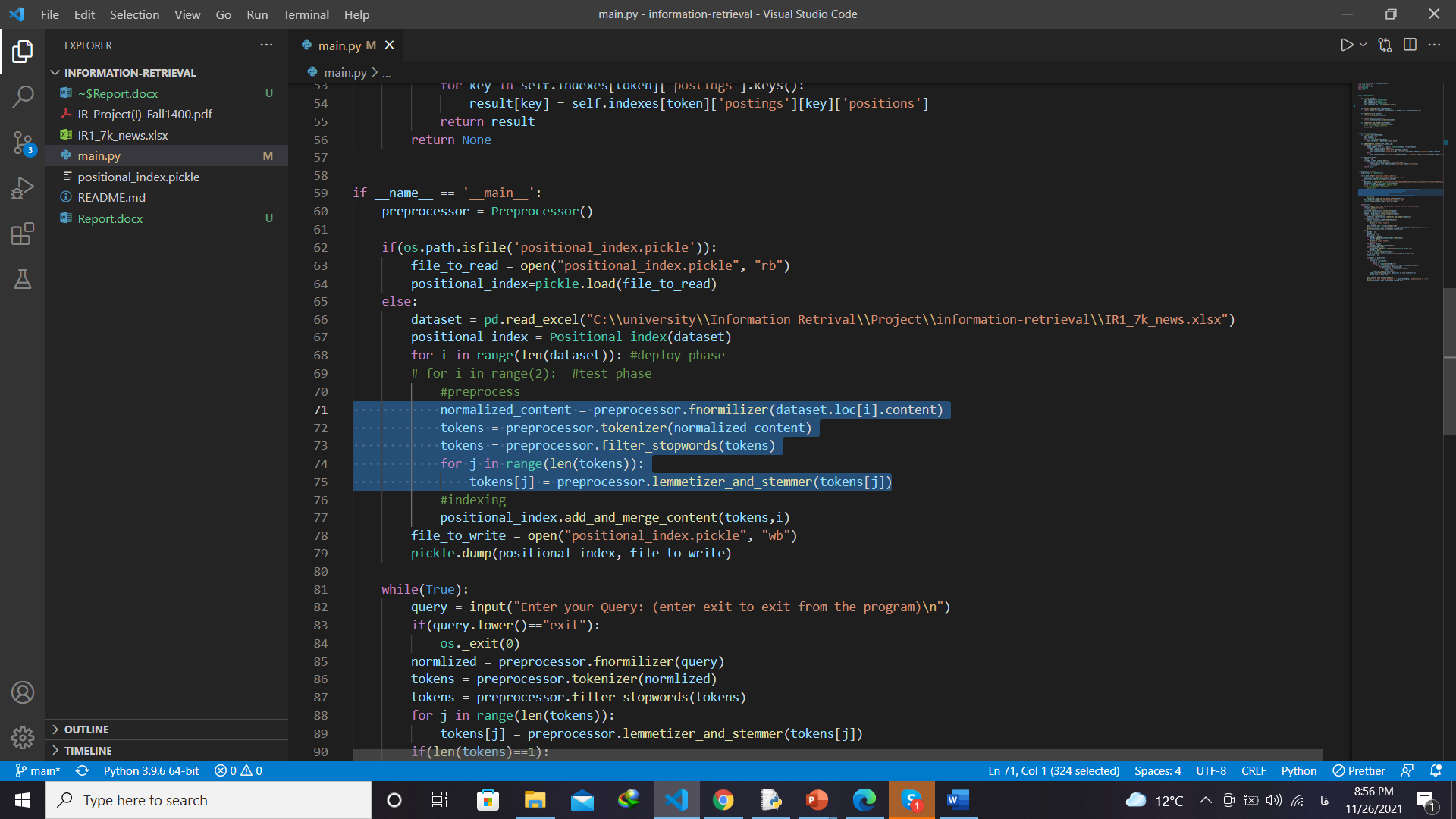
حسن کاظمی طهرانی

گزارش فاز اول پروژه بازیابی اطلاعات

9629041

1-

پیش پردازش های انجام شده در قطعه کد زیر صورت می گیرد:



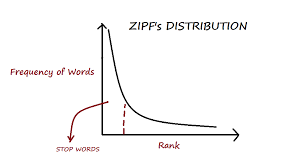
برای اینکار از کتابخانه hazm استفاده شده است. در مرحله اول متن را نرمال سازی می کنیم. نرمال سازی برای این می باشد که در سراسر متن به طور مثال نیم فاصله رعایت شود و به یک شکل باشد و این مرحله کلمه ای مانند "پیش نویس" را به "پیش‌نویس" تبدیل می کند. اگر این کار انجام نشود ممکن است یک جا "پیش نویس" به صورت جدا نوشته شده باشد و جزو کلمات مرتبط به حساب نیاید زیرا که "پیش‌نویس" در اصل با نیم فاصله می باشد.

در مرحله بعدی توکن سازی انجام می دهیم تا متن به صورت کلمه کلمه در آید. این مرحله برای ساخت شاخص مکانی کاربرد دارد. در این مرحله جمله "من علی هستم" به توکن های "من" و "علی" و "هستم" تبدیل می شود.

در مرحله بعد کلمات پرتکرار زبان فارسی از متن حذف می شود. این اقدام از این جهت صورت می گیرد که چنین کلماتی تاثیری در مرتبط بودن محتوی متن ندارند و پردازش آن ها صرفا بار اضافی ای بر سیستم می باشد. در این مرحله کلماتی مانند "در"، "و"، "." و ... حذف می شوند و سرعت ساخت شاخص به شدت افزایش پیدا می کند.

