هدف تمرین:

- آشنایی با ساخت مدل زبانی (Language model)
- آشنایی با ساخت مدل زبانی (Language model) در NLTK
- آشنایی با محاسبه Perplexity برای مدل‌های زبانی ایجاد شده در NLTK

مراحل:

۱. در این تمرین شما با فایل متنی Shakespeare.txt کار می‌کنید. نسخه NLTK استفاده شده در این تمرین 3.5 است.

۲. در ابتدا میخواهیم مراحل پیش‌پردازش را روی پیکره‌ی داده شده اعمال کنیم.

a. فایل پیکره داده شده را باز کنید.

b. متن مذکور را به جملات آن تجزیه کنید.

c. هر جمله را به کلمات آن تجزیه کنید به صورتی که فقط اعداد و کلمات بمانند و ابتدا و انتهای هر جمله توکن‌های <s> و </s> را اضافه کنید.

d. تعداد tokenها و typeهای پیکره را بدست آورید و نمایش دهید.

e. در ادامه کلمات پر تکرارِ جملاتی که از قسمت c بدست آمده است را حذف کنید و در یک متغیر جدید ذخیره کنید. می‌خواهیم یک بار با استفاده از پیکره با stopwords و یک بار با استفاده از پیکره بدون stopwords مدل‌های زبانی را ایجاد کنیم.

۳. در ادامه، میخواهیم n-gramهای متن را ایجاد کنیم.

a. با کمک کتابخانه‌ی NLTK، Unigram، Bigram، Trigram و Quadrigramهای مربوط به متن را استخراج کنید. این کار را یک بار برای متن با stopwords و یک بار برای متن بدون stopword انجام دهید.

۴. با کمک کتابخانه NLTK، تعداد تکرار هر n-gram به خصوص در متن را بدست آورید و ۵ n-gram پر تکرار از هرکدام از n-gramهای ایجاد شده در قسمت ۳ را نمایش دهید. این کار را یک بار برای متن با stopwords و یک بار برای متن بدون stopword انجام دهید.

۵. در این بخش می‌خواهیم مدل‌های زبانی را با در نظر گرفتن add-one smoothing ایجاد کنیم.

a. با کمک n-gramهای ایجاد شده و فرمول‌های ارائه شده در درس، مدل‌های زبانی را (به صورت دستی) محاسبه کنید. خروجی این قسمت باید مدل‌های زبانی Unigram، Bigram، Trigram و Quadrigram باشد. **(راهنمایی: برای هر n-gram، احتمال آنرا با استفاده از فرمول ارائه شده در درس محاسبه کنید.)**

b. این کار را یک بار برای متن با stopwords و یک بار برای متن بدون stopword انجام دهید.

۶. حال می‌خواهیم با استفاده از مدل‌های ایجاد شده، کلمات جدید را با استفاده از دنباله‌ای از کلمات ایجاد شده پیش بینی کنیم. برای این کار تابعی طراحی کنید که مدل و دنباله کلمات را ورودی بگیرد و کلمه جدید را به عنوان خروجی برگرداند. کلمات بعد از دنباله کلمات زیر را در خروجی با استفاده از مدل‌های ایجاد شده، نمایش دهید.

a. این کار را یک بار برای متن با stopwords و یک بار برای متن بدون stopword انجام دهید.

b. به ازای هر مدل زبانی در جدول زیر، دو نمونه دنباله کلمه مثال بزنید که برای مدل زبانی ایجاد شده با stopwords و مدل زبانی بدون stopwords نتیجه متفاوت داشته باشد و این دو نوع مدل زبانی را با یکدیگر به صورت مختصر مقایسه کنید.

مدل زبانی	دنباله کلمات
Bigram	William, Tattered
Trigram	(ragged, hand), (sweetly, chide)
Quadrigram	(beguile, the, world), (in, thy, noon)

۷. تابعی طراحی کنید که با گرفتن مدل زبانی و یک طول به صورت عدد، با استفاده از مدل گرفته شده جمله‌ای به طول گفته شده ایجاد کند. به ازای Unigram، Bigram، Trigram و Quadrigram جملاتی با طول ۲۰ ایجاد کنید. **(راهنمایی: از متد ایجاد شده در قسمت ٦ کمک بگیرید.)**

۸. حال می‌خواهیم با استفاده از کتابخانه NLTK به ایجاد مدل زبانی بپردازیم.

a. بررسی کنید چگونه می‌توان در کتابخانه NLTK مدل زبانی آموزش داد.

b. برای Unigram، Bigram، Trigram و Quadrigram مدل‌های زبانی با NLTK ایجاد کنید. (این کار را صرفا برای n-gramهای با stopwords انجام دهید.)

c. بررسی کنید چگونه می‌توان با استفاده مدل‌های زبانی ایجاد شده، جملات با طول دلخواه ایجاد کرد. به ازای مدل‌های ایجاد شده جملات با طول ۲۰ ایجاد کنید و در خروجی چاپ کنید.

d. بررسی کنید چگونه می‌توان perplexity مدل‌های ایجاد شده را برای یک جمله دلخواه محاسبه کرد و perplexity جمله‌ی Then let not winter's ragged hand deface را برای مدل زبانی unigram حساب کنید.

- نتیجه نهایی را در قالب یک فایل Jupyter notebook و یک گزارش PDF در یک فایل فشرده zip که نام آن با فرمت NLP_Name_Family_Ex3 (به جای Name اسم و به جای Family نام‌خانوادگی خودتان را قرار دهید) است قراردهید و به ایمیل abedi.a1997@gmail.com ارسال کنید.

موفق باشید