

1402-02-29

Natural Language Processing

Dr. Mohammad Bidoki

Homework #3

By: Mohammad Barzegar

Student ID: 4010724001

**Persian Gulf University**

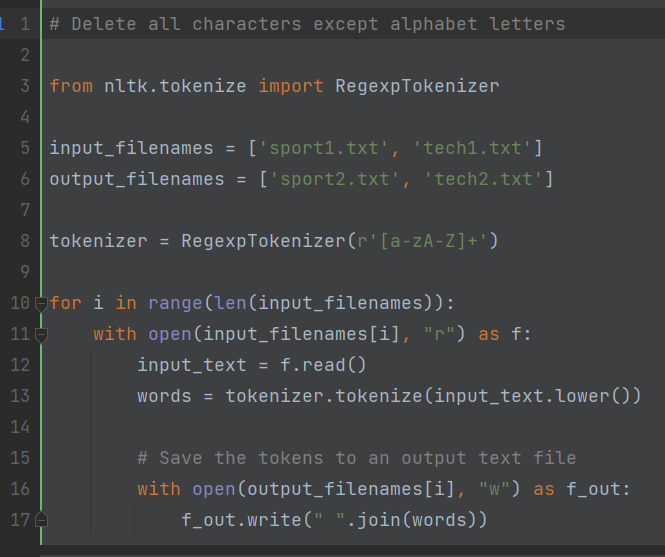
Faculty of Intelligent Systems Engineering and Data Science

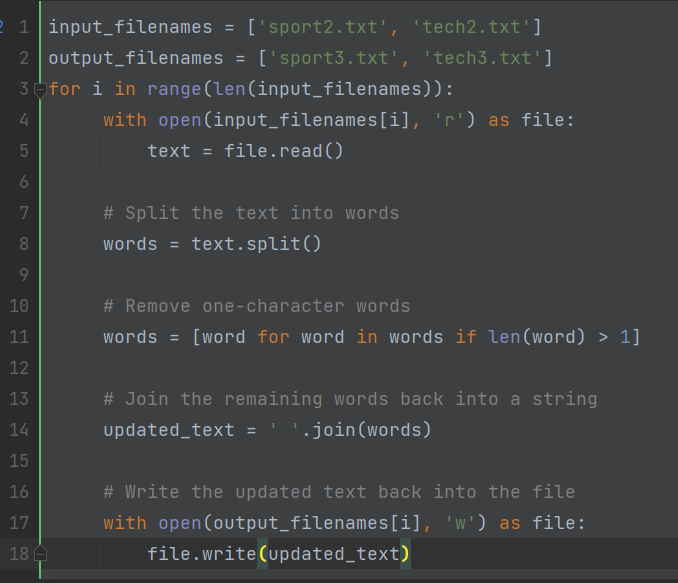
**تمرین a (توکنایز و حذف علامت های اضافی):**

ابتدا با استفاده از WhiteSpaceTokenizer، اسپیس ها و تب ها و لاین بریک ها را پاک کرده تا فایل یکدست شود:

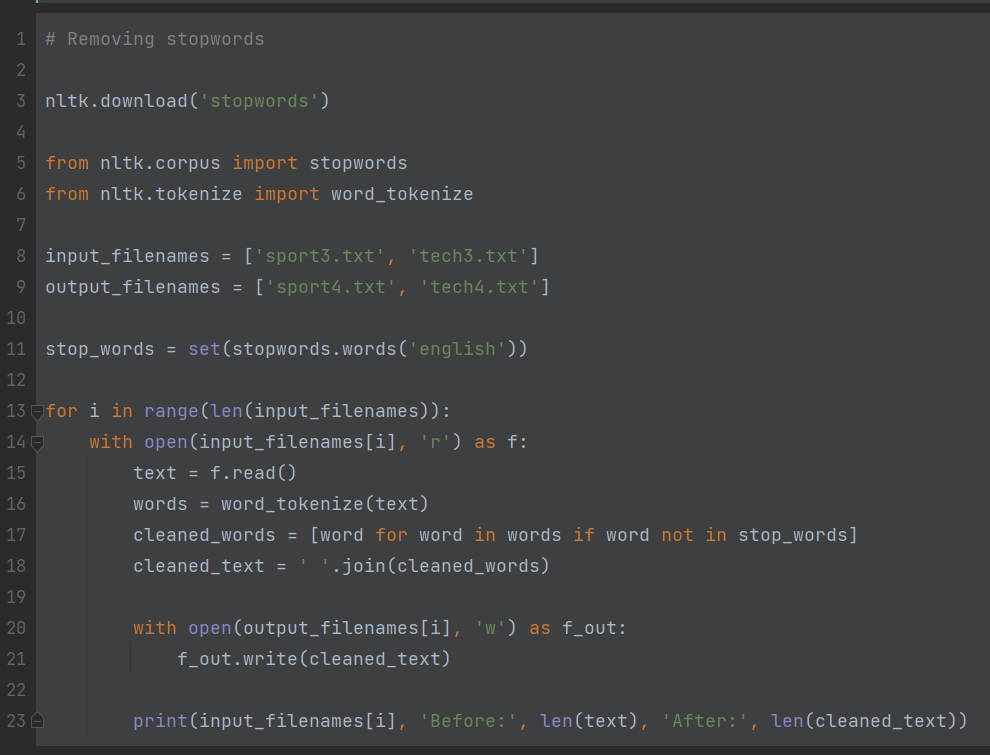


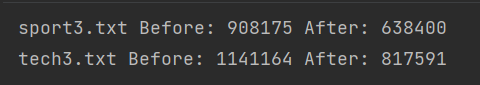
**تمرین b (حذف کردن stopword ها و حروف تک کاراکتری):**

****خروجی مرحله قبل، ورودی این مرحله است. با استفاده از RegexpTokenizer همه ی کاراکتر ها بجز حروف الفبای انگلیسی را حذف می کنیم:

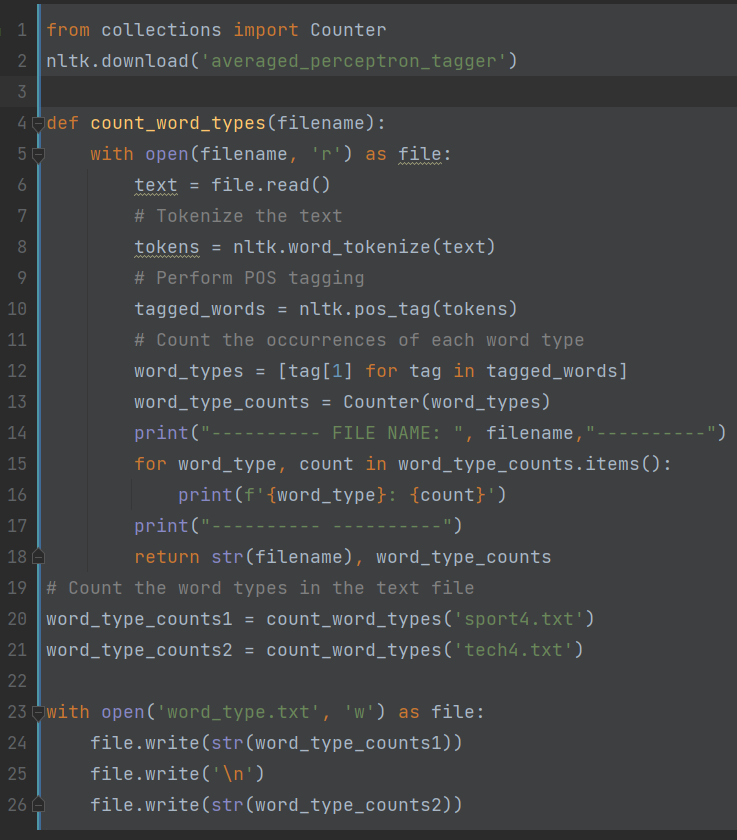
خروجی مرحله قبل را برداشته و سپس حروف تک کاراکتری را حذف می کنیم:

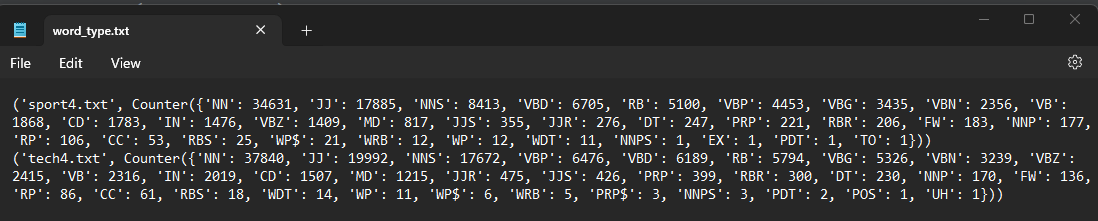
با استفاده از خروجی مرحله قبل، حروف stopword را نیز حذف میکنیم:





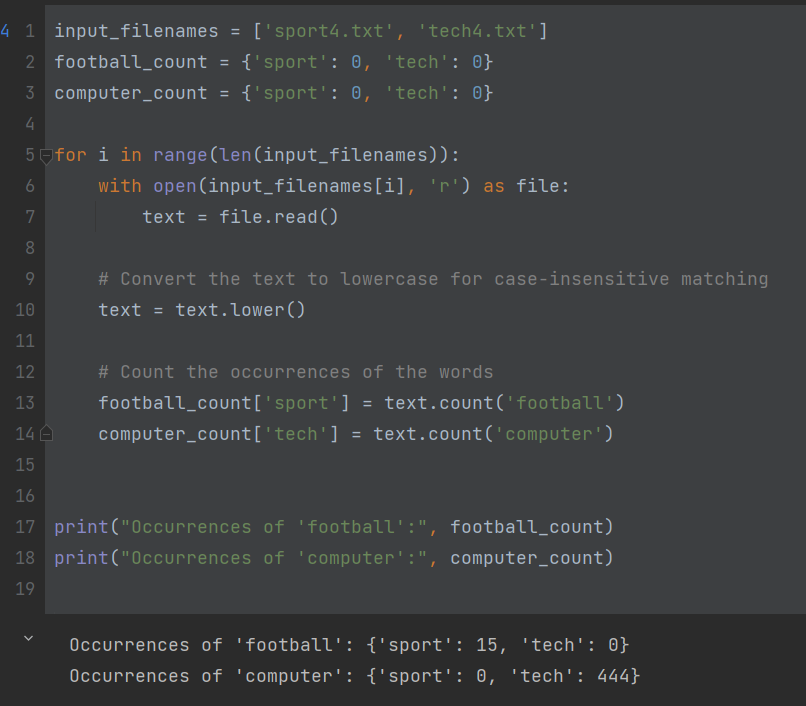
**تمرین c (type های موجود در هر متن):**

با استفاده از متد pos\_tag، تایپ هر کلمه را برای هر فایل مشخص میکنیم. خروجی را در فایل word\_type.txt می توانید مشاهیده بفرمایید.



**تمرین d (تعداد کلمه football و computer):**

تعداد تکرار هر کلمه در هر دو فایل را به دست می آوریم:

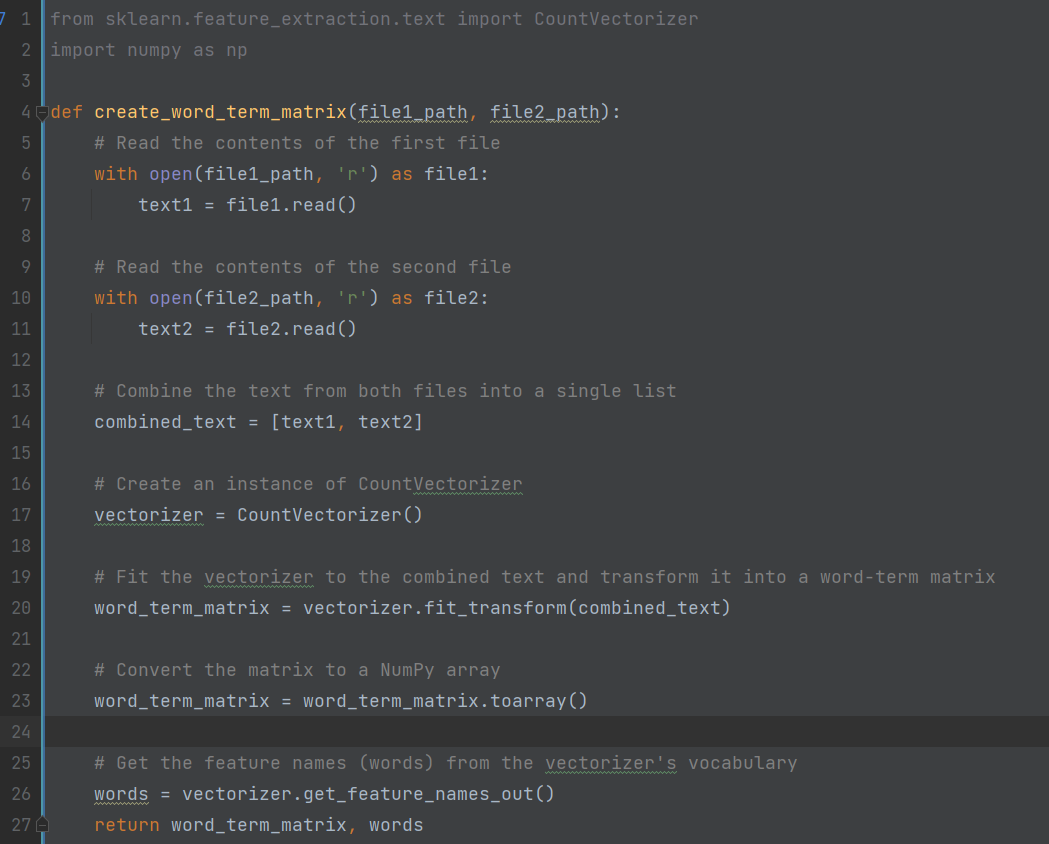


تعداد تکرار football در فایل sport 15 و در فایل tech 0 بار است.

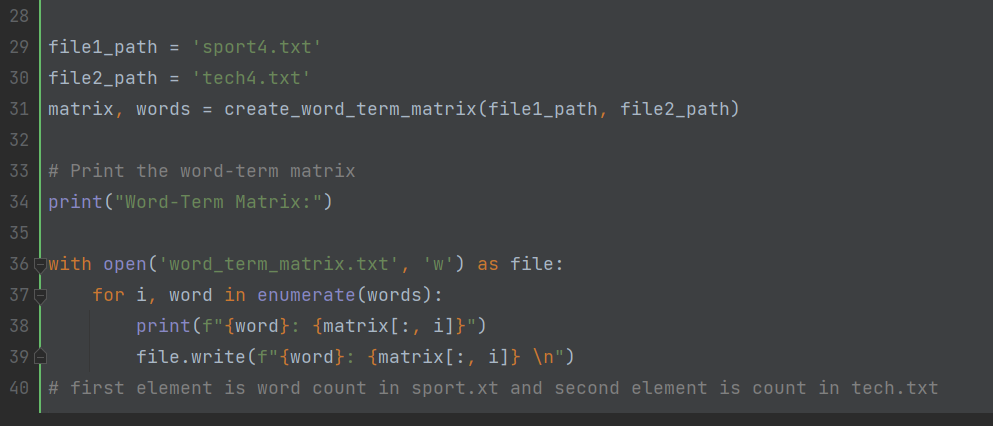
تعداد تکرار computer در فایل sport 0 و در فایل tech 444 بار است.

