سپس توضیح دهید که در چه شرایطی هر روش را ترجیح می‌دهید و چه معیارهایی در انتخاب بین آن‌ها مهم هستند.
- اگر پرس‌وجوها به فارسی و اسناد به انگلیسی باشند، چه رویکردهایی پیشنهاد می‌کنید؟ مزایا و معایب ترجمه ماشینی و تعبیه چندزبانه را مقایسه کنید.
- برای پرس‌وجوهایی که هیچ مثال مشابهی در training set ندارند، چه استراتژی‌هایی وجود دارد؟

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IIR-CA5-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 7 روز آن برای هر تمرین می‌توانید استفاده کنید، در صورتی که این 14 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 50ـ تعلق می‌گیرد و به استاد نیز گزارش می‌گردد.**
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:MohammadJavadRanjbarKalahroodi@gmail.com

تاریخ آپلود تمرین	۳۰ آذر ۱۴۰۴
مهلت تحویل بدون جریمه	۱۱ دی ۱۴۰۴
مهلت تحویل با تأخیر، با جریمه 10 درصد	۱۸ دی 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:

- کپی‌ـ‌پیست 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 کپی‌ـ‌پیست بدون تغییر)**: نمره ۵۰ـ در تکلیف
- **تخلفات مکرر**: نمره ۵۰ـ 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."