

1402-02-08

Natural Language Processing

Dr. Mohammad Bidoki

Homework #2

By: Mohammad Barzegar

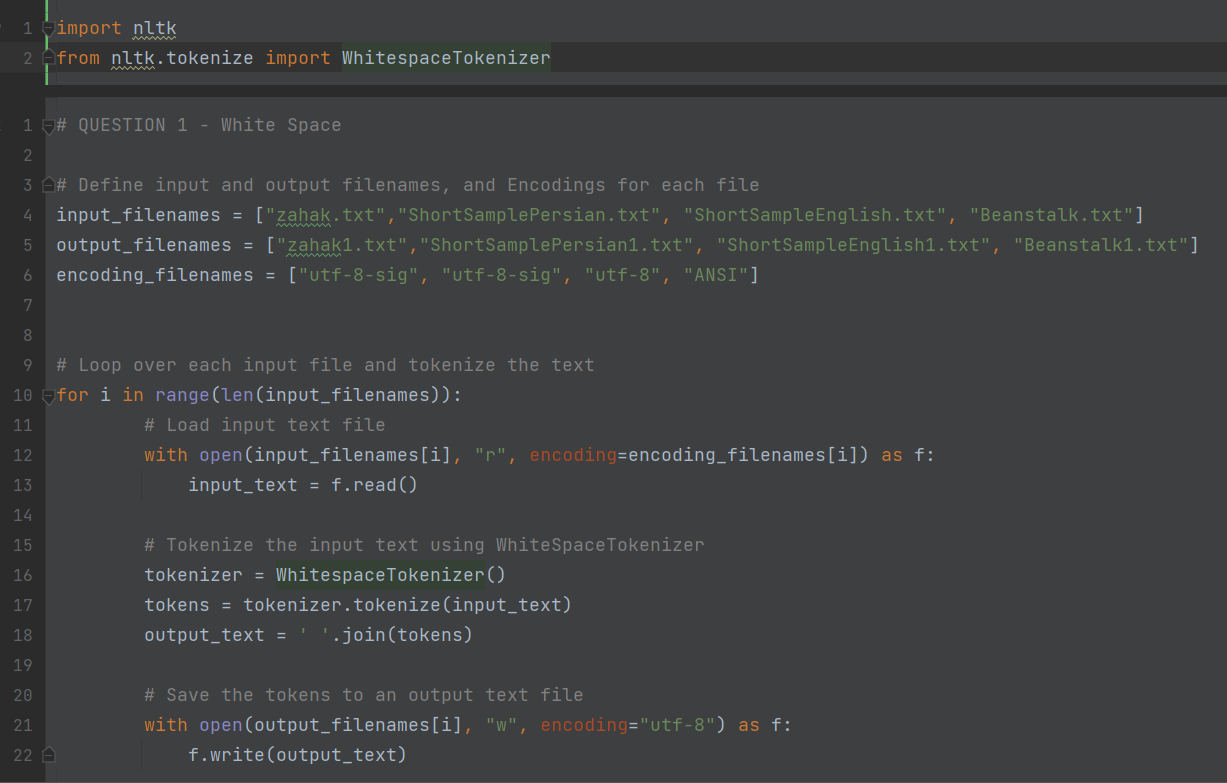
Student ID: 4010724001

**Persian Gulf University**

Faculty of Intelligent Systems Engineering and Data Science

**تمرین شماره 1 (WhiteSpaceTokenizer):**

با استفاده از متد WhiteSpaceTokenizer کاری میکنیم تا فاصله ها (اسپیس یا تب) و پاراگراف ها را حذف کرده و بین هر کلمه فقط یک اسپیس وجود داشته باشد.