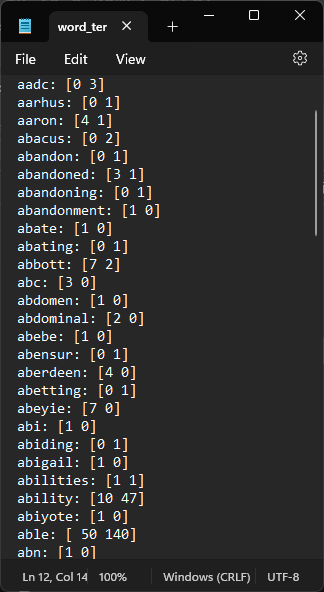
**تمرین e (ماتریس term-document):**

ابتدا تابعی مینویسیم که با گرفتن نام 2 فایل، ماتریس term-document را بسازد. از متد countervectorizer استفاده خواهیم کرد:

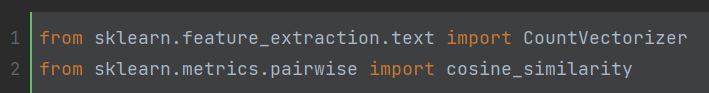


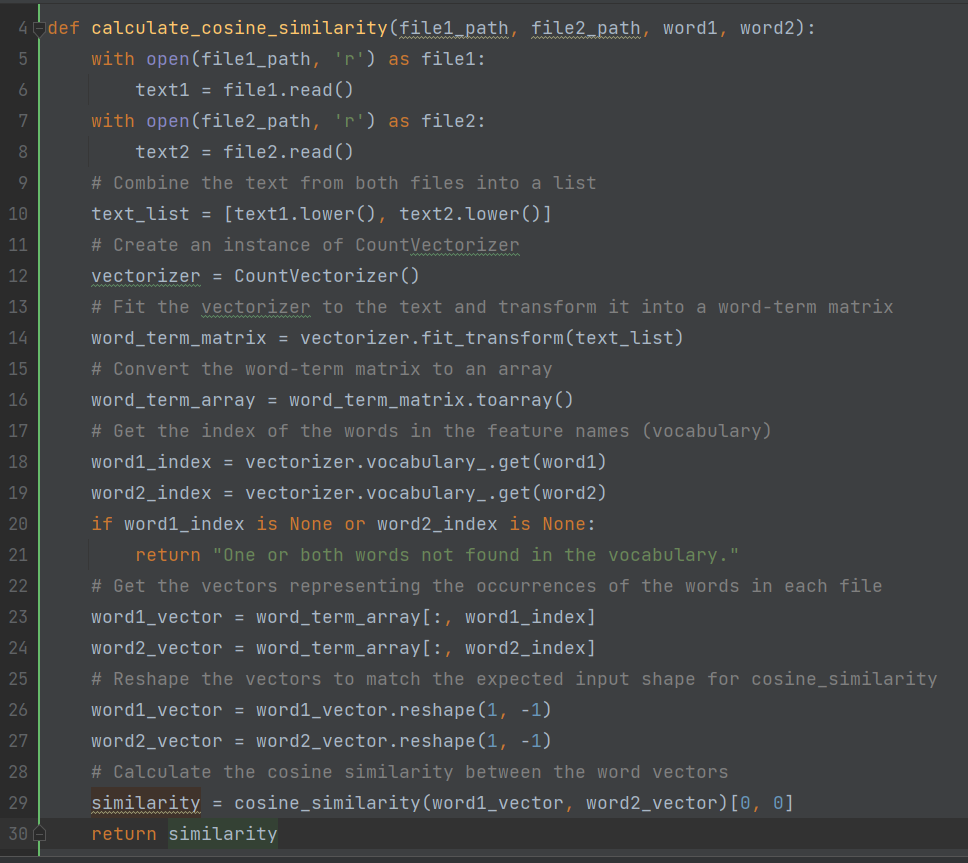
سپس با استفاده از این تابع، خروجی مورد نظر را در فایل word\_term\_matrix.txt ذخیره می کنیم که در فایل تحویلی موجود است. المان اول (چپ) تعداد کلمه در فایل sport و المان دوم (راست) تعداد در فایل tech است.



