بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۴-۰۳

مدرس: احسان‌الدین عسگری

تمرین دوم | مدل‌های زبانی | مهلت ارسال: ۱۶ خرداد

- **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

انتخاب موضوع تمرین

در ابتدای تمرین، هر گروه باید یکی از موضوعات زیر را به‌عنوان دامنه داده تمرین خود انتخاب کند. این انتخاب، نوع داده‌هایی را که در ادامه با آن‌ها کار خواهید کرد، تعیین می‌کند. تمام مراحل تمرین (شامل ساخت متون و سوالات آموزش، آموزش مدل و ارزیابی) باید بر پایه داده‌های مربوط به موضوع انتخاب‌شده انجام شود. انتخاب مناسب موضوع، با توجه به علایق و توانایی‌های گروه، می‌تواند تأثیر بسزایی در کیفیت نهایی پروژه داشته باشد. در ادامه، فهرست مجموعه‌داده‌هایی که می‌توانید انتخاب کنید، آمده است:

۱. **غذا:** این مجموعه‌داده شامل اطلاعات مربوط به غذاهای محلی از تمرین قبلی است و حدود ۲۰۰۰ رکورد داده دارد.

۲. **مشاهیر:** این مجموعه‌داده شامل اطلاعات مربوط به مشاهیر (ادبی، علمی، فرهنگی، ورزشی و...) از تمرین قبلی است و حدود ۹۰۰۰ رکورد داده دارد.

۳. **تاریخ:** این مجموعه‌داده شامل اطلاعات مربوط به وقایع، جنگ‌ها و قراردادهای تاریخی از تمرین قبلی است و حدود ۱۳۰۰ رکورد داده دارد.

۴. **منابع طبیعی و سایت‌های گردشگری:** این مجموعه‌داده شامل اطلاعات مربوط به مکان‌های گردشگری و منابع طبیعی چند استان ایران از تمرین قبلی است و حدود ۷۰۰ رکورد داده دارد.

۵. **شعر:** داده‌های مربوط به اشعار فارسی شامل نام شاعر، بیت، موضوع، سبک و دوره تاریخی از مجموعه‌داده‌هایی مانند گنجور.

۶. **نظرات کاربران دیجی‌کالا:** شامل نظرات کاربران درباره محصولات مختلف دیجی‌کالا، با عناصر ساخت‌یافته مانند امتیاز، نظر متنی، نقاط قوت و ضعف، که به‌صورت عمومی منتشر شده‌اند.

توجه داشته باشید که مجموعه داده‌های مربوط به تمرین اول دارای حقوق مالکیت هستند و صرفاً جهت استفاده در چارچوب این تمرین آموزشی در اختیار شما قرار گرفته اند. انتشار عمومی یا اشتراک‌گذاری آن‌ها با افراد خارج از این چارچوب مجاز نیست.

مراحل تمرین

در این تمرین، دانشجویان با فرآیند کامل ساخت یک مدل retrival از داده‌های ساختار یافته آشنا خواهند شد. مراحل تمرین شامل آماده‌سازی داده‌ها، تولید سوالات، آموزش مدل زبانی با استفاده از داده‌های تولیدشده، طراحی مجموعه ارزیابی، اجرای مدل‌های مختلف، تهیه خروجی‌های متنوع و در نهایت ارزیابی انسانی با ابزار Label Studio خواهد بود. هدف نهایی این تمرین، درک دقیق‌تری از نحوه عملکرد مدل‌های زبانی در درک و ایجاد بازنمایی مناسب در کاربردهای عملی است.