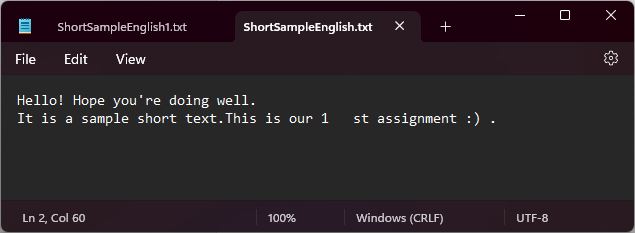


پس از ایمپورت کردن کتابخانه های مورد استفاده، اسم فایل های ورودی را در input\_filenames و اسم مورد نظر برای فایل های خروجی را در لیست output\_filenames نوشته ایم. انکودینگ هر فایل فرق دارد که در encoding\_filenames نام انکودینگ هر فایل نوشته شده است ( ایندکس مشابه input\_filenames).

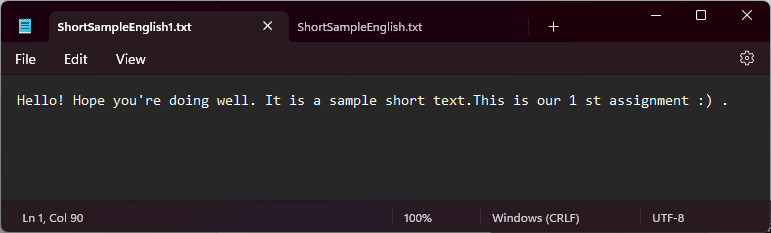
سپس در یک حلقه، تک تک هر فایل ورودی را لود کرده و متد WhiteSpaceTokenizer را روی آن انجام میدهیم و سپس با استفاده از دستور join کلمات را طوری به هم اتصال میدهیم که فقط یک اسپیس بین آن ها باشد. سپس فایل را ذخیره می کنیم.

برای هر تمرین، ورودی و خروجی کد را برای فایل ShortSampleEnglish می بینیم و سایر فایل ها در فولدر تمرین موجود می باشد.

فایل ShortSampleEnglish قبل از اعمال این کد:



فایل ShortSampleEnglish پس از اعمال این کد:



**دقت فرمایید فایل خروجی برای هر سوال، عدد آن سوال را در آخر اسم فایل خود دارد، مثلا zahak1 خروجی فایل zahak پس از انجام دادن دستورات تمرین 1 است.**

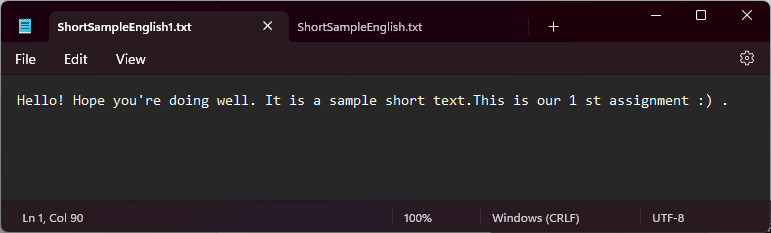
**تمرین شماره 2 (Lowercase):**

در این سوال، خروجی سوال قبلی را با استفاده از متد lower تمامی حروف را به حروف کوچک انگلیسی تبدیل می کنیم. بدیهی است که این کار برای فایل های فارسی نیاز نیست.

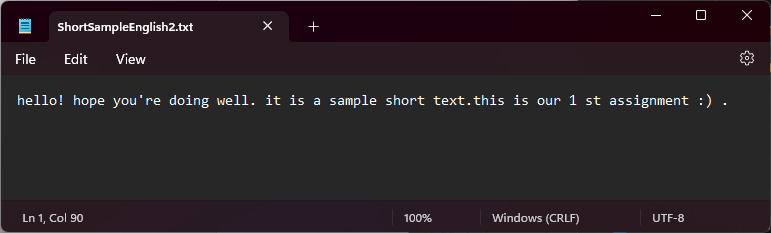
روال کلی مشابه سوال قبلی است:



فایل ShortSampleEnglish قبل از اعمال این کد:

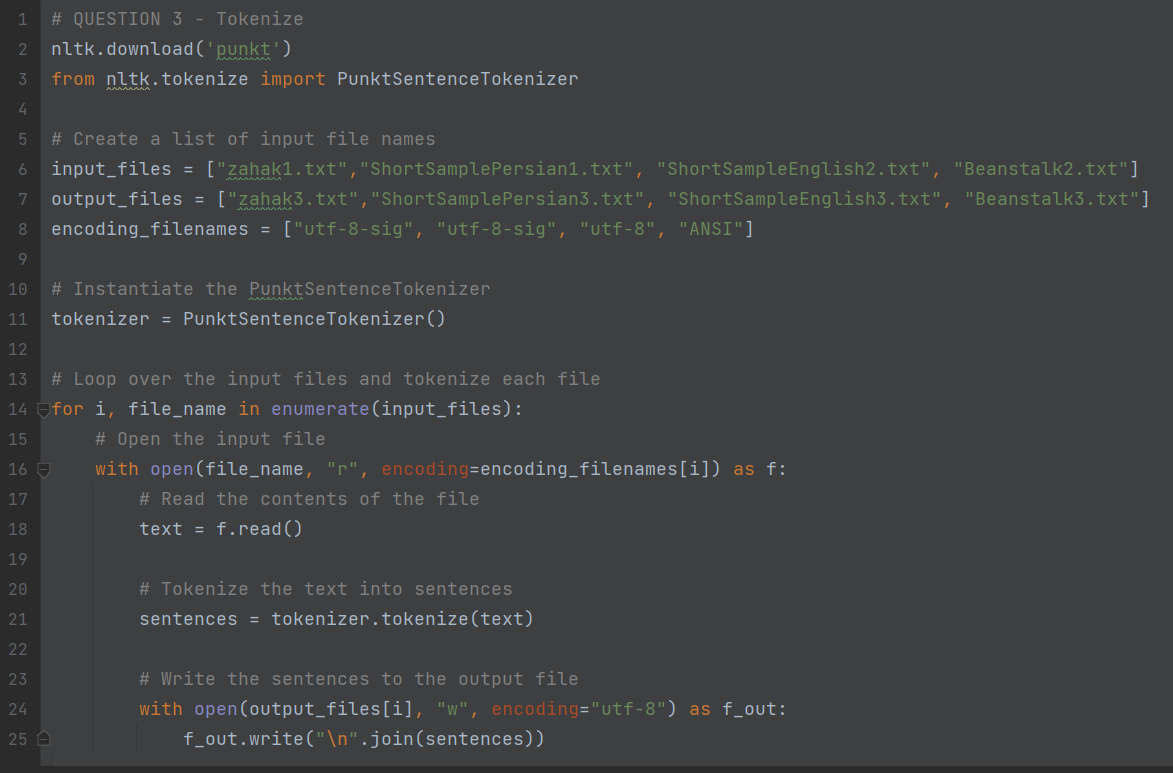


فایل ShortSampleEnglish پس از اعمال این کد:

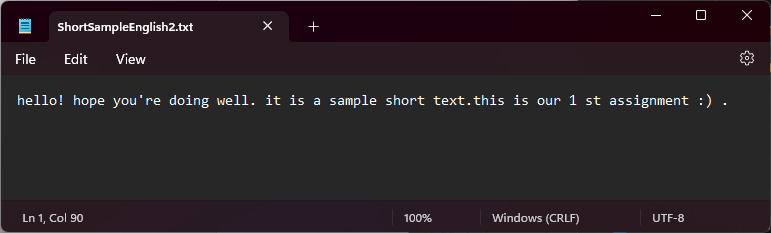


**تمرین شماره 3 (استخراج جملات و توکن ها):**

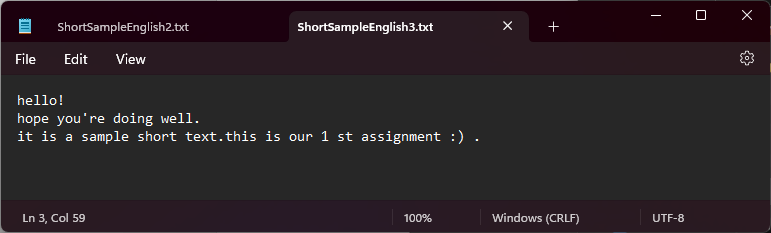
با استفاده از PunkSentenceTokenizer، جملات هر فایل را شناسایی و جداسازی انجام می دهیم. در انتها در هر خط یک جمله قرار خواهد داشت. ورودی متون انگلیسی از تمرین دوم و متون فارسی از تمرین اول است.



