حسن کاظمی طهرانی

9629041

گزارش فاز دوم پروژه بازیابی اطلاعات

در این فاز ما به جای استفاده از مکان کلمات برای رتبه بندی، از شاخص tfidf استفاده کردیم. قسمت هایی از کد که تغییر کرده است شامل :

محاسبه tf و idfو tfidf برای هر کلمه در هر سند در مرحله indexing.

محاسبه طول بردار tfidf که بعدا برای محاسبه شباهت کسینوسی لازم می شود.

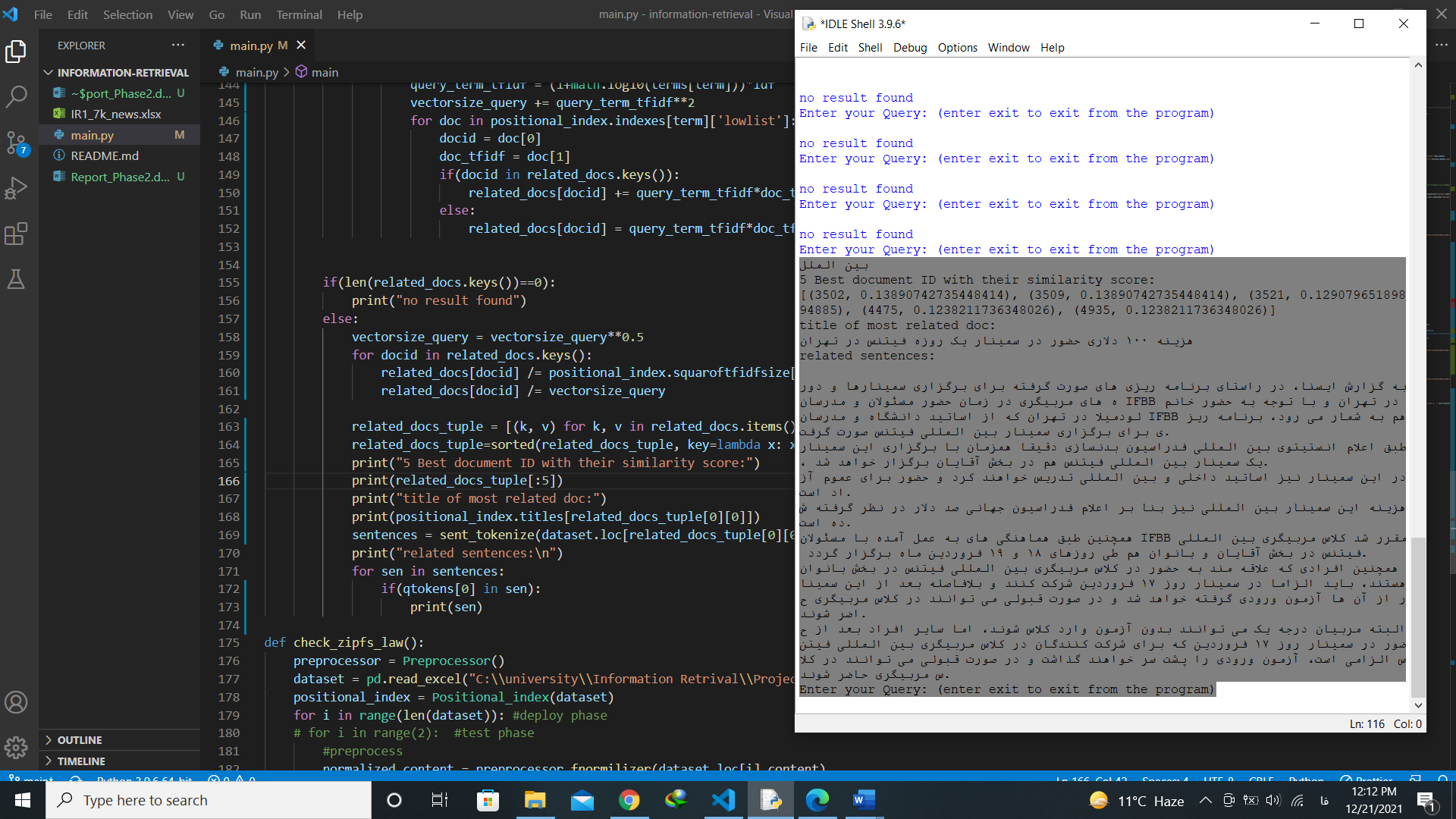
مرتب سازی سند های هر کلمه بر اساس tfidf برای درست کردن championlist.

