
 <p>دانشگاه صنعتی امیرکبیر</p>	<p>دوره آموزشی پردازش زبان طبیعی (NLP)</p> <p>تمرین ششم</p> <p>مهلت تحویل: شنبه ۳ مهر ۱۴۰۰</p>	 <p>آکادمی همراه</p>
-----------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------

مقدمه:

در این تمرین یک چت بات ساده را در چند حالت پیاده سازی خواهیم کرد.

داده مورد نظر دیتاست دیلی دیالوگ Daily dialogue هست. این دیتاست در نت به نسخه‌های مختلف موجود است. که برای این تمرین می‌توانید از داده آموزش، ولیدیشن و تست به صورت زیر استفاده کنید:

```
!pip install datasets
from datasets import load_dataset
dataset = load_dataset('daily_dialog')
dataset
```

این دیتاست مجموعه از دیالوگ‌های روزمره است. اگر یک دیالوگ دلخواه را از دیتاست را انتخاب کنیم، شخص a جملات زوج شامل (۰ و ۲ و...) و شخص b جملات فرد (۱ و ۳ و...) این دیالوگ را گفته است.

هدف طراحی شبکه‌ای است که بدون در نظر گرفتن تاریخچه دیالوگ به یک turn مناسب‌ترین پاسخ را بدهد. از آنجایی که ترکیب پیش‌پردازش و مدل‌های مختلف دقت‌های متفاوتی دارند، معیار پیاده‌سازی خود را ویدئو آموزشی قرار دهید. همچنین پیشنهاد می‌شود برای هر turn مقدار MAX_LENGTH را معادل ۴۰ قرار دهید (با توجه به اینکه دیتاست به اندازه کافی بزرگ هست حذف نمونه‌های جملات بزرگ تأثیر منفی مشهودی نخواهد داشت).

موارد پیاده‌سازی برای تمرین:

۱- یک مدل ترنسفورمر با حداقل دولایه برای encoder و دولایه برای decoder پیاده‌سازی کنید. در لایه embedding از بازنمایی word2vec استفاده کنید و حالت trainable لایه بازنمایی را در به دو صورت فعال و غیرفعال تست کنید. مدل‌ها را با یکدیگر مقایسه کنید و نظر خود را در مورد مقایسه بنویسید. در سایر قسمت‌ها منظور از ترنسفورمر بهترین مدل این بخش هست.

۲- یک مکالمه ایجاد کنید که turn اول آن جمله زیر باشد و ۱۰ turn آن را با مدل بهترین مدل سوال ۱ ادامه دهید.

"What do you do for your weekend?"

۳- سؤال یک را با یک مدل از خانواده Seq2Seq مانند LSTM یا هر مدل دیگر پیاده‌سازی کنید و نتایج را به‌دست آورید.

۴- برای سوالات زیر جواب بهترین مدل سوال ۳ را چاپ کنید.

```
"Where have you been?"  
"What is your name?"  
"Where are you from?"  
"What do you want to drink?"
```

۵- برای هردو مدل ترنسفورمر و RNN دو نمودار خطای آموزش و ولیدیشن را رسم کنید. دقت مدل‌ها را بر اساس accuracy و زمان آموزش مقایسه کنید.

نکات عمومی تمرین برای تصحیح سریع‌تر:

نکته ۱: در بالاترین سلول از نوت‌بوک یک متغیر usePretrained استفاده کنید و مقدار آن موقع تحویل True باشد. برای بندهای ۲ و ۴ تمرین باید بدون آموزش مدل‌ها کار کنند و دسترسی خاصی به درایو شما هم نیاز نباشد.

نکته ۲: متغیر دیگری در همین سلول باشد به نام DriveAccess و مقدار آن هنگام تحویل False باشد. هدف این متغیر حذف قسمت‌های هست که ذخیره مدل انجام می‌شود.

نکته ۳: آدرسی که فایل‌های خود را ذخیره می‌کنید هم به اینصورت باشد.

```
'/content/drive/MyDrive/HW-06/{Google-userName}-*. * '
```

مثلا برای academy@gmail.com و ذخیره مدل:

```
'/content/drive/MyDrive/HW-06/academy-model-best.h5 '
```