

دوره آموزشی پردازش زبان طبیعی (NLP)

تمرین چهارم

مهلت تحویل:

۶ شهریورماه ۱۴۰۰



سوال ۱) جملات زیر را به عنوان دادههای ورودی در نظر بگیرید و با سه کتابخانه مختلف پیشپردازش متن فارسی، این جملات را Tokenize و Normalize کنید هر کتابخانه کدام مورد را صحیح پیشپردازش کرده و برای کدام موارد اشتباه کردهاست.

جمله ۱: انتخابات هفته آینده رادر تهران برگزار میکنیم ، اما دسته بندی حوزهها برای رأی گیری از این هفته آغاز شده است

جمله ۲: توانتشار خبرمهم در رسانه ها را تایید کردی

جمله ۳: پسردکتراحمدی درکارنامه حرفه ای خود ۱۸ سریال تلویزیونی و ۱۲فیلم سینمایی دارد

جمله ۴: نقّاشی ساختمان به پایان رسیدددددد

جمله ۵: اللهٔ اکبر گویان وارد مسجد شد

جمله ۶: دکتر جان مممممممممممممنونم که مشکلات حرکتی فرزندم را درمان کردید

نکته ۱: برای کتابخانهها می توانید از دو کتابخانه hazm و parsivar استفاده کنید و کتابخانه سوم را با جستجوی خود پیدا و استفاده کنید.

نکته ۲: توجه کنید که بهعنوان مثال از یک Tokenizer خوب انتظار داریم در جملات بالا موارد زیر را اصلاح کند:

رادر – را در

دسته بندی - دستهبندی

نکته ۳: توجه کنید که بهعنوان مثال از یک Normalizer خوب انتظار داریم در جملات موارد زیر را اصلاح کند:

نقّاشي - نقاشي

رسیددددددد - رسید

سوال ۲) برای تمام شش جمله ورودی سوال ۱ مراحل پیش پردازش زیر را نیز با استفاده از هر کتابخانه انجام دهید و خروجیهای آنها را نیز بررسی و مقایسه کنید.

الف) Stemming

ب) Lemmatization

Stopword Removal (7

سوال ۳) در سوال ۲ جملات اصلی ذکر شده در سوال ۱ بهعنوان ورودی هر یک از بخشها استفاده شد. در این سوال همه مراحل پیش پردازش سوال ۲ را تکرار کنید با این تفاوت که بعد از اعمال Tokenization و Normalization جملات را بهعنوان 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 پیش پردازش کدام کتابخانه را بهتر میدانید و ترجیح می دهید از آن استفاده کنید.

سوال 0) با استفاده از دادههای train و test تمرین هفته سوم یک دیکشنری از تمام کلمات با تکرار بالای 7 کلمه بسازید. سپس برای هر یک از کلمات خطای زیر فاصله لونشتاین 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 7 در نظر بگیرید.)

۱. اختصاد

۲. سادرات

٣. فوتكال

۴. مسابغاط

۵. واز دات

۶. مشرکت

۷. کشوور

٨. منجلسه

نکته ۵: مثالی از خروجی این سوال در ادامه آورده شدهاست. 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), (اشصت', ۱), (اشخصی', ۲)
```

سوال ۴) با استفاده از داده train یک مدل زبانی بایگرم آموزش دهید. حال فرض کنید برای هریک از کلمات خطای سوال ۵ می دانیم که آن واژه در چه متنی رخ داده است. با توجه به عبارت متنی موجود برای هر یک از ۵ کلمه ای که به عنوان 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 شده است. برای هر عبارت، کلمه منتخب و احتمال آن را در خروجی چاپ کنید.

۱. اختصاد: رشد اختصاد و تحرک زندگی اجتماعی

۲. سادرات : حجم سادرات ایران

٣. فوتكال : فدراسيون فوتكال كشور

۴. مسابغاط : در جریان انعکاس مسابغاط صبح

۵. وازدات : اقلام عمده وازدات كشور

۶. مشرکت: اصل مشرکت مردمی

۷. کشوور : وزارت کشوور جمهوری اسلامی ایران

٨. منجلسه : جلسه علني ديروز منجلسه شوراي اسلامي

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 لونشتاین 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.

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

نکته V: توجه کنید که حرف 2 و 2 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 2 و 2 در مجموعه داده به صورت عربی تایپ شده است.