تغییر در پیدا کردن اسناد که در این فاز به کمک محاسبه بردار tfidf و محاسبه شباهت کسینوسی انجام پذیرفته است.

گزارش های خواسته شده: ( برای مقایسه دقیقا همان کلمات مورد استفاده در فاز 1 استفاده شده است )

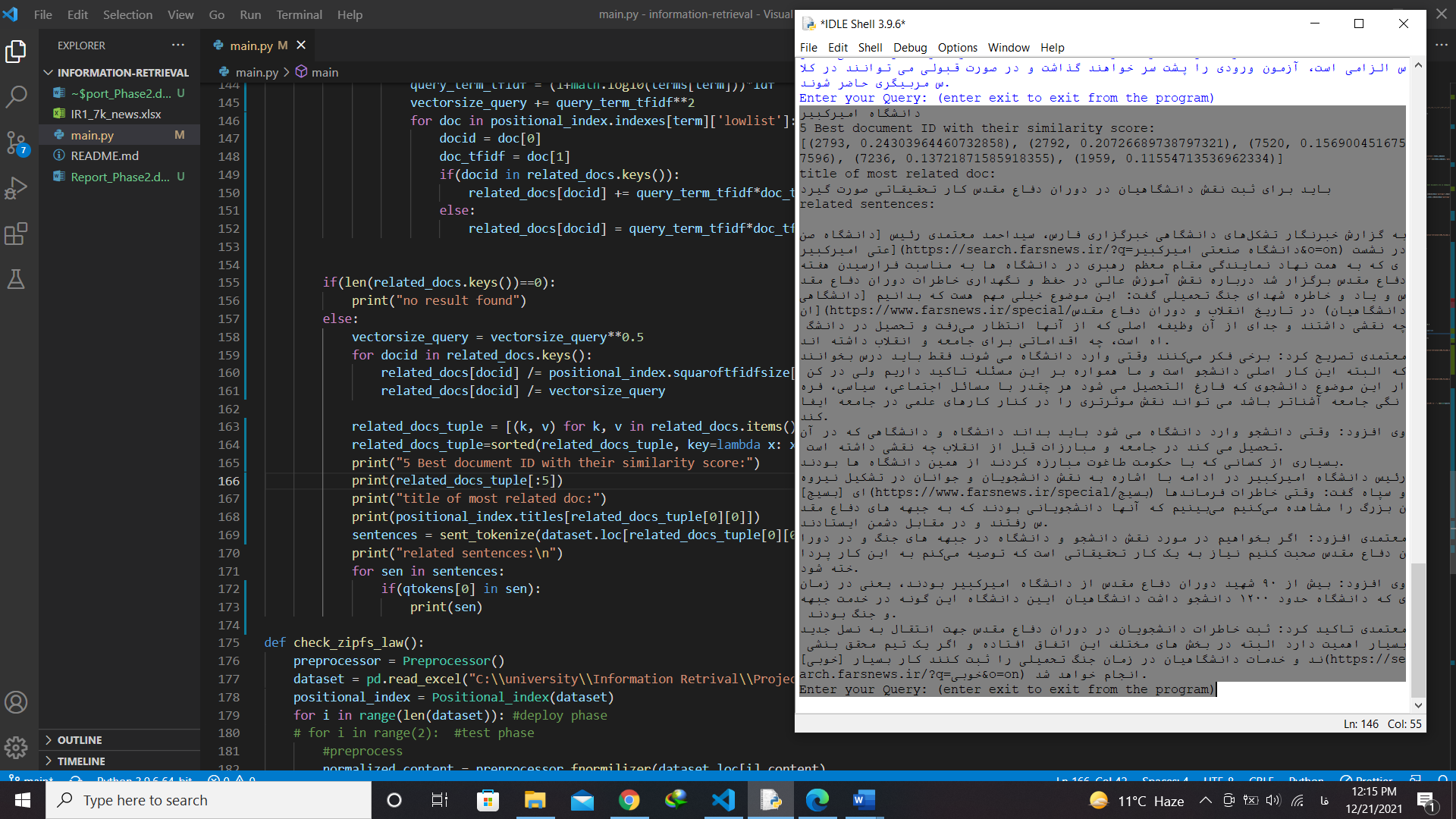
1- پاسخ به پرسمان در حالت های زیر:

الف ) یک پرسمان از کلمات ساده و متداول تک کلمه ای: (بین الملل)



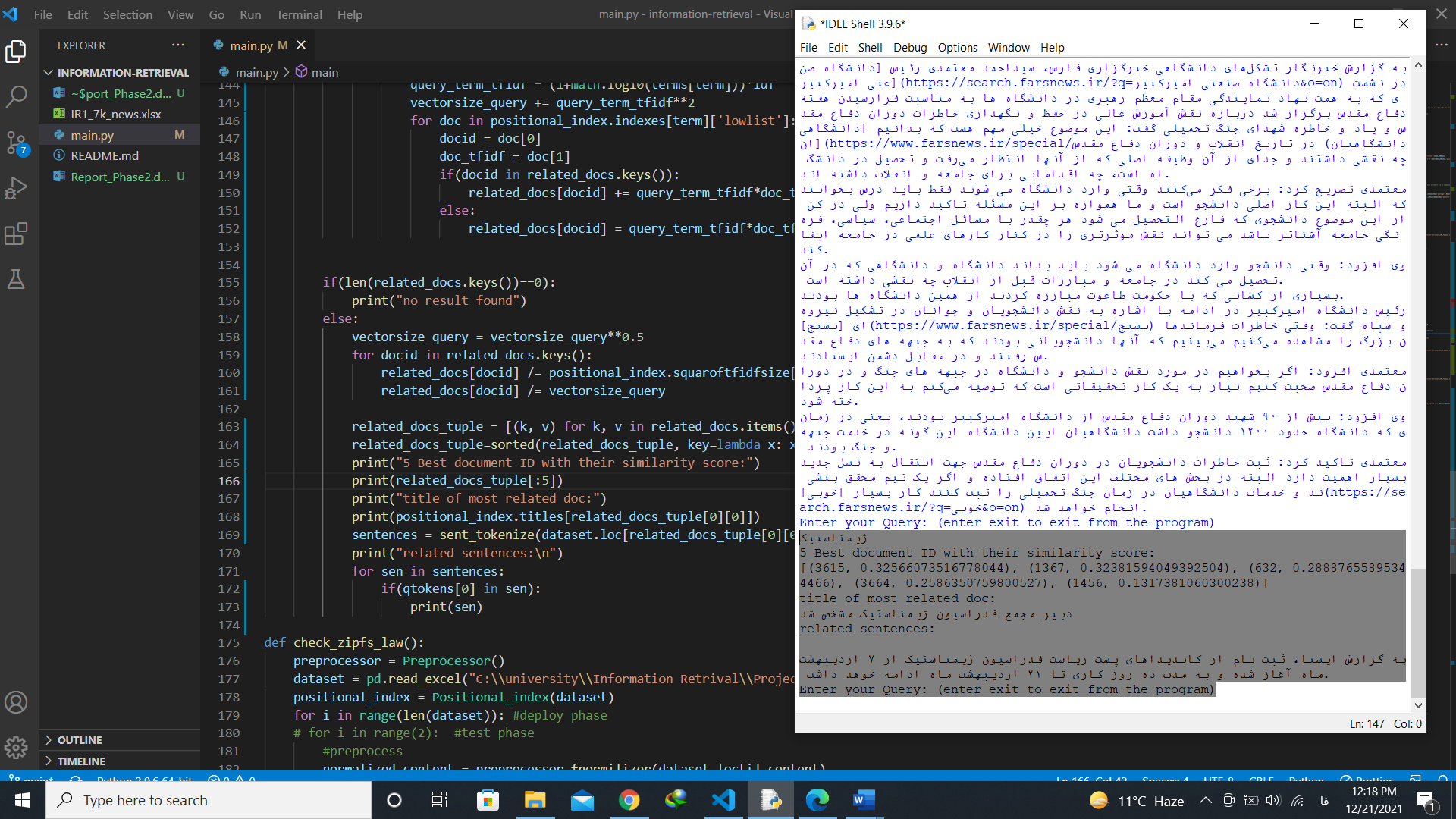
بله مرتبط است و در این فاز اسنادی پیدا می شود که تعداد بیشتری کلمه منحصر به فرد از کوئری را دارا می باشند. این خبر واکشی شده مربوط به یک سمینار بین المللی با حضور مربی های بین المللی می باشد.