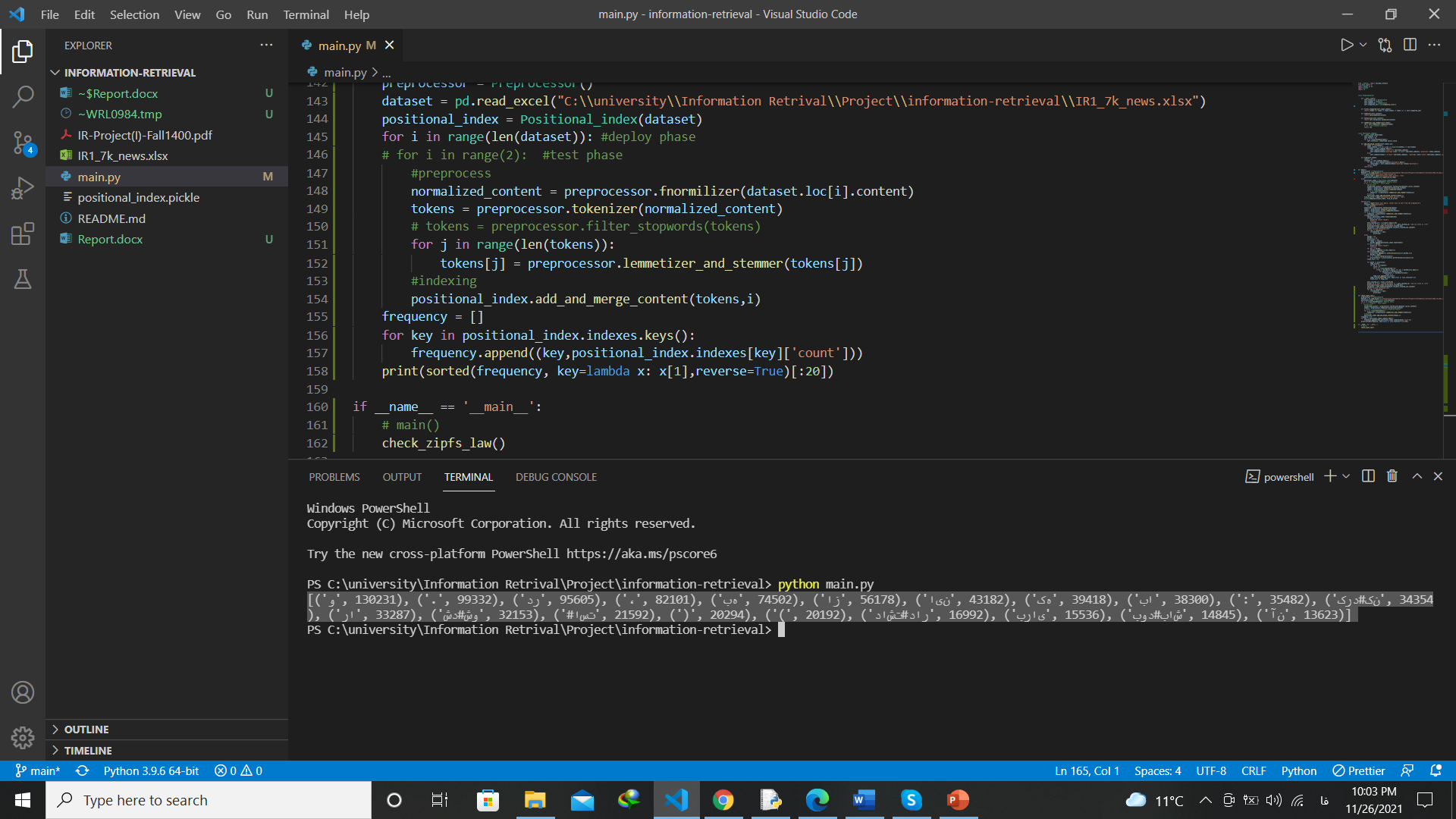
در مرحله آخر ریشه یابی صورت می گیرد. این امر از این جهت می باشد که هنگامی که کاربر کلمه ای مانند "درخت" را جستجو کرد متن هایی با محتوی "درخت ها" نیز به آن ها نمایش داده شود. زیرا که متن حاوی "درخت ها" قطعا به درخت مرتبط است و در صورت عدم ریشه یابی کاربر باید حتما کلمه درخت ها را جستجو کند تا سند به آن نمایش داده شود. همچنین در این مرحله فعل ها به صورت مصدر ماضی خود در می آیند. مانند "می روی" که تبدیل به "رفت" می شود.

2-

به کمک قانون زیف می توان کلمات پرتکرار با بار اطلاعاتی کم را شناسایی کرد و آن ها را حذف کرد و به این صورت تحیل های ما روی متن بسیار efficient تر خواهد بود.

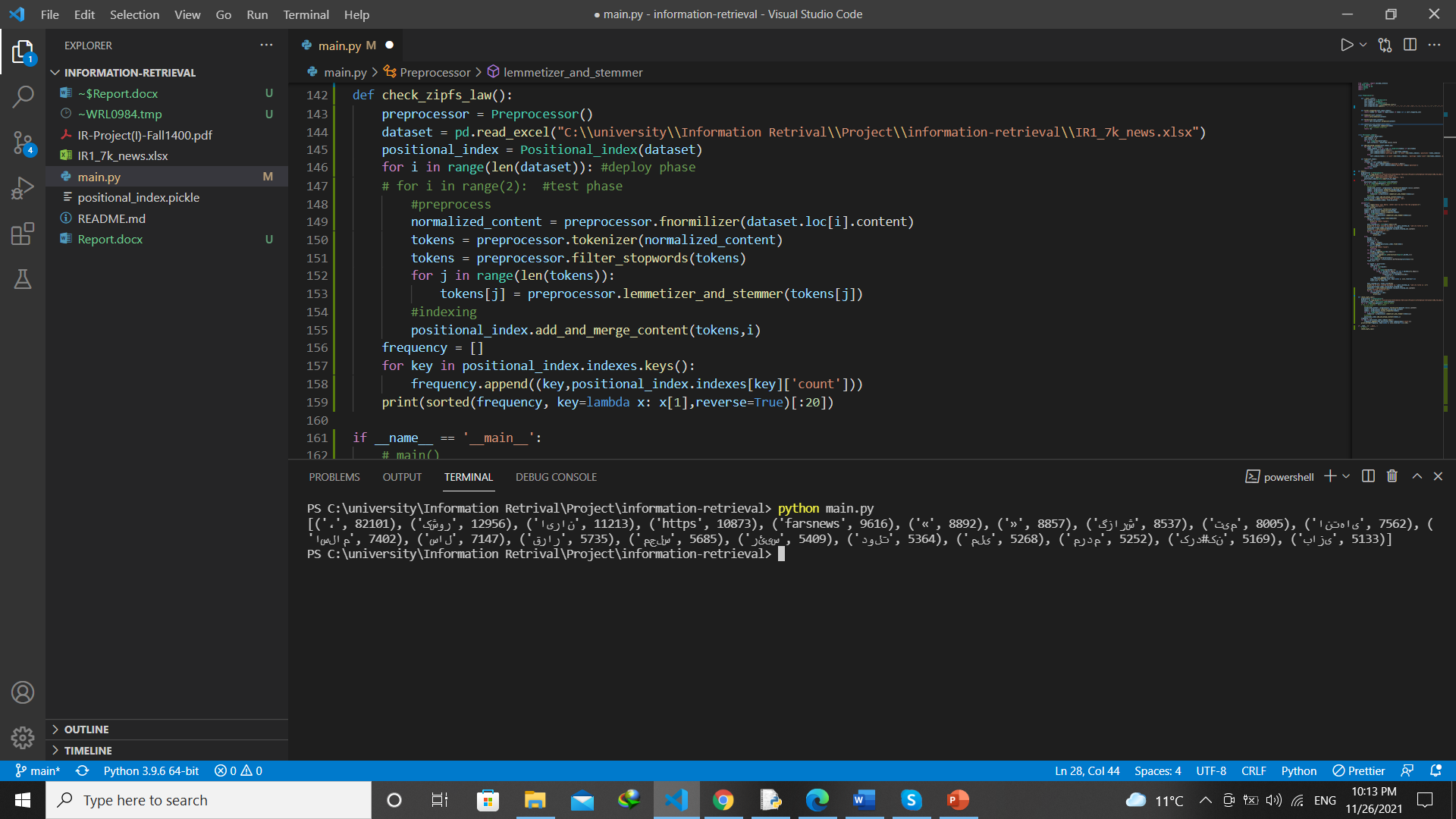
بدون حذف کلمات پرتکرار:

طبق تکرار کلمات مرتب شده اند و 20 تای اول آن ها به نمایش در آمده است. قانون زیف می گوید در یک متن طبیعی فراوانی i امین کلمه پرتکرار رابطه عکس با ضریب 1/i با پرتکرار ترین کلمه دارد.



بدون حذف پرتکرار ترین کلمه "و" می باشد با مقدار 130231 و بعد از آن کلمه "در" با 95605 تکرار. طبق قانون زیف باید 95605 = k/2 باشد. و همچین 74502 که مربوط به کلمه پرتکرار بعدی یعنی "به" است باید با k/3 برابر باشد. طبق رابطه اول k=191210 و طبق رابطه دوم k=223506. تقریبا قانون زیف رعایت شده است که علت این امر وجود کلمات پرتکرار می باشد.

با حذف کلمات پرتکرار:

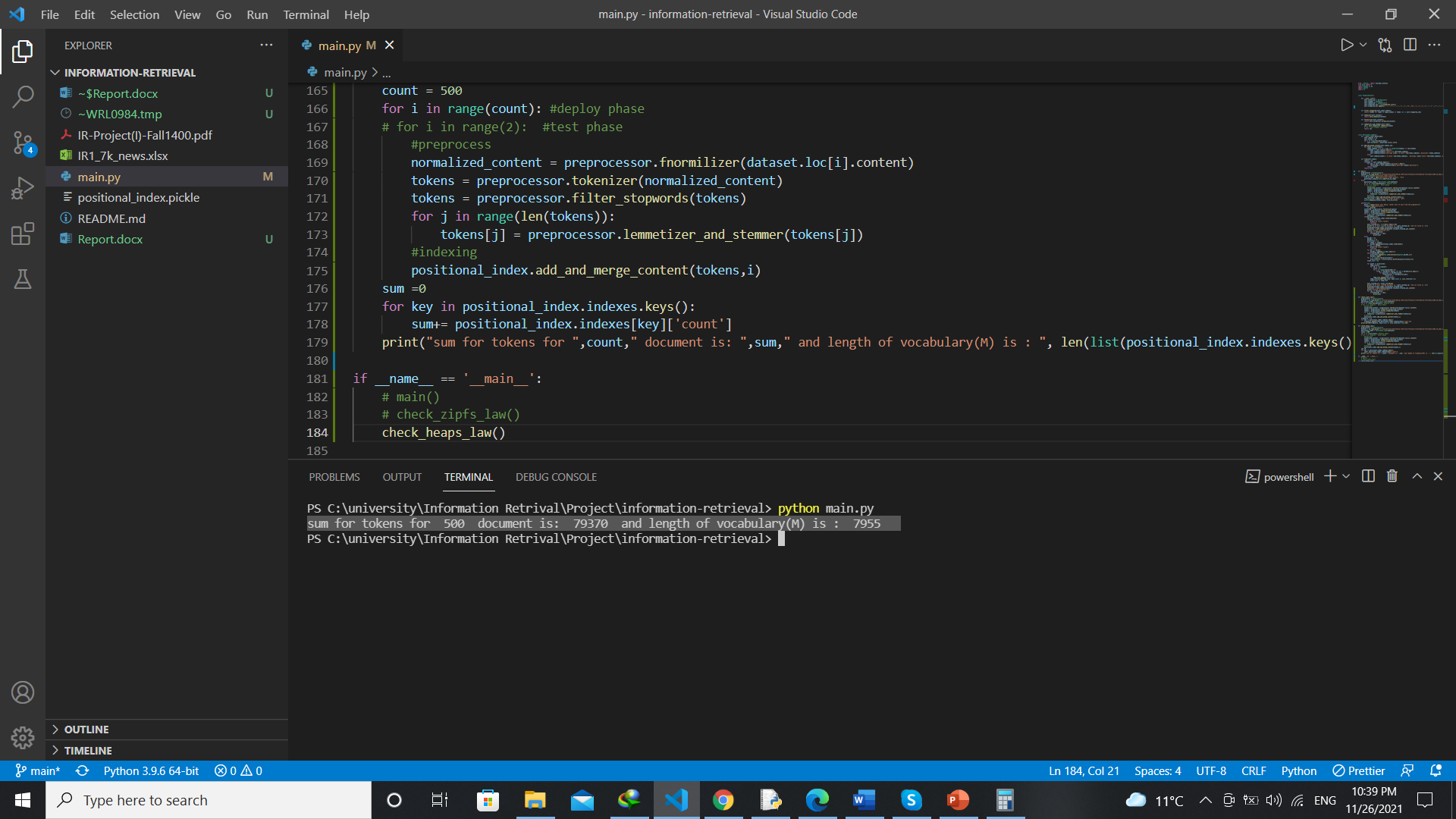


با حذف کلمات پرتکرار نظم حاکم در متن طبیعی از بین می رود و قانون زیف رعایت نمی شود. در این حالت k برای i=2 برابر است با 22426 و برای i=3 برابر است با 32000 است. این امر نشان می دهد که ما توانسته ایم درصد زیادی از کلمات که بار اطلاعاتی کمی دارند را حذف کنیم.

