تمرین چهارم پردازش زبان طبیعی

توجه:

- **کد باید فقط در زبان پایتون باشد.**
- **استفاده از کتابخانه‌های از پیش آماده مجاز است. در گزارش کتابخانه ها ذکر شوند.**
- **لطفاً علاوه بر ضمیمه کردن کد، نتایج را تحلیل و در فایل گزارش خود ضمیمه کنید.**
- **فایل گزارش به فرمت pdf و به زبان فارسی باشد.**

فایل مجموعه داده IMDB، برای مسئله تشخیص احساسات [1]، را از این لینک دانلود کنید. مجموعه داده IMDB به چهار بخش داده، بدون برچسب، آموزشی، اعتبار سنجی و تست تقسیم شده است. برچسب هر نمونه عدد صفر یا یک است، که به ترتیب نشانگر احساسات منفی و مثبت اند. در این مسئله هدف این است که با استفاده از داده بدون برچسب (unsupervised) نمایش مناسبی بدست بیاورید که در تشخیص احساسات به کار ببرید.

۱. پس پیش پردازش های لازم را بر روی متن خام انجام دهید. (پیش پردازش حداقل باید شامل کوچک کردن حروف انگلیسی و حذف علائم و اعداد باشد.)

۲. با استفاده از الگوریتم BPE متن را واحد سازی کنید (آموزش واحدساز را بر روی داده بدون برچسب انجام دهید.)

۳. کدگذار مبدل را بر روی داده آموزشی آموزش دهید. ابرپارامتر های مختلفی را برای تعداد لایه ها، تعداد سر های ماژول خود_توجه و اندازه کدگذاری ماژول خود_توجه مورد بررسی قرار دهید و تاثیر هر کدام را بر روی دقت داده آموزشی و ارزیابی بررسی کنید.

۴. مدل بهینه را انتخاب و گزارش کنید و بر روی داده تست پیشبینی را انجام دهید و آن را در فرمت مشابه "y_train" با نام "y_test.txt" ذخیره و ضمیمه کنید.

۵. مدل برت را با استفاده از تکنیک self-supervised learning بر روی مجموعه داده بدون برچسب آموزش دهید (جملات را استخراج کنید و MLM را بر روی آن پیاده سازی کنید). تعداد سر های ماژول خود_توجه را ۴، تعداد لایه ها را ۴ و اندازه کدگذاری را ۱۲۸ در نظر بگیرید.

۶. اکنون با استفاده از شبکه پیش آموزش دیده به عنوان استخراج کننده ویژگی، مدل را آموزش دهید (شبکه پیش آموزش دیده در فرآیند آموزشی ثابت است). دقت را بر روی داده آموزش و اعتبار سنجی گزارش کنید. نتایج بر روی داده تست را در فایلی با نام "y_test2.txt" ذخیره کنید.

۷. اکنون شبکه بخش قبل را مجدد آموزش دهید، اما در این قسمت شبکه پیش آموزش دیده هم آموزش دهید (نرخ یادگیری کوچک بگذارید). دقت را بر روی داده آموزش و اعتبار سنجی گزارش کنید. نتایج بر روی داده تست را در فایلی با نام "y_test3.txt" ذخیره کنید.

[1] Sentiment Analysis