ب) یک پرسمان از عبارات ساده و متداول چند کلمه ای ( دانشگاه امیرکبیر)



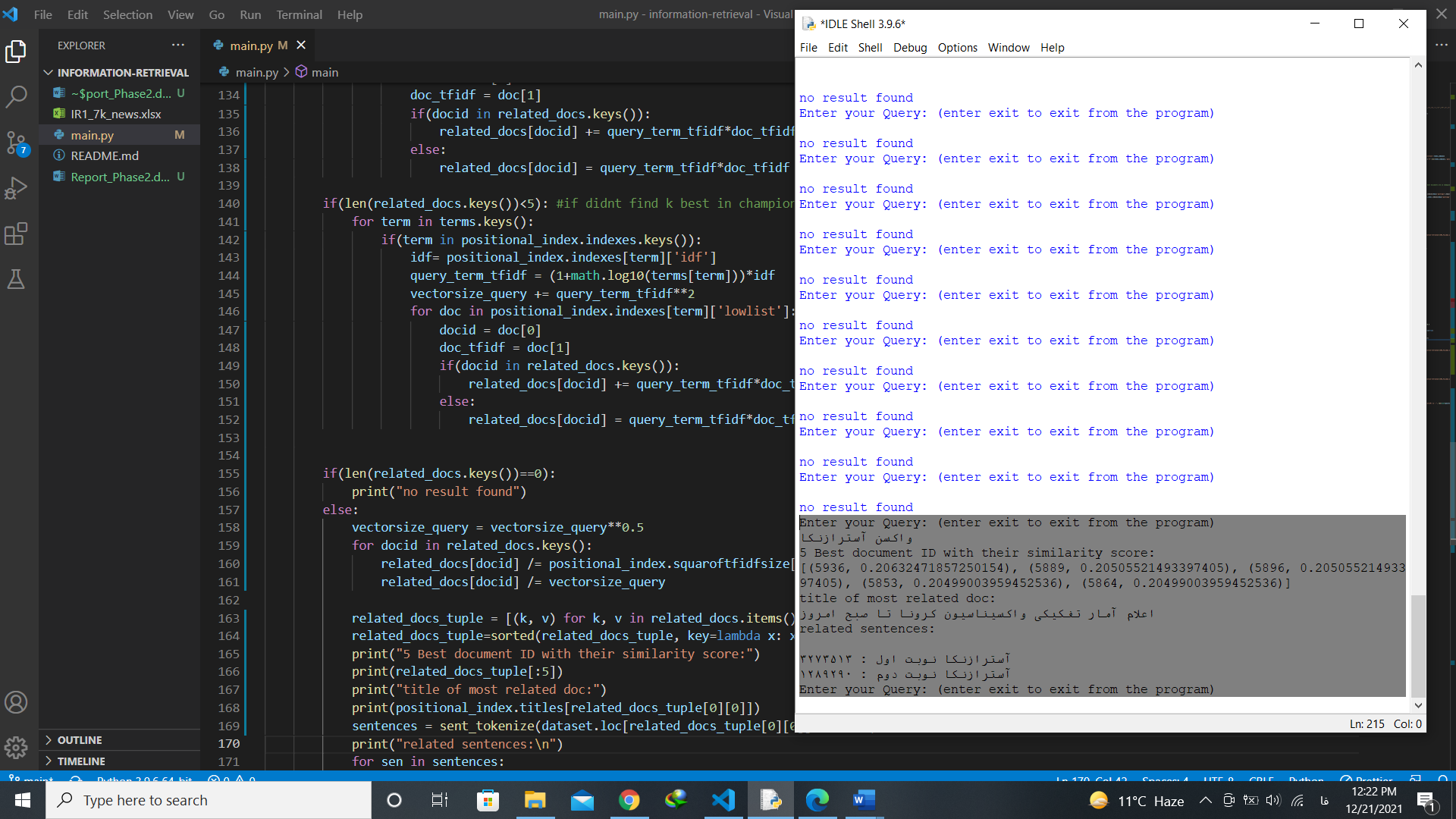
بله مرتبط است. خبر حاوی تعداد زیادی جمله حاوی "دانشگاه امیرکبیر" می باشد و خبر واکشی شده مربوط به سخنرانی رئیس دانشگاه امیرکبیر می باشد که کاملا مرتبط می باشد.

