بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۴-۰۳

مدرس: احسان‌الدین عسگری

تمرین سوم | Multimodal NLP | مهلت ارسال: ۲۹ مرداد

- **مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.**
- **انتخاب چالش در کوئرا** : در تمرین‌هایی که چند چالش دارند، فقط یک نفر از هر گروه در کوئرا باید چالش مورد نظر گروه را انتخاب کند. امکان تغییر چالش تا قبل از زمان ددلاین انتخاب چالش وجود دارد. البته ذکر این نکته ضروری است که هر چالش محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند دارد. بنابراین در اسرع وقت برای انتخاب چالش اقدام کنید.
- **امکان ارسال با تأخیر و قوانین آن:** در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز (با سقف ۵ روز برای هر تمرین) وجود دارد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- **قابلیت بازاجرای کامل نوت‌بوک‌ها:** توجه داشته‌باشید که نوت‌بوک‌های شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در همان نوت‌بوک قرار داده شود.
- **نحوه آپلود فایل‌های پروژه و مدل‌ها:** تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند، باید به شکل فایل فشرده در **CW** آپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره‌شده یک مدل در صورتی که بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی (نظیر Google Drive) آپلود کرده و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
  کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتماً در گزارش کار، نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها ذکر شود.
- **اهمیت گزارش کار:** بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری که کرده‌اید را توضیح دهید؛ بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. چند نمونه از خروجی‌های مساله را در گزارش بیاورید و بر اساس آن، رفتار برنامه‌تان را تحلیل کنید. در صورتی که پارامتری مانند دقت، صحت یا هر معیار دیگری خواسته شده باشد، آن‌ها را محاسبه کرده و در گزارش خود ارائه دهید.
- **تأثیر موارد امتیازی:** دقت داشته باشید، موارد امتیازی که در این تمرین آمده است، صرفاً بر روی امتیاز همین تمرین اثر دارد و روی نمرات تمرین‌های دیگر تأثیرگذار نخواهد بود.
- **نحوه پرسش سؤال و رفع ابهام:** در صورت وجود هرگونه ابهام یا مشکل، در کوئرا مطرح کنید و از ارسال پیام مستقیم به تیم تدریس خودداری نمایید.

توضیحات تمرین

توضیحات کلی

در سومین تمرین درس، دانشجویان با یکی از پرکاربردترین موضوعات روز حوزه هوش مصنوعی یعنی **پیاده‌سازی سیستم‌های بازیابی و تولید چندوجهی (Multimodal Retrieval-Augmented Generation یا Multimodal RAG)** آشنا می‌شوند. این تمرین به صورت گروهی انجام می‌شود و لازم است اعضای هر گروه مسئولیت‌ها را به‌درستی میان خود تقسیم کرده و همه اعضا در پیشبرد پروژه به‌طور فعال مشارکت داشته باشند. هدف این

تمرین ایجاد یک زنجیره کامل از گردآوری داده‌های چندوجهی تا ساخت، آموزش، ارزیابی و تحلیل یک سیستم RAG کاربردی است که بتواند به سوالات متنی و تصویری پاسخ‌های دقیق و مرتبط ارائه دهد. در این تمرین تأکید بر استفاده از خروجی‌های تمرین‌های پیشین است؛ یعنی هر گروه باید داده‌هایی که در حوزه‌های فرهنگی، تاریخی، گردشگری، ادبی، مشاهیر، غذاها، یا سایر حوزه‌های تعریف‌شده تولید و پردازش کرده‌اند را پایه کار خود قرار دهد.

با رشد سریع مدل‌های چندوجهی و کاربردهای آن‌ها در سامانه‌های پرسش و پاسخ، موتورهای جستجو، سامانه‌های توصیه‌گر و ابزارهای تحلیل محتوا، توانایی یکپارچه‌سازی و پردازش اطلاعات متنی و تصویری اهمیت فراوانی پیدا کرده است. مدل‌های **Multimodal RAG** با استفاده همزمان از داده‌های متنی و تصویری امکان بازیابی و تولید پاسخ‌های دقیق‌تر و غنی‌تر را نسبت به مدل‌های تک‌وجهی فراهم می‌کنند. در این تمرین، دانشجویان باید با مسائل مربوط به جمع‌آوری، پاک‌سازی، تطبیق و ترکیب داده‌های چندوجهی آشنا شوند و بتوانند با اتکا به داده‌های ساختاریافته و متون تولیدشده در تمرین‌های قبلی یک مجموعه داده چندوجهی با کیفیت ایجاد نمایند.

فرآیند تمرین شامل چند گام اصلی است. ابتدا لازم است داده‌های متنی و تصویری مرتبط به صورت ساختارمند و پاک‌سازی شده گردآوری شوند. هر نمونه داده باید به گونه‌ای تهیه شود که ارتباط معنایی مناسبی بین بخش‌های متنی و تصویری وجود داشته باشد. سپس باید **مدل بازیابی و تولید چندوجهی انتخاب و پیاده‌سازی** گردد. انتظار می‌رود دانشجویان با مدل‌های پایه و متن‌باز مانند CLIP، BLIP، LLaVA، MiniGPT-4 یا سایر مدل‌های مشابه آشنایی پیدا کرده و بتوانند آن‌ها را متناسب با داده‌های فارسی تنظیم و آموزش دهند. مدل نهایی باید قادر باشد با دریافت یک ورودی متنی یا تصویری، اطلاعات مرتبط را از میان داده‌های چندوجهی بازیابی کرده و پاسخ مناسب تولید کند. در ادامه، **ارزیابی عملکرد مدل** از اهمیت بالایی برخوردار است. دانشجویان باید مجموعه‌ای از سوالات متنی و تصویری متنوع تهیه کنند و عملکرد مدل را با استفاده از معیارهای کمی مانند دقت و Recall و همچنین ارزیابی انسانی بررسی نمایند. نتایج مدل چندوجهی باید با روش‌های پایه مانند بازیابی صرفاً متنی یا صرفاً تصویری مقایسه شود تا نقاط قوت و ضعف مدل به دقت مشخص گردد. تحلیل کمی و کیفی خروجی مدل‌ها، بررسی چالش‌ها و ارائه پیشنهاد برای بهبود، بخش مهمی از گزارش نهایی را تشکیل می‌دهد.

در تمامی مراحل انجام تمرین **دقت در مستندسازی و گزارش‌دهی** مورد تاکید است و لازم است کلیه کدها، داده‌ها و مستندات پروژه به گونه‌ای ارائه شوند که امکان بازتولید و ارزیابی مجدد نتایج فراهم باشد.

انتخاب موضوع تمرین

همان طور که می دانید، در unimodal RAG متنی، ابتدا باید متون مرجع را embed کرده و در vector-db ذخیره کنید. در این تمرین باید ابتدا embedding یک modality دیگر را در کنار embedding متنی ذخیره کنید. دقت کنید که modality دوم شما باید حتما یکی از موارد زیر باشد: صوت، تصویر، ویدئو، مکان

در این تمرین باید یکی از موضوعات تمرین قبل را به اختیار خود ادامه دهید؛ یعنی دادگان غذاهای محلی، دادگان مشاهیر، دادگان منابع طبیعی و سایت های گردشگری، دادگان شعری و دادگان نظرات کاربران دیجیکالا. کاربردهایی (واقعی) از multimodal RAG در ادامه آمده اند که میتوانید یکی از آن‌ها را انتخاب کنید و یا اگر ایده‌ی جدیدی دارید، با توضیح کامل آن در مستندات، سراغ پیاده سازی آن بروید.

۱. معرفی مکان گردشگری بر اساس ۴ ویژگی مکانی «آب و هوا، ارتفاع، جمعیت، موقعیت توپوگرافیک»

* توضیح: انتظار می رود ۴ تصویر نقشه ایران بر اساس ۴ ویژگی گفته شده دانلود شود و در گزارش کار حتما بر اساس این تصاویر خروجی ها توضیح داده شوند.
* توضیح: موقعیت توپوگرافیک مانند دشت، ساحلی، بیابانی، کوهستانی، ...

۲. پرسش از ویژگی‌های مکان‌های گردشگری بر اساس تصویر آن‌ها

- مثلا بر اساس تصاویر مکان‌های گردشگری اصفهان، جاهایی که مناسب اطراق با چادر هستند را معرفی کن.