۱. ساخت دیتاست از داده‌های ساختار یافته

ابتدا، براساس دامنه انتخاب‌شده، مجموعه‌ای از داده‌های ساختاریافته در اختیار شما قرار می‌گیرد. وظیفه شما تبدیل این داده‌ها به متنی روان و قابل فهم برای انسان است، به‌گونه‌ای که بتوان سوالاتی از آن استخراج کرد.
به‌عنوان مثال، در مجموعه‌داده دیجی‌کالا، لازم است تعدادی محصول انتخاب کرده و از نظرات آن‌ها متنی استخراج کنید که بیانگر ویژگی‌های محصول باشد. یا در داده‌های تمرین قبلی، از بخش‌های مختلف فایل JSON متنی مانند نمونه زیر بسازید یا استخراج کنید:

```
{
  "name": "مریم میرزاخانی",
  "birth_year": 1977,
  "field": "ریاضیات",
  "award": "مدال فیلدز"
}
```

تبدیل به متن: «مریم میرزاخانی، متولد ۱۹۷۷، یکی از برجسته‌ترین ریاضی‌دانان ایرانی بود که موفق به دریافت مدال فیلدز شد.»
پس از ساخت متون، باید از هر متن حداقل ۵ سوال استخراج شود که جنبه‌های مختلف اطلاعات موجود را پوشش دهند. این جنبه‌ها شامل موارد ذکر شده در داده‌ها هستند (مثلا برای غذا می تواند دستور پخت، وعده غذایی، مواد تشکیل دهنده و ... باشد). سوالات می‌توانند به صورت دستی تولید شوند یا با کمک مدل‌های زبانی و اصلاح انسانی تهیه شوند. تنوع در ساختار پرسش‌ها و نوع اطلاعات هدف، نشانه کیفیت این بخش از تمرین است.
برای تبدیل داده به متن و طراحی سوالات می‌توانید از روش‌های مختلفی مانند الگوهای نگارشی (template-based)، برنامه‌نویسی مبتنی بر قاعده (rule-based)، یا استفاده از مدل‌های زبانی (مانند llama3 یا Gemini) بهره بگیرید. مهم

است که متون نهایی واضح، روان و دارای اطلاعات باشند، به گونه‌ای که بتوان بر اساس آن‌ها پرسش‌های طبیعی و متنوع طرح کرد.

۲. ساخت مجموعه داده‌های ارزیابی

حال که متون خود را آماده کردید باید برای ارزیابی دقیق مدل‌ها، یک مجموعه ارزیابی مستقل شامل ۵۰ سوال به صورت انسانی تهیه شود. این سوالات بهتر است به صورت کلی روی دامنه داده شما طرح بشوند. مثلا یک نمونه در دامنه غذا می‌تواند سوال "با گوجه، بادمجان و تخم‌مرغ چه غذایی می‌توانم تهیه کنم؟" باشد. و یا می‌توانید برخی از سوالات رو مستقیم از روی متن داده‌ها طراحی کنید. دقت کنید که سوالات شما نباید در آموزش استفاده شده باشند.

۳. آموزش مدل زبانی

پس از آماده‌سازی مجموعه داده‌ها، مدل زبانی پایه GLOT500 به عنوان مدل اولیه انتخاب می‌شود. این مدل یک مدل چندزبانه است که توسط HuggingFace معرفی شده و قابلیت اجرای وظایف زبانی مختلف را دارد. مطالعه مقاله این مدل خالی از لطف نیست. در این مرحله باید این مدل را با استفاده از داده‌های پرسش و متن متناظر با آن‌ها، آموزش مجدد دهید (fine-tuning) تا بازنمایی‌های سوالات و متن‌های متناظر با آن‌ها نزدیک به هم شوند. می‌توانید از روش contrastive learning استفاده کنید. با ابتکار در روش‌های آموزش می‌توانید به نتایج بهتری برسید. برای مثال شاید در ابتدا انجام یک آموزش اولیه مانند Masked Language Modeling بر روی متون بتواند کمک‌کننده باشد.

