



الف Stemming

ب Lemmatization

ج Stopword Removal

**سوال ۳)** در سوال ۲ جملات اصلی ذکر شده در سوال ۱ به عنوان ورودی هر یک از بخش‌ها استفاده شد. در این سوال همه مراحل پیش‌پردازش سوال ۲ را تکرار کنید با این تفاوت که بعد از اعمال Tokenization و Normalization جملات را به عنوان ورودی مورد استفاده قرار دهید.

نکته ۴: توجه کنید که ممکن است بعضی از کتابخانه‌ها تمام قابلیت‌های فوق را نداشته باشند. مواردی که در یک کتابخانه خاص موجود نیست را خالی بگذارید.

**سوال ۴)** با توجه به خروجی‌های سوال ۱ تا ۳ جمع‌بندی کنید که کیفیت پیش‌پردازش کدام کتابخانه را بهتر می‌دانید و ترجیح می‌دهید از آن استفاده کنید.

**سوال ۵)** با استفاده از داده‌های train و test تمرین هفته سوم یک دیکشنری از تمام کلمات با تکرار بالای ۳ کلمه بسازید. سپس برای هر یک از کلمات خطای زیر فاصله لונشتاین را با تمام کلمات دیکشنری محاسبه کنید و پنج کلمه اول با کمترین فاصله را همراه با مقدار فاصله چاپ کنید. (هزینه insert و delete را ۱ و هزینه substitute را ۲ در نظر بگیرید.)

۱. اختصار

۲. سادرات

۳. فوتکال

۴. مسابغات

۵. وازدات

۶. مشرکت

۷. کشورور

۸. منجلسه

نکته ۵: مثالی از خروجی این سوال در ادامه آورده شده است. فرض کنید کلمه اصلی 'شخصیت' باشد که به اشتباه 'شخصت' نوشته شده باشد. (یعنی دارای خطای delete باشد). پنج کلمه اول با کمترین فاصله برای کلمه 'شخصت'، همراه با مقدار فاصله آن‌ها نوشته شده است.

شخصت: ('شخصیت', ۱), ('شخص', ۱), ('شخصت', ۱), ('شخصی', ۲), ('مشخص', ۲)

**سوال ۶)** با استفاده از داده train یک مدل زبانی بایگرم آموزش دهید. حال فرض کنید برای هر یک از کلمات خطای سوال ۵ می‌دانیم که آن واژه در چه متنی رخ داده‌است. با توجه به عبارت متنی موجود برای هر یک از ۵ کلمه‌ای که به‌عنوان کلمات صحیح هر واژه غلط در سوال ۴ مشخص کردید، محاسبه کنید کدام یک احتمال بایگرم بیشتری برای حضور در عبارت مشخص شده برای هر کلمه دارد و آن را انتخاب نمایید. عبارت مربوط به هر کلمه کنار آن نوشته شده‌است. برای هر عبارت، کلمه منتخب و احتمال آن را در خروجی چاپ کنید.

۱. اقتصاد : رشد اقتصاد و تحرک زندگی اجتماعی

۲. صادرات : حجم صادرات ایران

۳. فوتکال : فدراسیون فوتکال کشور

۴. مسابغات : در جریان انعکاس مسابغات صبح

۵. وازدات : اقلام عمده وازدات کشور

۶. مشرکت : اصل مشرکت مردمی

۷. کشور : وزارت کشور جمهوری اسلامی ایران

۸. منجلسه : جلسه علنی دیروز منجلسه شورای اسلامی

نکته ۶: کلمه 'شخصت' مثال سوال قبل را همراه با عبارت زیر به عنوان ورودی این سوال در نظر بگیرید.

شخصت: گوشه دیگری از شخصت این بانوی گرانقدر

از بین ۵ کلمه‌ای که به‌عنوان کلمات صحیح برای کلمه 'شخصت' با استفاده از روش لונشتاین به دست آمده، با استفاده از مدل بایگرم احتمال بیشتری برای حضور کلمه 'شخصیت' در عبارت 'گوشه دیگری از شخصت این بانوی گرانقدر' به جای واژه دارای خطای 'شخصت' به دست می‌آید. بنابراین کلمه 'شخصیت' می‌تواند به عنوان بهترین کلمه صحیح از بین ۵ واژه سوال قبل برای کلمه دارای خطا 'شخصت' باشد. خروجی سوال برای این ورودی به صورت زیر خواهد بود. در کنار کلمه منتخب، احتمال آن نوشته شده‌است.

شخصت: 'شخصیت',  $5.9132 \times e^{-18}$

نکته ۷: توجه کنید که حرف 'ی' و 'ک' را برای کلمات دارای این حرف در کولب به صورت عربی تایپ کنید تا برای پیدا کردن کلمات در دیکشنری و همچنین محاسبه احتمال دچار مشکل نشوید. زیرا حرف 'ی' و 'ک' در مجموعه داده به صورت عربی تایپ شده‌است.