3-

* Heaps’ law: *M = kTb*

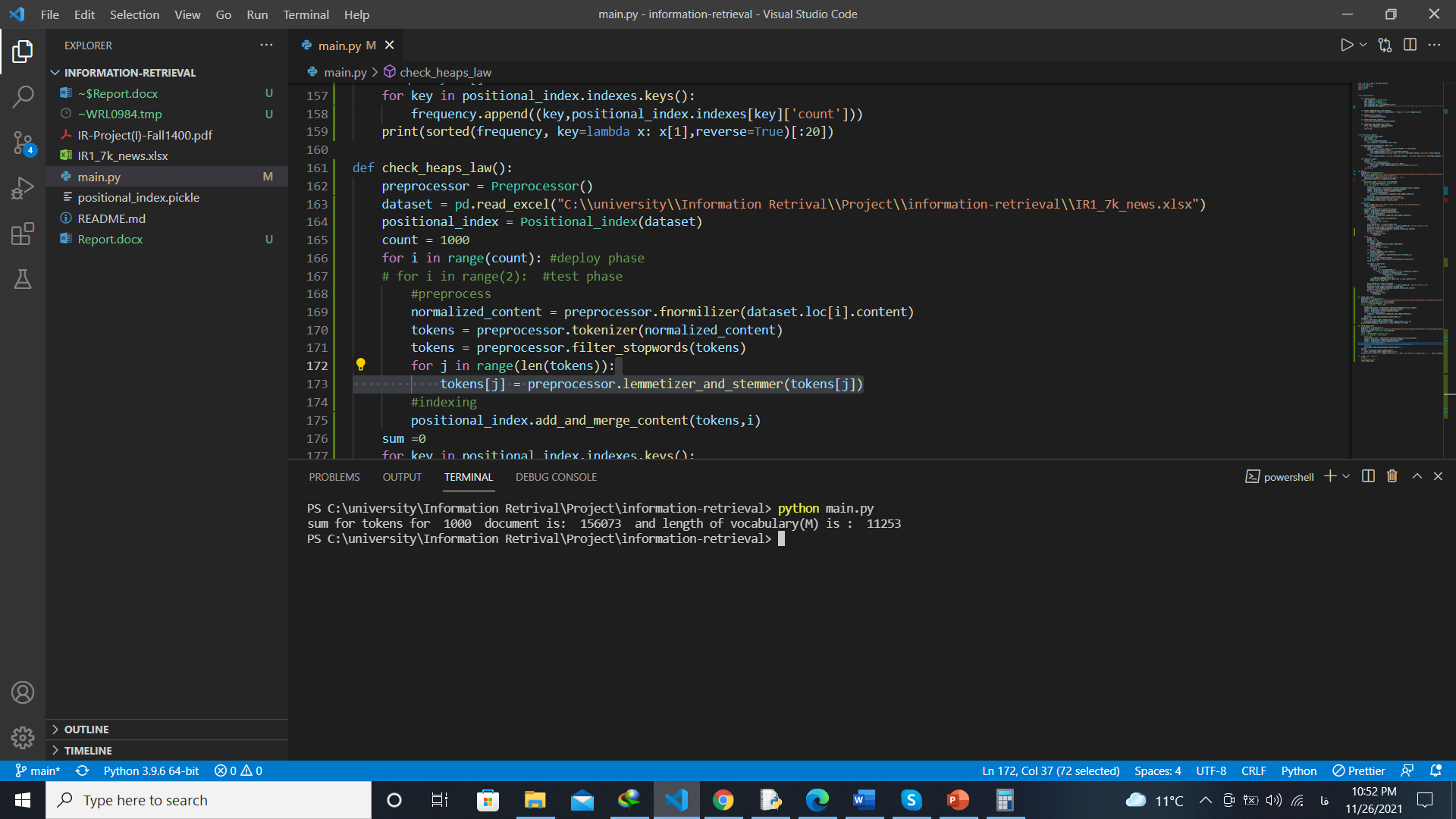
حالت با ریشه یابی: تعداد 500



طبق محاسبات : M=8000 و T=80000

طبق معادله قانون هیپس :k=30 و b=0.5

حالت با ریشه یابی تعداد 1000:



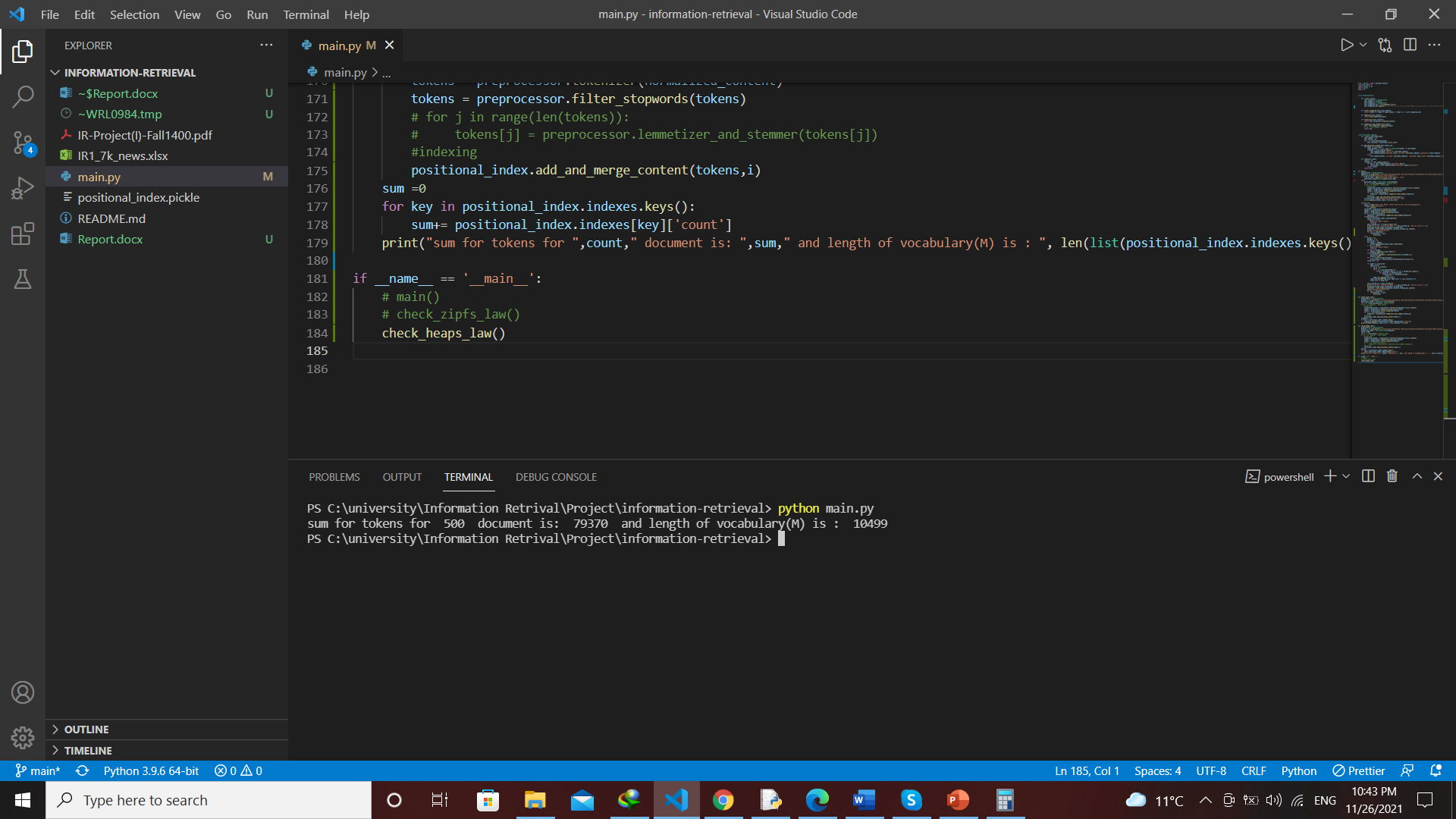
حال تعداد از 500 به 1000 تبدیل شده است. تعداد توکن برابر 156073 می باشد. طبق قانون هیپز با اطلاعات موجود در قسمت قبل k=30 و b=0.5 .

kTb=30\*1560730.5 =11850 . تعداد کلمات تخمینی

تعداد کلمات واقعی 11253.

به طور تقریبی میتتوان گفت قانون هیپز رعایت شده است و به کمک آن توانستیم تعداد کلمات را تخمین بزنیم.

حالت بدون ریشه یابی:



مشاهده می شود تعداد کلمات منحصر به فرد افزایش پیدا کرده است اما همچنان تعداد توکن ها ثابت مانده است. این امر نشان می دهد که بدون ریشه یابی تنها تعداد کلمات منحصر به فرد بیشتر می شود و توکن ها ثابت می ماند که این امر تنها بار حافظه ای مارا زیاد می کند.

4-

در روند ریشه یابی فعل ها به صورت مصدر ماضی خود در می آیند و به طور مثال فعل "رفتم" به فعل "رفت" تبدیل می شود. در روند ریشه یابی یکی از اقدامات این می باشد که کلماتی که با "م" ختم می شوند، حرف "م" از انتهای آن ها حذف می شود. همچنین در روند ریشه یابی باید فعل "رفتی" به "رفت" تبدیل شود. در نتیجه ناخواسته طبق روند ریشه یابی کلماتی که فعل نیستند و انتهای آن ها حرف"ی" می باشد تغییر می کنند.همچنین فعل های امری مثل فعل "برویم" که به "رفت" تبدیل می شود.

نمونه ها:

سلامتی – سلامت

مسلمی – مسلم

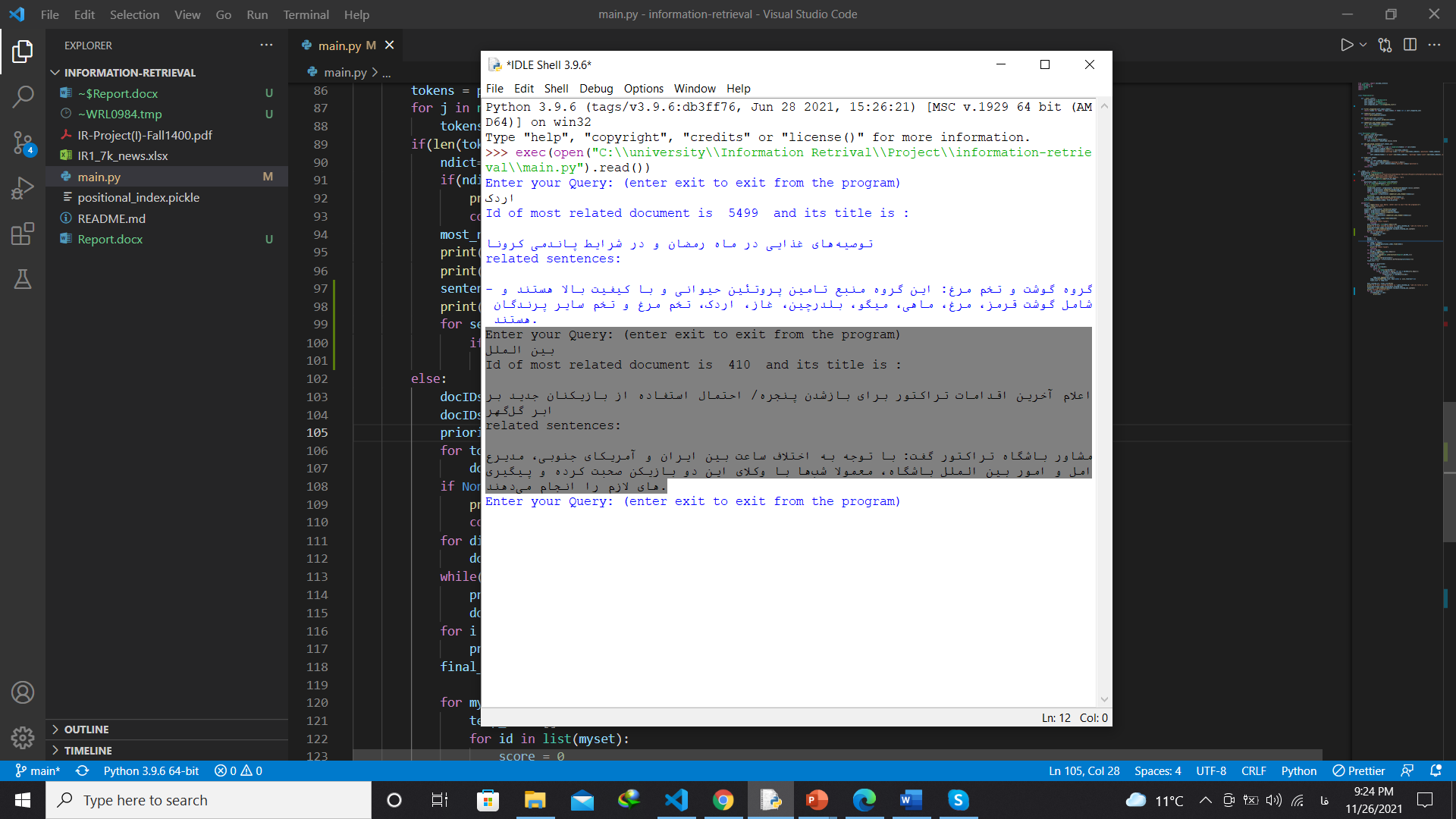
تلاطم – تلاط

بهنام - هنام

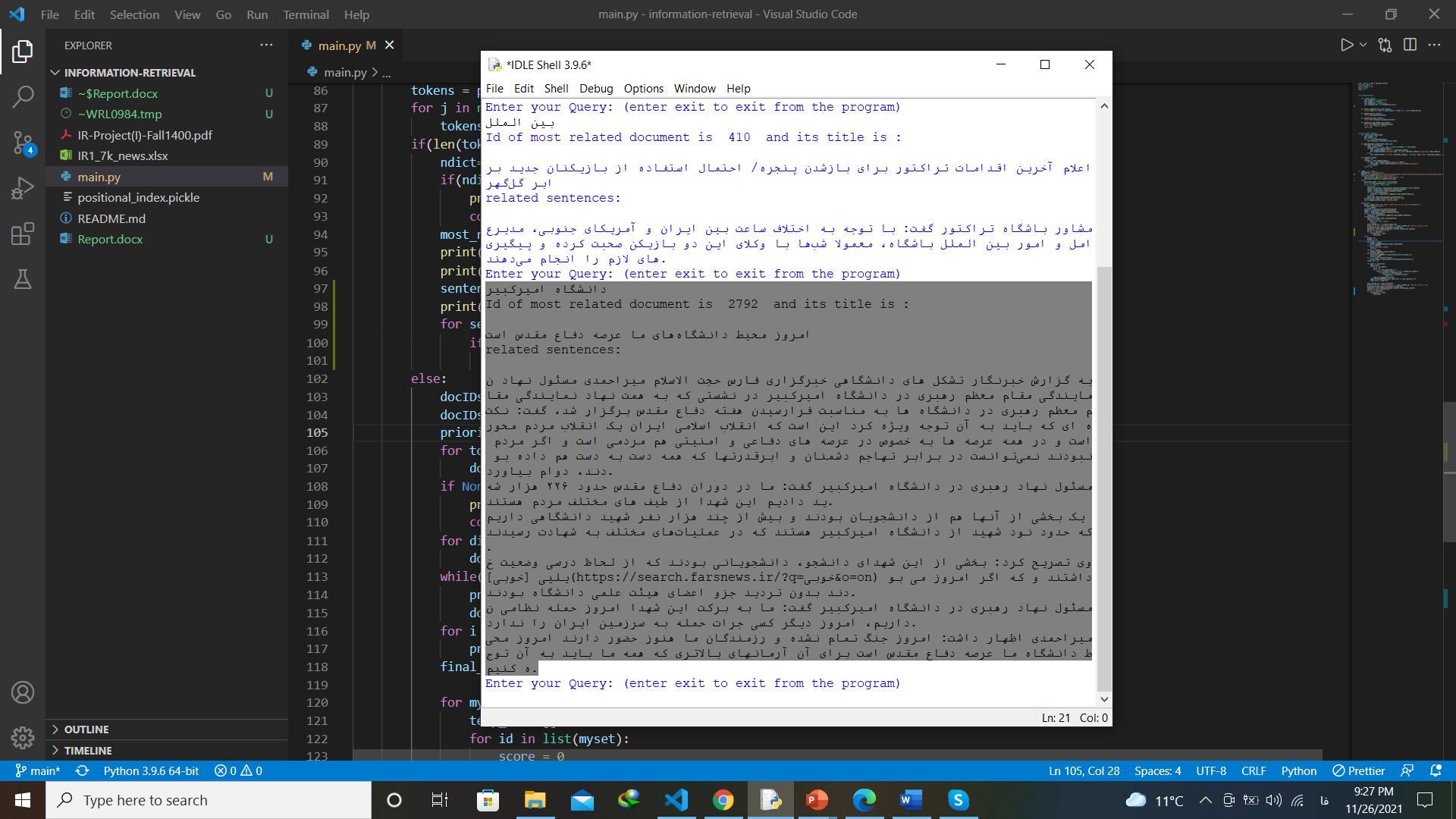
برای جلوگیری از این اتفاق باید قبل از ریشه یابی کلمات ، برچسب گذاری صورت گیرد. به این صورت اگر کلمه ای فعل باشد ریشه یابی صورت می گیرد اما اگر اسم باشد اتفاقی نمی افتد.

5- از آنجایی که سند های مرتبط تعدادشان زیاد است تنها مرتبط ترین سند نمایش داده شده است. در صورت نیاز می توان لیست تمام سند های مرتب اولویت بندی شده را چاپ کرد.