۴. خروجی گرفتن از مدل‌ها

در این بخش باید سه مدل زیر را بر روی ۵۰ سوال ارزیابی خود، بررسی کنید. برای هر یک از مدل‌ها باید سه خروجی برتر که بیشترین شباهت به سوال بیان شده دارند را مشخص و آن‌ها را ثبت کنید:

الف) یک روش پایه آماری، مانند استفاده از TF-IDF

ب) مدل پایه GLOT500 به صورت zero-shot ، بدون هیچ آموزشی

ج) مدل GLOT500 که شما در قسمت قبل آموزش داده‌اید.

در مجموع برای هر سوال، ۳ مدل و ۳ پاسخ از هر مدل، بین ۳ تا ۹ خروجی متفاوت خواهیم داشت که باید در مرحله بعد مورد ارزیابی انسانی قرار گیرند.

۵. ارزیابی انسانی با استفاده از Label Studio

در این بخش باید مدل‌های بخش قبل را ارزیابی کنید. در قسمت قبل برای هر یک از ۵۰ سوال ارزیابی خود ۳ تا ۹ خروجی متفاوت بدست آوردید. حال باید به ازای هر سوال خروجی‌ها را با توجه به میزان ارتباط آن‌ها به سوال رتبه‌بندی کنید. به این صورت که پاسخی که در جایگاه اول قرار می‌گیرد بیشترین ارتباط را با سوال دارد. دقت شود که اینکه هر خروجی برای چه مدلی است باید از دید ارزیاب پنهان بماند تا عدالت در مقایسه رعایت شود و ارزیاب دچار سوگیری نشود.

برای انجام این ارزیابی باید از نرم‌افزار Label Studio استفاده کنید و برای خروجی‌های خود یک پروژه ایجاد کنید. حداقل دو نفر باید هر سوال را به صورت مستقل ارزیابی کنند. ارزیابی‌ها باید به صورت ساختاریافته ذخیره شده و خروجی پروژه از Label Studio در فرمت JSON یا CSV برای تحلیل‌های بعدی استخراج شود.

۶. تحلیل نتایج و مقایسه عملکرد مدل‌ها

در پایان، نتایج ارزیابی باید به صورت آماری و کیفی تحلیل شوند. تحلیل شما باید شامل جداول مقایسه‌ای بین مدل‌ها، نمودارهای توزیع دقت مدل‌ها و بررسی رفتار مدل‌ها در انواع مختلف سوالات باشد. به‌طور خاص باید بررسی شود که مدل‌ها در پاسخ به چه نوع سوالاتی عملکرد بهتری دارند و در کدام موارد ضعف دارند.

همچنین تحلیل‌هایی از جمله تفاوت میان مدل پایه و فاین‌تیون‌شده، مقایسه بین روش آماری و مدل زبانی در گزارش نهایی گنجانده شود. تحلیل کیفی ارزیابی‌های انسانی می‌تواند دید خوبی از عملکرد واقعی مدل در کاربردهای عملی ارائه دهد.

مواردی که باید تحویل داده شوند:

- اسکریپت‌ها و کدها: شامل مراحل آماده‌سازی متن‌ها، تولید سوالات، آموزش مدل، تولید خروجی‌ها و ارزیابی.
- فایل‌های داده: شامل داده‌های متنی، پرسش تولیدشده و خروجی‌های هر مدل.
- پروژه Label Studio: فایل خروجی ارزیابی انسانی.
- گزارش تحلیلی: شامل تحلیل عددی، نمودارها، توضیحات روش، چالش‌ها، نقاط قوت و ضعف مدل‌ها.
- مستندات اجرا: شامل توضیح گام‌به‌گام روش‌ها، تنظیمات آموزش، پارامترهای مدل و تجربه اجرای پروژه.

امتیازی:
دانشجویانی که مایل به دریافت امتیاز اضافه هستند می‌توانند یک بخش اضافه طراحی کنند که در آن از مدل فاین‌تیون‌شده خودشان برای شناسایی نمونه‌های تکراری (Duplicate Detection) در مجموعه داده استفاده کنند.