بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۳-۰۲

مدرس: احسان‌الدین عسگری

تمرین چهارم | **تولید متن** | مهلت ارسال: ۲۷ تیر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.
- در تمرین‌هایی که چند چالش دارند، فقط یک نفر از هر گروه در گوگل فرم 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. بنابرین در اسرع وقت برای انتخاب چالش اقدام کنید.
- در طول ترم امکان ارسال با تاخیر برای هر تمرین ۵ روز و مجموع زمان مجاز تاخیر ۱۲ روز است. محل بارگزاری جواب تمرین‌ها بعد از ۵ روز، بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد. لازم به ذکر است که به دلیل تداخل زمان مجاز تاخیرها بین اعضای گروه در تمارین گروهی تمرین اول شامل تاخیر مجاز **نمی‌باشد**.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و مواردی از این دست) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- دقت داشته باشید، موارد امتیازی که در این تمرین آمده است، صرفا بر روی امتیاز همین تمرین اثر دارد و بر روی نمرات تمارین و یا بخش‌های دیگر درس، تاثیر ندارد.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به تیم تدریس خودداری کنید.

ساز و کار تمرین (ابتدا این بخش را به صورت کامل مطالعه نمایید.)

در این تمرین هر گروه یکی از موضوع‌های پیشنهادی را انتخاب خواهد کرد. درصورتی که در هرکدام از موضوعات تمایل دارید تا روی مجموعه دادگان دیگری کار کنید، توضیحات و آدرس مجموعه دادگان مدنظر را برای تیم تدریس در کوئرا ارسال کنید تا پس از بررسی و تایید تیم تدریس، بتوانید بر روی آن تمرین خود را انجام دهید.
برای موضوعاتی که در ترم‌های قبل نیز توسط دانشجویان مورد بررسی قرار گرفته است، بهترین کد آن‌ها در اختیار شما قرار داده خواهد شد تا از دوباره‌کاری جلوگیری شود. به تبع انتظار می‌رود تلاش شما روی آن موضوع باعث بهبود عملکرد کدهای قبلی شود.

در این تمرین نیاز است که از مدل‌های بر پایه ترنسفورمر استفاده نمایید.

برای موضوعات این تمرین معیارهای {۱،۲،۳،۴}-Bleu، Rouge و Bert-Score را گزارش کنید و اگر معیار دیگری لازم بود در خود ترک گفته شده است.

تولید عنوان برای مقاله خبری

در عصر اطلاعات، انتشار روزانه هزاران مقاله خبری امری رایج است و دسترسی به مهم‌ترین اخبار در کمترین زمان به یک چالش جدی تبدیل شده است. تولید عناوین دقیق برای مقالات خبری، به خوانندگان این امکان را می‌دهد تا به سرعت، اطلاعات کلیدی را دریافت کنند.
هدف این تمرین تولید عنوان مناسب از روی متن مقاله خبری می‌باشد.

داده‌ها

برای انجام این تمرین می‌توانید از دیتاست موجود در این لینک استفاده نمایید. این دیتاست شامل مقالات خبری با موضوعات مختلف، جمع آوری شده از ۶ منبع خبری مهر، ایرنا و ... می‌باشد. این دیتاست شامل ۹۳۲۰۷ نمونه می‌باشد که به سه دسته آموزش، ارزیابی و آزمون تقسیم شده است و هر نمونه علاوه بر متن خبر شامل اطلاعات دیگری همچون عنوان خبر، موضوع(دسته خبری) و ... می‌باشد. اطلاعات بیشتر راجع به این دیتاست در این لینک موجود است.

توجه:

دیتاست معرفی شده در قالب دیتاست‌های HuggingFace است بنابراین استفاده از آن برای فاین تیون کردن مدل‌های موجود روی HuggingFace راحت‌تر است ولی شما می‌توانید به جای این دیتاست از هر دیتاست مرتبط دیگری برای تولید عنوان استفاده نمایید.

نکته:

این لینک که از معماری‌های مختلف مثل LSTM برای تولید عنوان استفاده کرده، می‌تواند دید خوبی نسبت به این مسئله به ما بدهد ولی توجه داشته باشید که در این تمرین می‌بایست از مدل‌های مبتنی بر ترنسفورمر استفاده کرد.

