به نام خدا

درس پردازش زبان طبیعی

دکتر مرضیه داود آبادی

تمرین سری سوم

طراحان تمرین:

محمود فاضلی

مهلت تحویل:

۱۴۰۴/۰۲/۱۳

نکات تکمیلی

۱. پاسخ سوالات را به صورت کامل در یک فایل PDF و به همراه کدهای سوالات در فرمت ipynb. در یک فایل فشرده به شکل StudentNumber_FirstName_LastName_HW3.zip قرار داده و تا زمان تعیین شده بارگذاری نمایید.

۲. برای پیاده سازی ها زبان پایتون پیشنهاد می شود، لازم به ذکر است توضیح کد ها و نتایج بدست آمده، باید در فایل PDF آورده شوند و به کد بدون گزارش نمره ای تعلق نخواهد گرفت.

۳. به ازای هر روز تاخیر ۵۰ درصد از نمره تمرین کسر خواهد شد.

۴. لطفا برای انجام تمرین زمان مناسب اختصاص داده شود و انجام آن را به روزهای پایانی موکول نکنید.

۵. بد نیست منابع استفاده شده در حل هر سوال را ذکر کنید.

۶. خلاقیت نمره اضافی دارد

موفق باشید

سوال ۱

تا به اکنون شما با ترانسفورمرها و نیز تسک طبقه بندی متون مثل sentiment analysis آشنا شده‌اید. در این تمرین قصد داریم تا شما را با fine-tune کردن یک مدل BERT از پیش آموزش دیده شده برای تسک sentiment analysis با استفاده از transformer PyTorch Trainer مربوط به Hugging Face آشنا کنیم. لطفاً به نوت بوک این سوال یعنی فایل (Assignment 1.ipynb) مراجعه نمایید و قسمتهای خواسته شده را تکمیل فرمایید. تمامی توضیحات مربوط به هر بخش در اختیار شما قرار گرفته است. (توجه! بهتر است در گوگل کولب اجرا کنید.)

در گام اول بایستی دیتاست را از این بخش دانلود کنید (از دیتاست amazon_cells_labelled.txt استفاده کنید):

https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

برای بخش آموزش و ست کردن پارامترهای مدل، پارامترهای مختلفی را تست و ارزیابی کنید تا به یک مینیمم محلی بهینه دست یابید. (البته دقت مدل با پارامترهای پیشفرض بالای ۹۰ درصد خواهد بود)

*** به عنوان کار تشویقی و امتیازی در این قسمت می توانید استراتژیهای مختلفی را برای hyperparameter tuning استفاده کنید. (مثل grid search, optuna، روش های بیزین و ...)**

بعد از آموزش مدل، نوبت می رسد به ارزیابی عملکرد مدل. بدین منظور لطفاً متریک های مختلفی مثل دقت، فراخوانی، f1 score، AUC و ... را محاسبه و در گزارش خود ذکر کنید. راستی، از کجا میفهمید که مدل شما overfit یا underfit نشده است؟

***کار امتیازی: در این تمرین ما از tokenizer و مدل sequence classifier برت (BERT) برای قسمت های مختلف تسک استفاده کردیم. شما می توانید از بین لیست tokenizer ها و مدل های موجود در hugging face، موارد مختلفی را تست کنید و نتایج آن ها را با هم مقایسه کنید.**

سوال ۲

در این سوال قصد داریم تا شما را با تسک word analogy آشنا کنیم. بدین منظور:

۱. ابتدا بایستی از یکی از مدل ها یا همان word vector های از قبل آموزش دیده شده در hugging face استفاده کنید. (مدل پیشنهادی: https://huggingface.co/fse/word2vec-google-news-300/tree/main)

۲. با استفاده از معیار شباهت کسینوسی، شباهت ها (similarity) را محاسبه کنید.

۳. از word embedding های بدست آمده برای حل مسائلی مثل:

Man is to Woman as King is to _____ (answer: queen)

استفاده کنید.

۴. (اختیاری) word embedding ها را مدیفای کنید تا باياس embeddingها را کاهش دهید.

به نوت بوک سوال (Assignment 2.ipynb) مراجعه کنید و قسمت های خواسته شده را تکمیل کنید. برای هر بخش، نکات، درسنامه و راهنمایی های مرتبط قرار گرفته است. ضمناً بهتر است کد را در گوگل کولب اجرا کنید.