۳. سوال از چهره مشاهیر: ویژگی های مشاهیر با استفاده از چهره آن‌ها

- مثلا کدام شاعر قرن ۹ چهره گندم گون داشته است؟

۴. «دوست داریم خودمون ایده بزنیم :) »

جمع‌آوری داده

این بخش در جهت طراحی، گردآوری، پاک‌سازی و سازمان‌دهی مجموعه‌ای ساختاریافته از داده‌های چندوجهی متنی و تصویری است که باید به‌طور مستقیم با موضوع انتخابی هر گروه مرتبط باشند. تمرکز اصلی بر کیفیت و انسجام معنایی بین متن و تصویر است؛ به‌گونه‌ای که داده‌ها بتوانند مبنای آموزش یک مدل چندوجهی دقیق و کاربردی قرار گیرند.
یکی از الزامات کلیدی این تمرین، استفاده از **خروجی تمرین‌های پیشین گروه** به عنوان داده‌ی پایه‌ای است. این داده‌ها می‌توانند در حوزه‌های فرهنگی، تاریخی، گردشگری، مشاهیر، غذا و ... باشند. توسعه این داده‌ها با افزودن تصاویر مناسب و یا گسترش متنی، توصیه می‌شود. در صورت تمایل به فعالیت در حوزه‌هایی جدید مانند پزشکی، حقوقی یا سایر موضوعات تخصصی، **هماهنگی قبلی ضروری** می‌باشد.
برای گردآوری داده‌ها، استفاده از منابع عمومی و قابل اعتماد از جمله Wikipedia، وب‌سایت‌های تخصصی فارسی، کتاب‌های دیجیتال، و مقالات معتبر توصیه می‌شود.
فرآیند **پیش‌پردازش داده‌ها** باید با دقت بالا انجام گیرد. برای متون فارسی، اعمال نرمال‌سازی شامل حذف نویزهای نگارشی، اصلاح فاصله‌ها و به‌کارگیری ابزارهایی چون Hazm ضروری است. در مورد تصاویر نیز، حذف نمونه‌های تکراری، اصلاح اندازه به ابعاد استاندارد باید لحاظ شود.
داده‌ها باید در قالب ساختاریافته، ترجیحاً در قالب‌های JSON یا CSV ذخیره گردند.
آموزش مدل گام اول: پیاده‌سازی و تنظیم مدل بازیاب (Retriever)
ابتدا به این نکته توجه بفرمایید که برای این بخش، می‌توانید از خروجی تمرین قبلی‌تان استفاده کنید.

هدف مدل بازیاب، دریافت یک پرسش (متنی یا تصویری) و یافتن مرتبط‌ترین زوج‌های داده از پایگاه داده شماست. این کار از طریق تولید و مقایسه بردارهای بازنمایی (Embeddings) انجام می‌شود.
انتخاب مدل Encoder: برای تبدیل داده‌های چندوجهی خود به یک فضای برداری مشترک، باید از یک مدل Encoder چندوجهی استفاده کنید.

- برای داده‌های تصویری: مدل CLIP (Contrastive Language-Image Pre-Training) یک گزینه استاندارد و بسیار قدرتمند است.
- برای داده‌های صوتی: مدل‌هایی مانند CLAP (Contrastive Language-Audio Pre-Training) با منطقی مشابه CLIP عمل می‌کنند.

شما می‌توانید از نسخه‌های از پیش آموزش‌دیده این مدل‌ها (مثلاً از طریق Hugging Face) به عنوان نقطه شروع استفاده کنید.
تولید و ذخیره‌سازی Embeddingها:
برای هر نمونه داده در مجموعه خود:

- بُعد غیرمتنی (تصویر، صوت و ...) را به ورودی Encoder مربوطه در مدل بدهید تا بردار بازنمایی آن تولید شود.
- متن مرتبط را به ورودی Text Encoder مدل بدهید تا بردار بازنمایی متن تولید شود.
- این بردارها را در یک پایگاه داده برداری (Vector Database) مانند FAISS یا ChromaDB ذخیره کنید. این کار جستجوی سریع و بهینه را در میان میلیون‌ها بردار ممکن می‌سازد.

آموزش مجدد (Fine-tuning) مدل بازیاب:
برای دستیابی به عملکرد بهتر در حوزه تخصصی خود، می‌توانید مدل Encoder را بر روی داده‌های خودتان آموزش مجدد دهید. این کار با استفاده از یادگیری مقابله‌ای (Contrastive Learning) انجام می‌شود؛ به این صورت که مدل یاد می‌گیرد بازنمایی زوج‌داده‌های مرتبط را به هم نزدیک کرده و بازنمایی نمونه‌های غیرمرتبط را از هم دور کند.

گام دوم: یکپارچه‌سازی و زنجیره کامل
در نهایت، دو مدل بازیابی و تولید متن (توصیه می‌شود از یک مدل Large Vision Language Model مانند Gemma3 استفاده کنید) باید در یک زنجیره کامل به یکدیگر متصل شوند تا یک سیستم سرتاسری (end-to-end) شکل گیرد. فرآیند پاسخگویی به یک پرسش جدید به این صورت خواهد بود:

۱. پرسش کاربر (متنی یا تصویری) دریافت می‌شود.

۲. پرسش توسط مدل Encoder به بردار بازنمایی تبدیل می‌شود.

۳. با جستجو در پایگاه داده برداری، K نمونه داده چندوجهی برتر بازیابی می‌شوند.

۴. پرسش اصلی به همراه داده‌های بازیابی‌شده به مدل مولد تحویل داده می‌شود.

۵. مدل مولد، پاسخ نهایی را تولید و ارائه می‌کند.

ارزیابی مدل

این بخش سه مرحلهٔ پیوسته را در بر می‌گیرد که هدف آن سنجش دقت مدل‌های پایه و مدل RAG بر مجموعه‌داده‌ای چندوجهی است.

مرحلهٔ اول: ساخت مجموعهٔ آزمون
یک مجموعه ارزیابی در دو بخش تهیه می‌شود: (۱) دست‌کم پنجاه سؤال چندگزینه‌ای (MQC) متنی؛ (۲) مجموعه‌ای از سؤال‌های چندگزینه‌ای که متن را با یک مدالیته دیگر (صوت، تصویر، ویدئو یا دادهٔ مکانی) ترکیب می‌کند. تمام سؤال‌ها در دسته‌های موضوعی مشخص سازماندهی می‌شوند تا گزارش دقت حوزه‌ای نیز ممکن شود.

مرحلهٔ دوم: ارزیابی مدل‌های پایه
دو مدلِ چندزبانه بدون RAG روی هر دو بخشِ مجموعه آزمون اجرا می‌شوند. ورودی هر مدل صرفاً متن پرسش، در سؤال‌هایِ چندوجهی، توصیف کوتاهی از مدالیتهِ همراه است. نتایج با کلید پاسخ مقایسه می‌شود و دقت کل، دقت به‌تفکیک دستهٔ موضوعی و تحلیل خرد/کلان خطا گزارش می‌شود. اگر بیش از یک گروه روی یک موضوع کار می‌کند، ساخت این بخش می‌تواند به صورت مشارکتی انجام شود تا از تکرار یا عدم‌توازن داده جلوگیری گردد.

مرحلهٔ سوم: ارزیابی مدل RAG
همان آزمون با نسخه RAG تکرار می‌شود؛ علاوه بر متن پرسش، داده بازیابی‌شده به مدل داده می‌شود. افزون بر دقت پاسخ، کیفیت بازیابی با معیارهایی نظیر Hit@k، Precision و Recall برای اسناد بازگردانده‌شده محاسبه می‌شود. ثبت خروجی بازیاب ضروری است تا مشخص شود خطا از کدام مرحله (بازیابی یا تولید) نشأت گرفته است.

تحلیل نتایج

در این بخش از گزارش، انتظار می‌رود دانشجویان با تمرکز بر رفتار مدل و کیفیت خروجی‌ها به تحلیل دقیق‌تری از عملکرد سیستم خود بپردازند. هدف درک بهتر نقاط قوت و ضعف و شرایط مؤثر بر کارایی مدل در سناریوهای متنوع است. در گزارش به سوالات زیر پاسخ دهید. در صورت امکان برای هر مورد، در کنار توضیحات نمونه‌هایی از خروجی مدل بیاورید:

- آیا خروجی‌ها در حالت تصویری، متنی یا ترکیبی تفاوت معناداری دارند؟
- آیا بازیابی موفق و مرتبط انجام شده؟
- نقش promptها یا تنظیمات مدل چه بوده است؟
- آیا خروجی تولیدشده مبتنی بر اسناد بازیابی‌شده بوده یا صرفاً حدس مدل بوده است؟
- چه چالش‌هایی در زمینه زبان، ساختار سوال یا محتوای چندرسانه‌ای مشاهده شده است؟