به نام خدا

فاز نهایی پروژه درس پردازش زبان و گفتار

دکتر بهروز مینایی

تابستان ۱۴۰۱

تیم حل تمرین:

محمد یارمقدم

امیرحسین امینی مهر

ثمین حیدریان

مهسا انوریان

توحید عابدینی

رضا قهرمانی

هدف این بخش از پروژه آشنایی و کسب تجربه شما در زمینه تحلیل، استخراج ویژگی‌ها و رده‌بندی میباشد.

ساختار پوشه‌ها و فایل‌های مورد نیاز:

- data: داده‌های جمع‌آوری شده و تمیز شده از مرحله اول پروژه بعلاوه پردازش‌های جدید این مرحله.
- src: تمام کدهای نوشته شده برای پروژه بدون استثناء.
- reports: تمام کدهای نوشته شده برای پروژه بدون استثناء.
- Models: کلیه مدلها آموزش داده شده.

بخش اول: تولید جملات

در این بخش مدل ParsBERT را روی هر کدام از دسته داده‌های خود به صورت جداگانه fine_tune کنید و در قالب مدلی به نام "label"_parsBERT.berm_lm ذخیره کنید. سپس از این مدل‌ها برای تولید جملات استفاده کنید.

برای این بخش می‌توانبد از لینک‌های زیر استفاده کنید:

https://huggingface.co/HooshvareLab/bert-fa-base-uncased

https://huggingface.co/docs/transformers/training

بخش دوم:

برای این بخش داده‌های خود را به دو بخش train/test با درصدهای ۸۰/۲۰ بصورت متوازن برای هر برچسب تقسیم کنید.

در این قسمت مطابق با نوع تسک خود که می‌تواند اعم از دسته‌بندی، ترجمه، تولید متن و ... باشد می‌بایست با استفاده از یک معماری ساده که نوع آن می‌تواند دلخواه باشد، مدل خود را آموزش داده و نتایج بدست‌آمده از آن را به طور کامل در داکیومنت خود شرح دهید.

بخش سوم:

در این بخش شما می‌بایست معماری ساده‌ای که در قسمت قبل استفاده کردید و نتایج آن را در تسک خود مشاهده کردید را بهبود داده و آن را با استفاده مدل های پیچیده‌تری پیاده سازی کنید. برای این کار توصیه می‌شود از مدل‌های به روزی مانند transformers استفاده کنید.

دقت کنید میزان بهبود و نوع نگاه شما به داده‌ها و انتخاب مدل و معماری مناسب بخش مهمی از نمره شما را دربر خواهد داشت.

در انتها نتایج این قسمت را با قسمت قبل به صورت کامل بررسی و مقایسه کنید.

موفق باشید