به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس بازیابی اطلاعات

تمرین ۴

آذر ماه ۱۴۰۲

*فهرست

مقدمه

در این تمرین، دانشجویان با یکی از وظایف موجود در پردازش متن به نام تحلیل احساسات آشنا می‌شوند و برای این منظور لازم است که به منظور آشنایی با چالش‌های احتمالی موجود در هنگام استفاده از مدل زبانی BERT از این مدل زبانی استفاده شود.

در بخش عملی، دانشجویان ابتدا با تئوری تحلیل احساسات و نقش آن در پردازش زبان‌های طبیعی آشنا میگردند.

برای این منظور لازم است که دانشجویان با استفاده از زبان برنامه‌نویسی پایتون و کتابخانه‌های پایتورچ یا تنسورفلو، داده‌های متنی را پردازش کرده و این وظیفه را انجام بدهند.

این تمرین به دانشجویان این امکان را می‌دهد تا مهارت‌های عملی در حوزه تحلیل احساسات با استفاده از یکی از مدل‌های زبانی، یعنیBERT ، را به دست آورند و با چگونگی عملکرد این مدل زبانی و معماری آن آشنا شوند.

همچنین، این تمرین به آنها این امکان را می‌دهد تا با مسائل و چالش‌های مرتبط با پردازش زبان‌های طبیعی و تحلیل احساسات آشنا شوند و نقدهای مثبت و منفی در متون را شناسایی و مدیریت کنند.

در قسمت تشریحی این تمرین، دانشجویان لازم است که به صورت مفهومی کاربرد مدل های زبانی مختلف را درک نمایند تا بتوانند از این مدل در وظایف دنیای واقعی استفاده نمایند.

مجموعه داده

دیتاست مورد استفاده مربوط به نظرات بیان شده توسط کاربران برای اپلیکیشن های موجود در گوگل پلی است که هر کدام از این نظرات دارای score هایی هستند و از این امتیازات میتوان در وظیفه تحلیل احساسات استفاده نمود.

دیتاست در پوشه Dataset قرار گرفته است.

سوال ۱ بخش عملی

در این بخش لازم است که وظیفه تحلیل احساسات بر روی دیتاست اشاره شده صورت بپذیرد. هدف از این وظیفه دسته‌بندی هر کدام از نظرات به یکی از سه دسته‌ی مثبت، منفی و خنثی است. برای این منظور گام های زیر صورت بپذیرد و گزارش هر بخش به صورت کامل ارائه گردد.

الف: در ابتدا پیش پردازشی بر روی دیتاست صورت بپذیرد، برای این منظور لازم است که ستون score به شکل categorical تبدیل گردد. برای این منظور تمامی نظرات با امتیاز کمتر مساوی ۲ به عنوان دسته‌ی منفی، امتیازات برابر ۳ با تگ خنثی و بیشتر از ۳ با تگ مثبت تبدیل گردد. پس از این کار با نمودار مناسب تعداد نظرات موجود در هر دسته را به تفکیک نمایش بدهید. آیا تعداد تگ های هر دسته متعادل است؟

ب: در این بخش لازم است که دیتاست به سه بخش تست، آموزش و اعتبارسنجی تفکیک گردد.

ج: مدل را برای وظیفه تحلیل احساسات در چندین ایپاک آموزش بدهید. لازم است که در این بخش به طور کامل نحوه ی آموزش مدل را در گزارش خود بیان نمایید. (این توضیحات باید شامل معماری مورد استفاده باشد.)

د: لازم است که نمایش دهید مدل آموزش داده شده دچار بیش برازش (overfit) نشده است، برای این منظور لازم است که از نمودار مناسب استفاده کنید.

ه: معیارهای f1-score و accuracy را برای داده‌های تست نمایش بدهید.

لازم به ذکر است فهم و درک مسئله و همین طور گزارش کامل در بخش‌های مختلف بخش اصلی نمرات این بخش را تشکیل میدهند و به تمرین‌هایی که صرفا پیاده‌سازی کد است و یا نتایج بدون توضیح تشریحی هستند، هیچ نمره‌ای تعلق نخواهد گرفت.

سوال ۲ بخش تشریحی (تحقیقی)

در سال‌های اخیر مدل های زبانی بزرگ مانند GPT3 با استقبال گسترده‌ای روبرو شده‌اند، زیرا این مدل‌ها دارای توانایی بالایی در تولید متن هستند و همین طور بازنمایی‌های تولید شده از این مدل‌ها در کاربردهای مختلفی قابل استفاده هستند. فرض کنید ما به دنبال پیاده‌سازی یک سیستم پرسش و پاسخ شخصی‌سازی شده بر بروی حجم زیادی از اطلاعات مانند چندین کتاب هستیم. لازم به ذکر است که این مدل زبانی فاقد اطلاعات موجود در کتاب‌های اشاره شده هستند(بر روی داده‌های این کتاب آموزش ندیده‌اند. اکنون به پرسش‌های زیر پاسخ بدهید. (لازم به ذکر است که ما توانایی آموزش مدل بر روی داده‌های مذکور را نداریم)

الف: چرا ما نمیتوانیم تمامی کتاب‌ها و هر سوال را به صورت یکجا به عنوان ورودی مدل زبانی بزرگ در نظر گرفته و پاسخ را دریافت کنیم؟

ب: راهکار پیشنهادی خود را در جهت حل مشکل قسمت الف بیان کنید؟

ج: معماری خود را در جهت پیاده‌سازی این سیستم پرسش پاسخ مطرح کنید. (این سیستم باید به صورت کامل بیان گردد به این صورت که کاربر سوال خود را از سیستم می‌پرسد و سیستم پاسخ مناسب کاربر را از بین متون کتاب بیان میکند.)

سوال امتیازی

تشخیص تقلب نوشتاری (plagiarism detection) یکی از تسک‌های چالشی در حوزه تحلیل داده‌های متنی است. هر نوع استفاده غیر مجاز از نوشته دیگران یا حتی خود فرد بدون ارجاع دهی مناسب تقلب نوشتاری در نظر گرفته می‌شود. در مساله تشخیص تقلب نوشتاری یک سند مظنون و مجموعه‌ای از اسناد منبع داریم و میخواهیم بخش‌هایی از سند مظنون را پیدا کنیم که از اسناد منبع گرفته شده‌اند.

مساله کشف تقلب نوشتاری معمولا در دو گام انجام میشود: (۱) بازیابی اسنادی که احتمال می‌رود منبع تقلب باشند و (۲) تشخیص بخش تقلب از سند مظنون و بخش متناظر آن از سند منبع.

با اسـتفاده از مطالبی که در بخش بازیابی عصـبی آموختید، راه حلی برای تشـخیص تقلب نوشـتاری ارائه دهید. تا حد ممکن سوال را دقیق و خلاصه پاسخ دهید.

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان IR_CA4_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.
- 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.
- تمرین تا یک هفته بعد از مهلت تعیین شده با جریمه تحویل گرفته می شود که جریمه تأخیر تحویل تمرین تا یک هفته به ازای هر روز ۱۵ درصد است.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارئه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با ایمیل زیر در ارتباط باشید:

mailto:mhmssadeghi74@gmail.com

مهلت تحویل بدون جریمه: ۲۴ آذر ماه ۱۴۰۲

مهلت تحویل با تأخیر، با جریمه ۱۵ درصد : ۱ دی ماه ۱۴۰۲