پ) یک پرسمان دشوار و کم تکرار تک کلمه ای ( ژیمناستیک )



بله مرتبط است. خبر کاملا مربوط به رشته ژیمناستیک می باشد که دبیر فدراسیون آن در خبر تغییر یافته است. گرچه این خبر تنها یک بار کلمه ژیمناستیک آمده است ولی به دلیل idf بالای این کلمه این خبر مرتبط شده است.

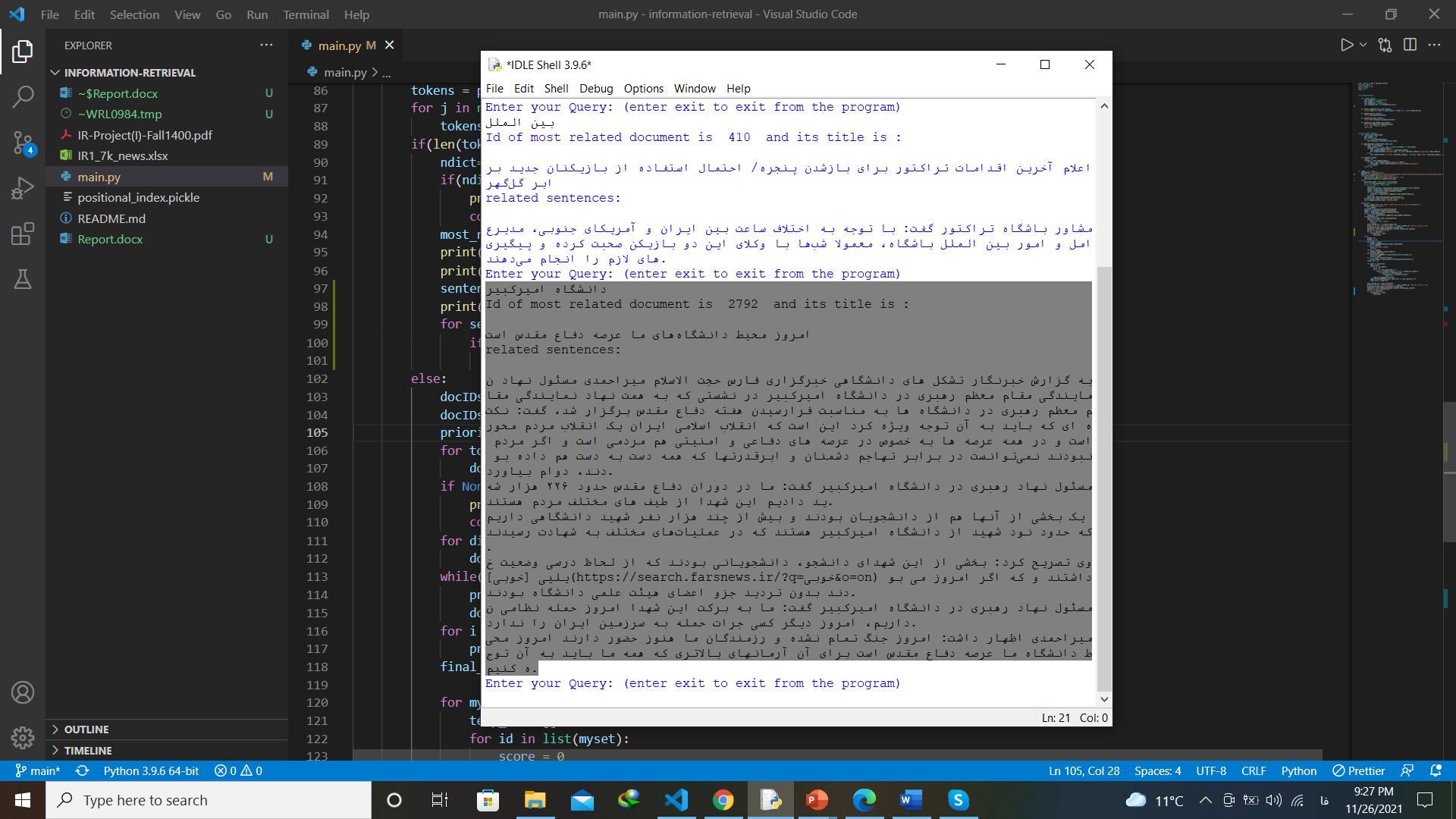
ت) یک پرسمان دشوار و کم تکرار چند کلمه ای (واکسن آسترازنکا)