فایل ShortSampleEnglish قبل از اعمال این کد:

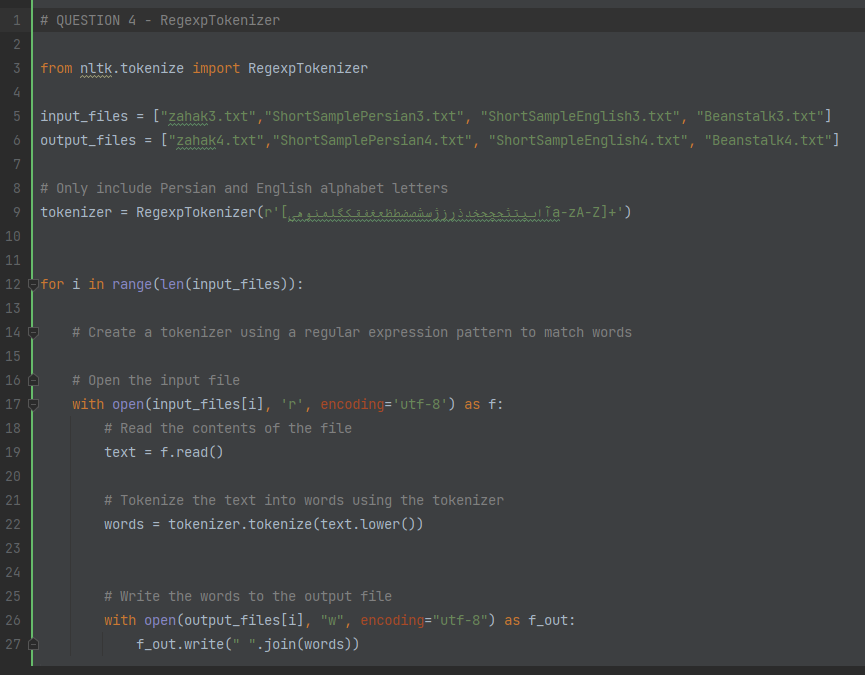


فایل ShortSampleEnglish بعد از اعمال این کد:



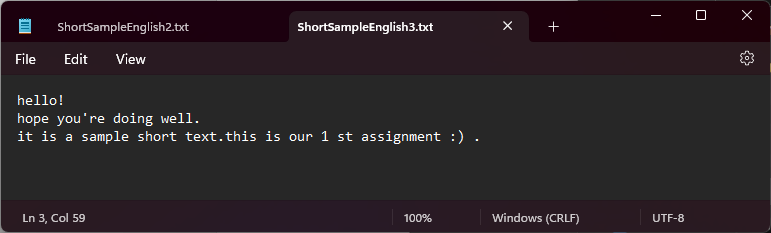
**تمرین شماره 4 (حذف علایم نگارشی):**

ورودی این مرحله، همان متون خروجی مرحله قبل است. اکنون میخواهیم همه علایم بجز حروف الفبا از متون حذف شود. با استفاده از RegexpTokenizer این کار را انجام می دهیم:

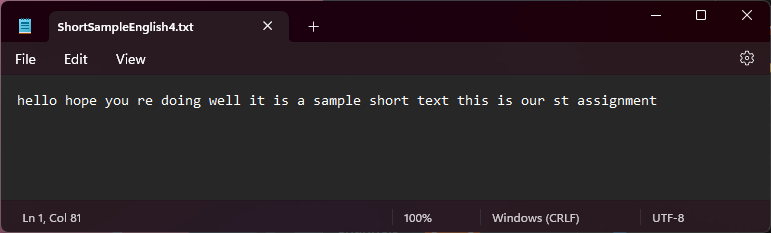


در لاین 9 به عنوان argument دستور RegexpTokenizer حروف الفبای فارسی و انگلیسی را قرار می دهیم، تا همه ی کاراکتر ها بجز 34 حرف فارسی و حروف بزرگ و کوچک انگلیسی a-z از متن حذف شود.