خروجی ها:

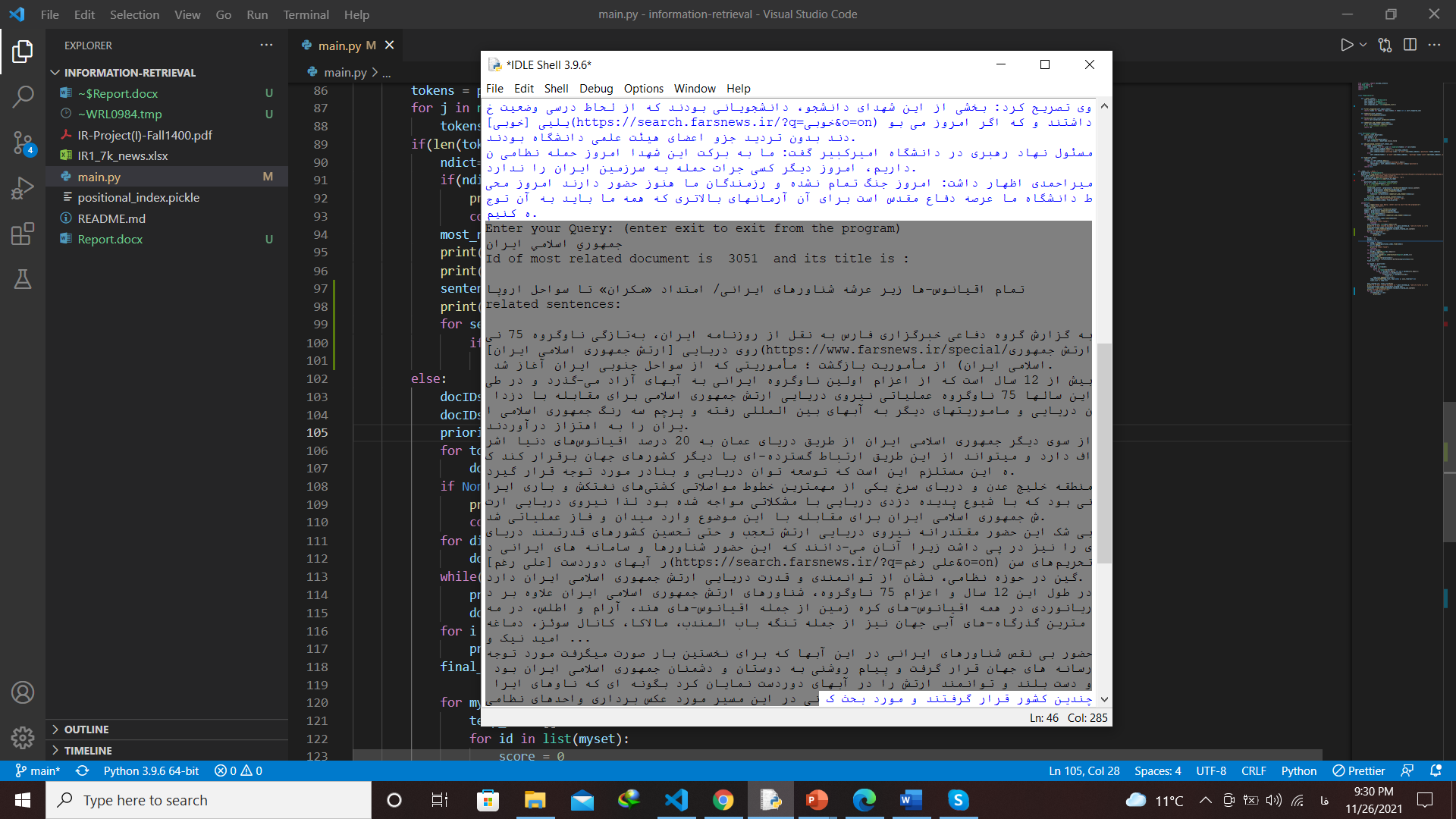
الف) خروجی برای یک پرسمان متداول تک کلمه ای: بین الملل

مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. زیرا که در متن امر بین الملل آمده است که با بین الملل مرتبط است.

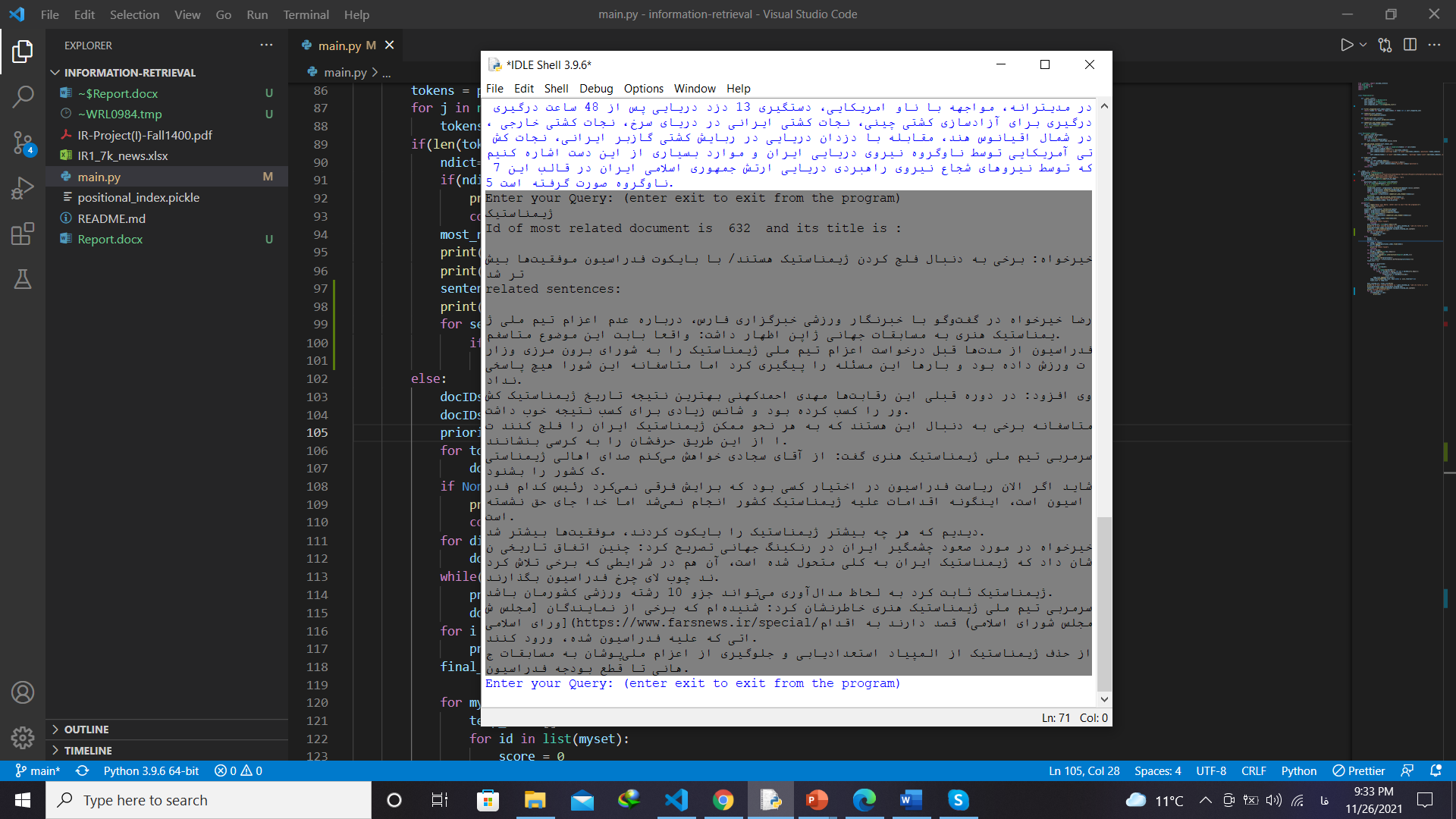
ب) متداول دو کلمه ای: دانشگاه امیرکبیر

مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. زیرا که در متن مسئول نهاد رهبری در دانشگاه امیرکبیر صحبتی کرده است و این مورد به دانشگاه امیرکبیر مرتبط است. و ترتیب رعایت شده است.

پ) متداول چندکلمه ای : جمهوری اسلامی ایران

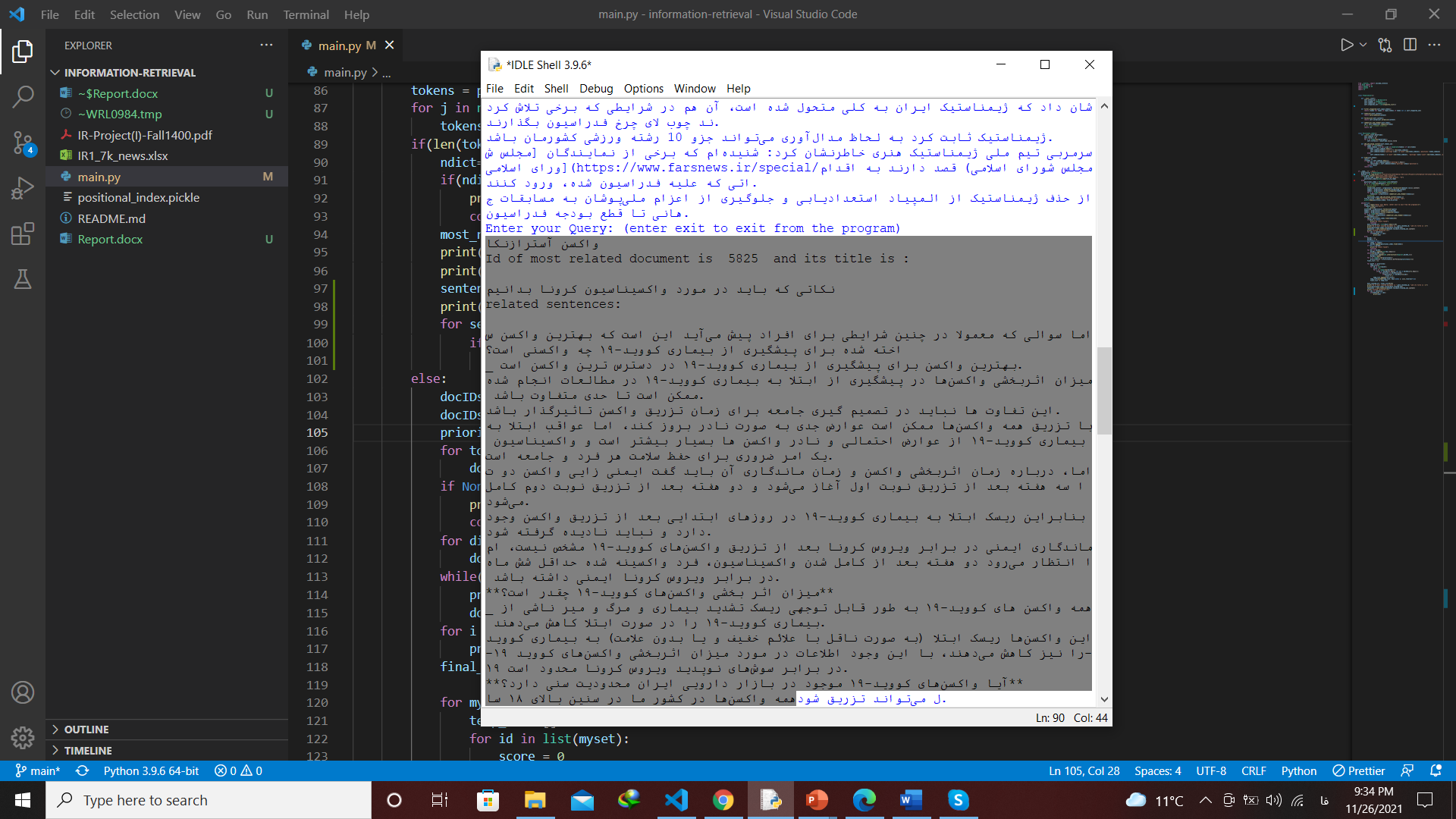


مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. زیرا که متن در ارتباط با جمهوری اسلامی ایران و ارتش آن می باشد و تعداد زیادی از این پرسمان با همان ترتیب در متن تکرار شده است و سیستم به خوبی توانسته این متن را در اولویت قرار دهد.

ت) دشوار و کم تکرار تک کلمه ای: ژیمناستیک

مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. متن کاملا در ارتباط با ورزش ژیمناستیک می باشد.

ث) دشوار کم تکرار چند کلمه ای: واکسن آسترازنکا



مرتبط است و سیستم به خوبی توانسته سند را بازیابی کند. متن در رابطه با واکسیانسون در زمان کرونا می باشد که واکسن آسترازنکا نیز یکی از این واکسن ها می باشد و کاملا مرتبط است و در متن با همان ترتیب تکرار شده است.