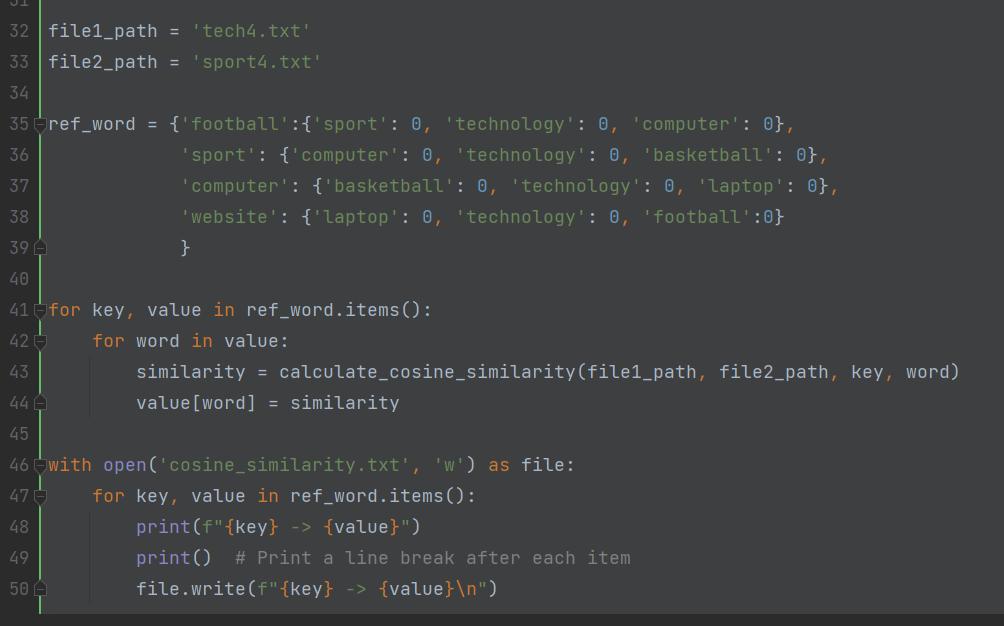


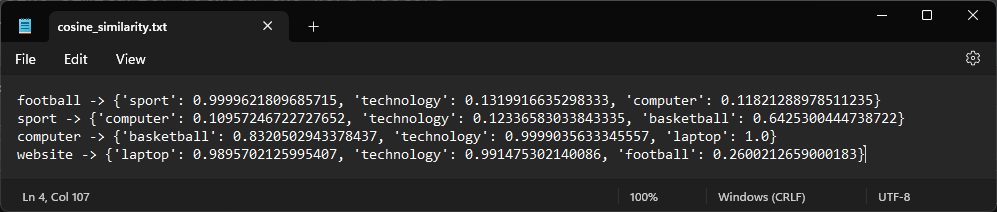
**تمرین f (شباهت کسینوسی):**

با استفاده از کتابخانه sklearn و متد های countervectorizer و cosine\_similarity تابعی می سازیم تا شباهت کسینوسی را محاسبه کند:



سپس با استفاده از تابع فوق، ورودی دو فایل تکست را داده و شباهت های کسینوسی مورد نظر را به عنوان خروجی در فایل cosine\_similarity.txt که در فولدر تحویل داده موجود است، درج می کنیم.





|  |  |  |
| --- | --- | --- |
| شباهت کسینوسی | کلمات | کلمه مرجع |
| 0.9999621809685715 | Sport | Football |
| 0.1319916635298333 | Technology |
| 0.11821288978511235 | Computer |
| 0.10957246722727652 | Computer | Sport |
| 0.12336583033843335 | Technology |
| 0.6425300444738722 | Basketball |
| 0.8320502943378437 | Basketball | Computer |
| 0.9999035633345557 | Technology |
| 1.0 | Laptop |
| 0.9895702125995407 | Laptop | Website |
| 0.991475302140086 | Technology |
| 0.2600212659000183 | Football |

همانگونه که نشان داده شده است، بیشترین شباهت کلمه football با کلمه ی sport، بیشترین شباهت کلمه sport با کلمه ی basketball، بیشترین شباهت computer با laptop، و بیشترید شباهت website با technology است.