فایل ShortSampleEnglish قبل از اعمال این کد:



فایل ShortSampleEnglish قبل از اعمال این کد:



**تمرین شماره 5 (StopWords):**

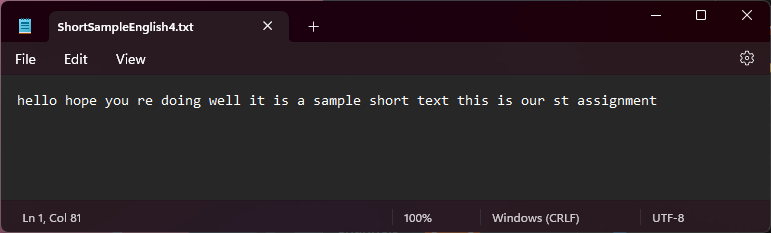
Stop Words اصطلاحا کلماتی هستند که در یک زبان با تکرار بسیار زیاد استفاده می شوند ولی معنای خاصی به متن ما اضاف نمی کنند. مثلا در زبان فارسی (آنچه، آنگاه، آن، آهان) و در انگلیسی (how, by, just).

لیست این کلمات انگلیسی در کتابخانه nltk موجود است و آن را ایمپورت میکنیم. ابتدا برای متن های انگلیسی این کار را انجام می دهیم:

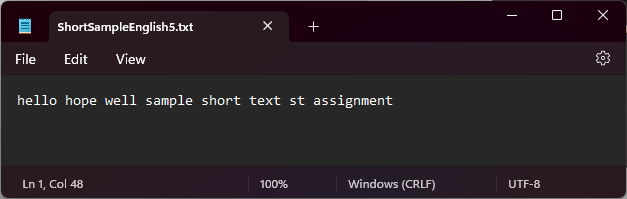


مشاهده می شود که یک فایل از 80 کاراکتر به 47 کاراکتر کاهش یافته، و فایل دیگر از 4265 به 2635 کاهش یافته است.

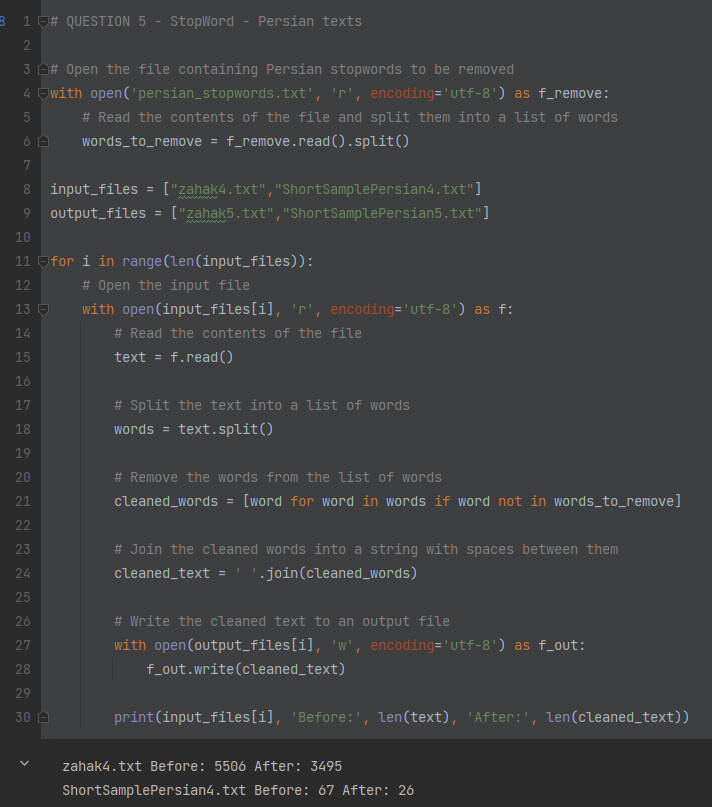
فایل ShortSampleEnglish قبل از اعمال این کد:



فایل ShortSampleEnglish بعد از اعمال این کد:

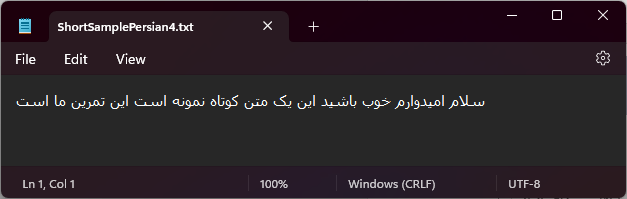


برای زبان فارسی در این کتابخانه stopword ها وجود ندارد و باید از منبعی دیگری آن ها را پیدا کنیم. یک برنامه نویس ایرانی به نام vahid kharazi لیستی از stopword های فارسی تهیه کرده و در گیتهاب قرار داده است، این لیست کلمات را استخراج و در فایل تکست ذخیره می کنیم (persian\_stopwords.txt).

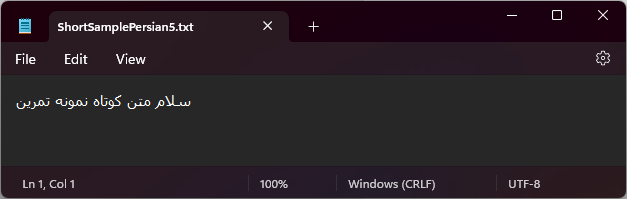
 لینک گیتهاب: <https://github.com/kharazi/persian-stopwords>

مشاهده می شود که یک متن از 5506 به 3495 کاراکتر و متن دیگر از 67 به 26 کاراکتر کاهش یافته است. این کار برای کاهش حجم فایل ها و همچنین حذف کلامتی که معنای خاصی به متن اضافه نمی کنند موثر است.

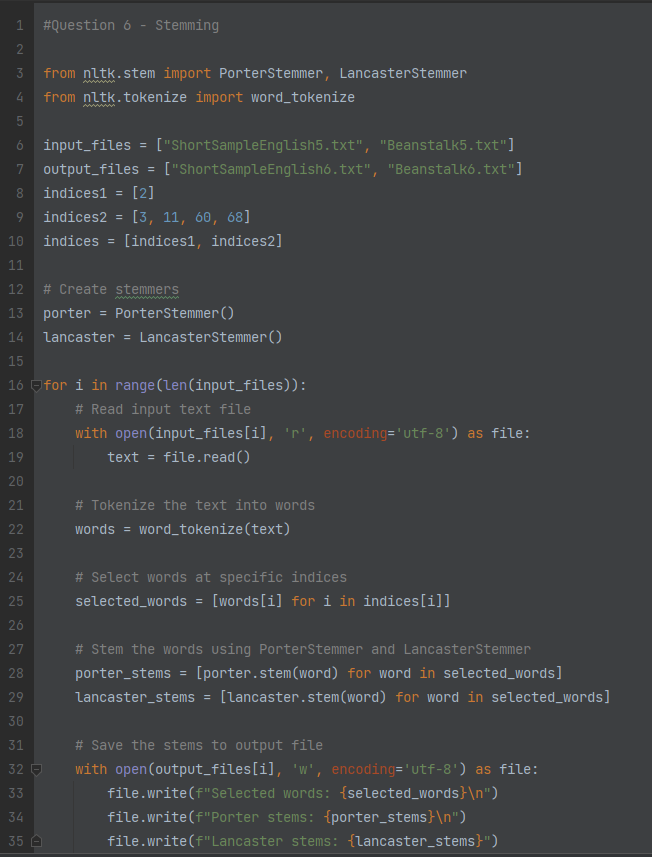
فایل ShortSamplePersian قبل از اعمال این کد:



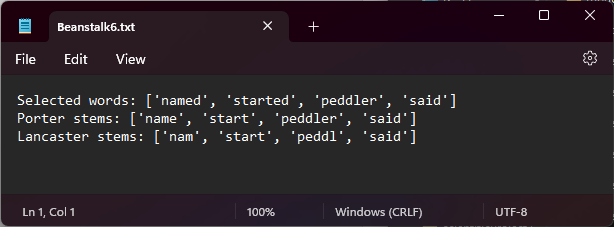
فایل ShortSamplePersian پس از اعمال این کد:



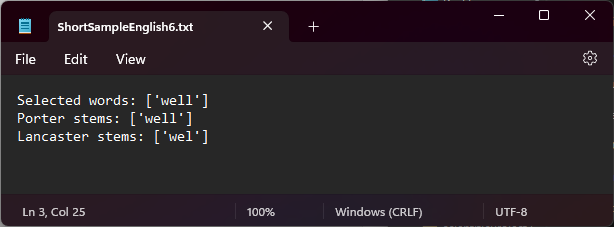
**تمرین شماره 6 (Stemming):**

در این قسمت با PorterStemmer و LancasterStemmer ریشه کلمات گفته شده در ایندکس های مورد نظر را استخراج خواهیم کرد.

خروجی برای BeansTalk:



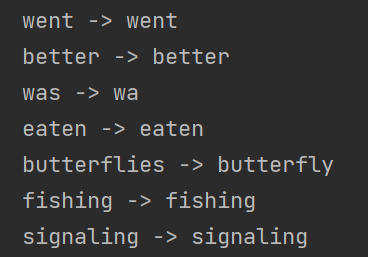
خروجی برای ShortSampleEnglish:

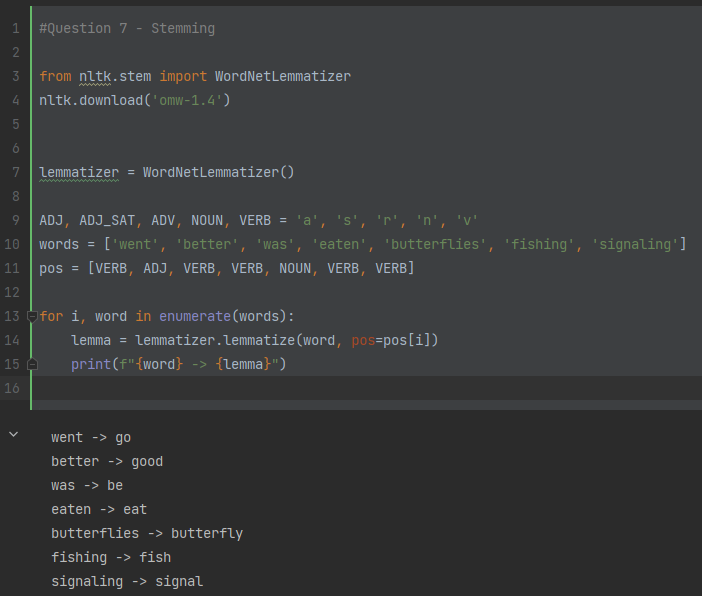


در سطر اول کلمات اصلی، در سطر دوم خروجی porter و در سطر سوم خروجی Lancaster نوشته شده است.

**تمرین شماره 7 (Lemmatization):**

با استفاده از WordNetLemmaztizer کلمات را به حالت نگارشی اولیه بر میگردانیم.

ابتدا با ورودی های پیشفرض متد lemmatize متوجه می شویم که پاسخ درست برای همه ی کلمات بر نمیگردد :

سپس با تعریف کردن آرگیومنت pos در متد lemmatize، برای هر کلمه تعریف می کنیم که این کلمه فعل، اسم، صفت یاچه نوع کلمه دیگری است، سپس مشاهده شد که خروجی ها درست تولید می شود: