تمرین شماره4 و پروژه نهایی

NLP

1- ابتدا در سایت apify.com ثبت نام کنید. وارد پنل کنسول خود شوید از قسمت store اکتور Instagram comment scrapper را انتخاب کنید، سپس کار کردن با آن را امتحان کنید.
یک موضوع تحقیقاتی انتخاب کنید(برای مثال پست‌های مرتبط با فیلم oppenheimer). و نظرات کاربران را در مورد این موضوع جمع آوری کنید. تعدادی پست مرتبط برای این موضوع انتخاب و آدرس پست هارا در قسمت urls وارد کنید. وارد بخش schedule شوید و برای این اکتور یک زمان بندی دلخواه درست کنید. سعی کنید در نهایت حداقل بین 1000 تا 5000 دیتا غیر تکراری جمع آوری کنید.
-استفاده از api این سایت برای جمع آوری داده‌ها و مدیریت آنها نمره اضافه دارد.

2- -مراحل پیش پردازش را بر روی داده های جمع آوری شده در تمرین 1 را انجام دهید سپس مدل های زبانی unigram, bigram, three gram را برای این داده‌ها پیاده سازی کنید. برای هر کدام مراحل smoothing را انجام دهید و در گزارش کار ذکر کنید که استفاده از smoothing چه فایده‌ای دارد.
-تابعی برای محاسبه perplexity بنویسید.
-تابعی برای ساخت جمله با طول دلخواه برای هر 3 مدل زبانی بنویسید. (راهنمایی: برای این کار نیاز به یک تابع دارید که بر حسب کلمه‌ای که وارد آن می‌شود، محتمل‌ترین کلمه بعدی را برگرداند. مثال: برای مدل Bigram تابع یک کلمه میگیرد و کلمه بعدی را برحسب این کلمه حدس میزند.)
- به وسیله هر کدام از مدل‌ها 5 جمله جنریت کنید، perplexity هرکدام از جملات را بدست آورید و با یکدیگر مقایسه کنید و گزارش دهید.

Unigram:

$$P(w_1, w_2, \dots, w_n) \approx \prod_i w_i$$

Bigram:

$$P(w_n | w_1, w_2, \dots, w_{n-1}) \approx \prod_{i=1}^{n} P(w_i | w_{i-1})$$

Trigram:

$$P(w_n | w_1, w_2, \dots, w_{n-1}) \approx \prod_{i=1}^{n} P(w_i | w_{i-2}\, w_{i-1})$$

استفاده از chatgpt برای تمرین2 مجاز است.

3- مراحل پیش پردازش را برای دیتاهای سکسیسم انجام دهید، داده ها را به 3 قسمت train, test, validation تقسیم کنید و در نهایت یک مدل LSTM برای دسته بندی آنها به دو گروه Sexism و nonSexism بسازید. و در گزارش کار خود حتما ذکر کنید که هر لایه چه وظایفی انجام می‌دهند.

9.Creating a simple LSTM Model

```
[ ]  from keras.models import Sequential
     from keras.layers import Embedding, LSTM, Dense

     # Build the neural network model
     model = Sequential()
     model.add(Embedding(10000, 20))
     model.add(LSTM(32))
     model.add(Dense(1, activation='sigmoid'))
     model.compile(optimizer='rmsprop', loss='binary_crossentropy', metrics=['acc'])

     # Train the model
     LSTMhistory = model.fit(X_train, y_train, epochs=10, batch_size=64, validation_data=(X_val, y_val))
```

- **در مدل میتوانید تغییراتی به وجود آورید. افرادی که بیشترین دقت را به دست آورده باشند نمره اضافه خواهند گرفت.**

برای دسترسی به داده های sexism مراحل زیر را انجام دهید

داده هارا از آدرس گیت https://github.com/rewire-online/edos.git کلون کنید. و فایل edos_labelled_aggregated.csv را با دستور pd.read_csv بخوانید. از این فایل فقط ستون های text , label_sexist و split را نیاز دارید.

موفق باشید.