تشخیص بیماری و پیشنهاد دارو

مدل‌های زبانی می‌توانند در حوزه پزشکی برای کاربران بسیار مفید باشند. در این ترک شما باید مدلی توسعه بدهید که علائم و نشانه‌های بیمار را دریافت کرده و ضمن تشخیص نوع بیماری، داروهای موثر را تجویز کند.

مدل و معیارهای سنجش

توجه داشته باشید برای این امر باید از یک مدل مبتنی بر ترنسفورمرها استفاده کنید که بر روی اطلاعات پزشکی آموزش دیده باشد و آن را بر روی داده‌هایی که در ادامه آمده است فاین تیون کنید.
بررسی عملکرد مدل بر اساس معیارهای مطرح در زمینه تولید متن قبل و بعد از فاین تیون شدن و مقایسه آن‌ها نیز مورد نیاز است.

داده‌ها

مجموعه دادگان مورد استفاده در این تمرین n2c2 نام دارد و به صورت عمومی در دسترس نیست. به همین جهت لطفا استفاده خود از مجموعه را به این تمرین محدود کنید.

مجموعه داده از طریق این لینک در دسترس بوده و رمز آن nlp1402 است.

سیستم پرسش و پاسخ پزشکی

در این ترک، هدف پیاده‌سازی مدلی برای پرسش و پاسخ در زمینه پزشکی، مبتنی بر بازیابی اطلاعات است. درواقع شما باید مدلی ارائه دهید که با دریافت یک سوال پزشکی و متون بازیابی شده مرتبط با سوال، پاسخ را بر اساس اطلاعات موجود در متن ورودی بازگرداند.

بازیابی متون مرتبط

قدم اول این است که پس از ورود سوال کاربر، شما متون پزشکی مرتبط با آن را بازیابی کنید. به این منظور به یک دیتاست از متون پزشکی و یک تکنیک بازیابی متون مرتبط (برای مثال cosine-similarity در حالت ساده) احتیاج دارید.یک دیتاست مناسب برای این منظور چکیده مقالات pubMed است.

تولید پاسخ

پس از بازیابی متون مرتبط، شما باید از یک مدل زبانی بزرگ به گونه‌ای استفاده کنید که با دریافت سوال و متون مرتبط بازیابی شده، پاسخ سوال را برای کاربر تولید کند. مدل انتخابی شما باید به گونه‌ای باشد که قابلیت فاین تیون کردن آن را با توجه به منابع محدود خود داشته باشید (برای مثال مدل T5 که می‌توان آن را با ریسورس‌های گوگل کولب فاین تیون کرد).

اینکه ترکیب سوال کاربر و متن بازیابی شده به گونه‌ای باشد که یک پرامپت مناسب برای مدل باشد، در کیفیت تولید پاسخ مدل موثر خواهد بود. این موضوع از موارد قابل توجه در نمره‌دهی این ترک خواهد بود. همچنین لازم است عملکرد مدل در پاسخگویی را قبل و بعد از فاین تیون شدن مورد بررسی قرار دهید.

دیتاست‌های پرسش و پاسخ

در این لینک چندین دیتاست پرسش و پاسخ پزشکی موجود است. استفاده از سایر دیتاست‌های موجود پرسش و پاسخ پزشکی مانند دیتاست های موجود در HuggingFace نیز مانعی ندارد.

خلاصه‌سازی مستندات داروها

مستندات دارویی غالباً پیچیده و مبهم هستند، و فهم آنها برای افراد غیر متخصص دشوار است. خلاصه‌سازی این اطلاعات با زبانی ساده و قابل فهم، به بیماران کمک می‌کند تا از نحوه عملکرد دارو، عوارض جانبی و روش مصرف آن آگاه شوند.
در این بخش از شما انتظار می‌رود مدلی برای خلاصه‌سازی مستندات تست داروها ارائه دهید که با دریافت توضیحات کامل مراحل تست یک دارو و نتایج آن، یک پاراگراف خلاصه شده شامل موارد مهم را در خروجی تولید کند.

مدل و معیارهای سنجش

توجه داشته باشید برای این امر باید از یک مدل مبتنی بر ترنسفورمرها استفاده کنید و آن را بر روی داده‌هایی که در ادامه آمده است فاین تیون کنید.
بررسی عملکرد مدل بر اساس معیارهای مطرح در زمینه تولید متن و خلاصه‌سازی (مشابه بخش خلاصه‌سازی متون چند زبانه) قبل و بعد از فاین تیون شدن و مقایسه آن‌ها نیز مورد نیاز است.

