تمرین سوم درس پردازش زبان و گفتار

استاد درس: دکتر حمیدرضا برادران کاشانی

دستیاران آموزشی: آیین کوپایی- هاجر مظاهری

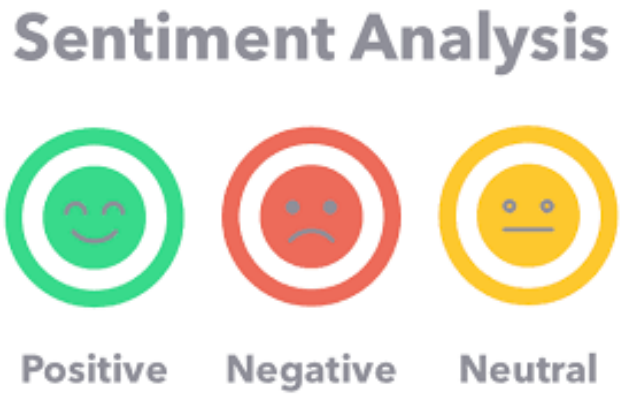
تاریخ بارگذاری تمرین: ۱۴۰۳/۰۲/۲۷

تاریخ تحویل تمرین: ۱۴۰۳/۰۳/۱۳

دانشگاه اصفهان

تحلیل احساسات

تجزیه و تحلیل احساسات فرآیندی است برای شناسایی و استخراج احساسات از درون زبان نوشتاری. تحلیل احساسات می‌تواند به عنوان یک کار طبقه‌بندی متن با هدف برچسب گذاری قطبیت یک متن (مثبت، منفی یا خنثی) مدل سازی شود. به کمک تحلیل احساسات کمپانی‌های مختلف قادر به رصد کردن نظر مشتریان در مورد محصولات و خدمات خود هستند. هدف از تمرین حاضر پیاده سازی یک سیستم تحلیل احساسات است. مجموعه داده مورد استفاده در این تمرین مجموعه داده sentiment140 است که از 1,600,000 توییت تشکیل شده است. این توییت‌ها به دو دسته مثبت و منفی تقسیم می‌شوند. در این تمرین با دو ستون text و sentiment کار می‌کنیم. احساس مثبت در این مجموعه داده با مقدار ۴ و احساس منفی با صفر نشان داده شده است. (مجموعه داده مورد نظر با عنوان sentiment140.csv در پوشه تمرین قرار گرفته است).



۱- پیش پردازش:

مراحل پیش پردازش زیر را **به ترتیب** انجام دهید.

۱-۱- برچسب توییت‌های مثبت را به عدد ۱ و برچسب توییت‌های منفی را به عدد صفر تبدیل کنید.

۱-۲- URL ها را با توکن URL ، منشن‌ها را با توکن MENTION و هشتگ‌ها را با توکن HASHTAG جایگزین کنید.

۱-۳- علائم نگارشی را حذف کنید.

۱-۴- هر توییت را به کلمات آن توکن‌بندی کنید.

۱-۵- عمل لم‌سازی را انجام دهید.

۱-۶- ۲۰ توییت اول را نمایش دهید.

۱-۷- ۸۰ درصد از مجموعه داده را برای آموزش و ۲۰ درصد باقیمانده را برای تست در نظر بگیرید.

۲- بردار سازی متن:

۲-۱- هدف از این بخش این است که به هر یک از کلمات منحصر به فرد یک عدد منحصر به فرد اختصاص دهیم و سپس آن کلمه را با عدد اختصاص داده شده جایگزین کنیم.

۲-۲- از آنجایی که برای پردازش نیازمند داده‌هایی با بعد یکسان هستیم، لذا بر روی دیتاست متد pad_sequences را هم اعمال کنید.

۳- تعبیه کلمات[1]:

در این تمرین تعبیه کلمات با استفاده از روش Word2Vec را پیاده سازی می‌کنیم. تعبیه کلمه یک بازنمایی آموخته شده برای متن است که در آن کلماتی که معنی یکسانی دارند بازنمایی مشابهی دارند. Word2Vec یک رویکرد محبوب است که از شبکه های عصبی برای یادگیری این تعبیه کلمات استفاده می کند. با استفاده از کتابخانه gensim از مدل Word2Vec برای تعبیه کلمات استفاده کنید. برای دانلود این مدل قطعه کد زیر را اجرا کنید.

```
# load google news word2vec
import gensim.downloader as api

w2v = api.load('word2vec-google-news-300')
```

۴- ساخت مدل دسته‌بند:

۴-۱- از مدل RNN با ساختار زیر استفاده کنید و مدل را مطابق با تنظیمات داده شده در جدول ۱ آموزش دهید.

Model Architecture:

- **Embedding layer**
- **RNN (Recurrent Neural Network)**

[1] Word Embedding

- **Dense layer :** with a Linear activation function

جدول ۱: تنظیمات اجرا

Optimizer	Criterion	Batch Size	Epoch
Adam	Cross-entropy	512	5

۴-۲- مدل آموزش دیده را بر روی داده تست آزمایش کنید و مقدار loss و accuracy را گزارش کنید. سپس با استفاده از کتابخانه scikit-learn و تابع classification_report مقادیر precision، recall و f1-score را گزارش دهید.

نکات تحویل

۱- پاسخ خود را در پوشه ای به اسم NLP_NAME_FAMILY_HW3 و در قالب zip بارگذاری نمایید.

۲- این پوشه باید حاوی موارد زیر باشد:

- کد نوشته شده در قالب یک فایل jupyter notebook
- فایل گزارش فنی در قالب یک فایل PDF

۳- لازم به ذکر است که رعایت قوانین نگارشی حائز اهمیت است.