داده‌ها

برای این بخش می‌بایست از داده‌های ClinicalTrials.gov استفاده نمایید که یک پایگاه داده رایگان و قابل دسترس برای عموم است که اطلاعات مربوط به آزمایشات بالینی در حال انجام و به پایان رسیده را برای طیف گسترده‌ای از بیماری‌ها و شرایط ارائه می دهد.

خنثی‌سازی متون حاوی بایاس جنسیتی

بایاس جنسیتی به تعصب و تبعیض علیه یک گروه خاص بر اساس جنسیت آن‌ها اشاره دارد. این بایاس می‌تواند خود را از طریق زبان جنسیتی نشان دهد و ممکن است به اشتباه، ویژگی‌هایی را بر اساس جنسیت به یک فرد نسبت دهد.
هدف این تمرین خنثی کردن متونی که حاوی چنین بایاس‌هایی باشند، است و برای این کار از مدل‌های زبانی مبتنی بر ترنسفورمر استفاده می‌کنیم.
انجام این تمرین شامل چندین قدم به شرح زیر است:

۱. در ابتدا نیاز به مجموعه داده‌ای داریم که دارای "متن حاوی بایاس جنسیتی" و متن متناظر با آن ولی بدون بایاس جنسیتی (خنثی‌شده) باشد. از آنجایی که درحال حاضر و باتوجه به دانش نویسنده، مجموعه داده‌ای به این شکل توسعه داده نشده است، نیاز است که به جمع آوری داده بپردازیم.

برای جمع‌آوری داده می‌توانید به توییت‌های شبکه اجتماعی X و یا سایر منابع مراجعه کنید. در این مرحله سعی کنید از انواع بایاس‌های جنسیتی مانند موارد زیر استفاده شود تا بتوان ارزیابی دقیق‌تری انجام داد. برای تولید دادگان و متون بدون بایاس آن می‌توانید از مدل‌های زبانی بزرگ استفاده نمایید.

جدول ۱: نمونه‌ای از دیتاست مورد نظر

متن اصلی	متن هدف	دسته بایاس جنسیتی
امشب با رفقا یه جشن درست و حسابی دعوتیم.	امشب با رفقا یه جشن درست و حسابی دعوتیم.	بدون بایاس جنسیتی
به نظرم فقط زن‌ها می‌توانند یه بچه رو خوب تربیت کنند.	هرکسی می‌تواند یه بچه رو خوب تربیت کند.	بایاس مبتنی بر کلیشه
فقط زن‌ها در مواقع جنگ دلسوزی و پشتیبانی دریافت می‌کنند.	هرکسی در مواقع جنگ دلسوزی و پشتیبانی دریافت می‌کند.	بایاس مبتنی بر عقیده

لازم به ذکر است که تعیین نوع بایاس جنسیتی از اهداف تمرین نبوده و صرفا برای اینکه شهودی آورده شود که چه نوع متونی مدنظر هستند ذکر شدند. همچنین برای اینکه بتوان مراحل بعدی را با دقت بهتری انجام داد، سعی کنید که اندازه دادگان خود را حداقل به ۱۰۰ برسانید.

۲. بعد از جمع‌آوری دادگان، نوبت به ارزیابی مدل‌های زبانی در انجام این وظیفه می‌رسد. در ابتدا یک مدل زبانی کم حجم که به صورت مبتنی بر دستورالعمل آموزش داده شده باشد را انتخاب کنید (به عنوان مثال مدل لاما) و با دادن یک پرامپت مناسب از مدل بخواهید متن ورودی را از لحاظ بایاس جنسیتی خنثی سازی کند.

۳. اکنون دادگان جمع آوری شده را به دو قسمت آموزش و ارزیابی تقسیم کنید و مدل زبانی مرحله قبل را روی قسمت آموزش دادگان تنظیم کنید. برای این کار پیشنهاد می‌شود از روش‌های تنظیم دقیق بهینه پارامتر مانند LoRA استفاده کنید.

۴. درنهایت عملکرد مدل را در دو حالت خام و تنظیم دقیق به کمک معیارهای معمول ارزیابی تولید متن مانند BLEU و BERTScore بررسی کنید.