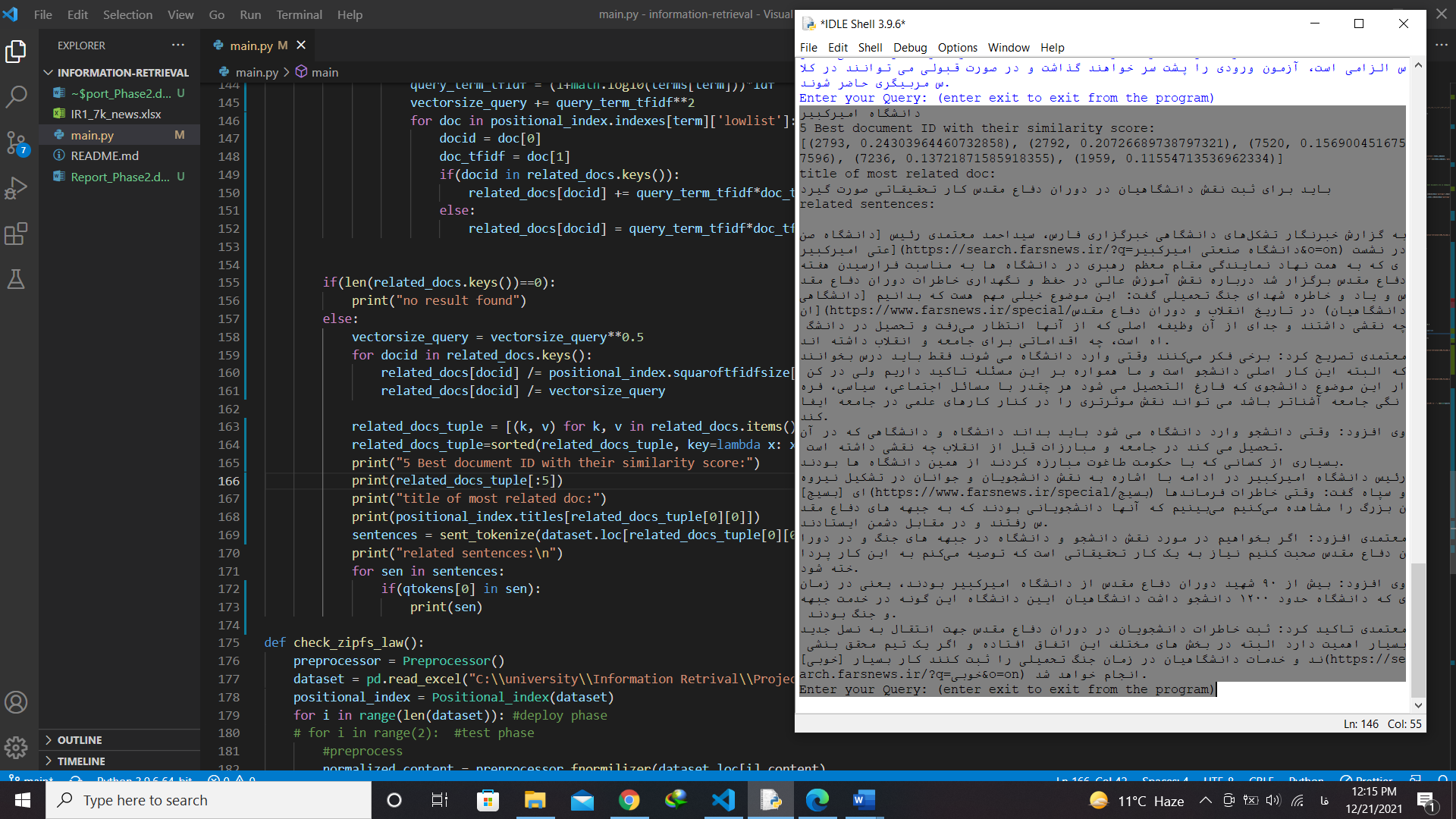
بله مرتبط است. خبر مربوط به امار واکسیناسیون کرونا می باشد که نوبت های واکسن آسترازنکا نیز در متن خبر آمده است.

2- مقایسه با فاز 1:

ب)

در فاز 1: 

در فاز2:

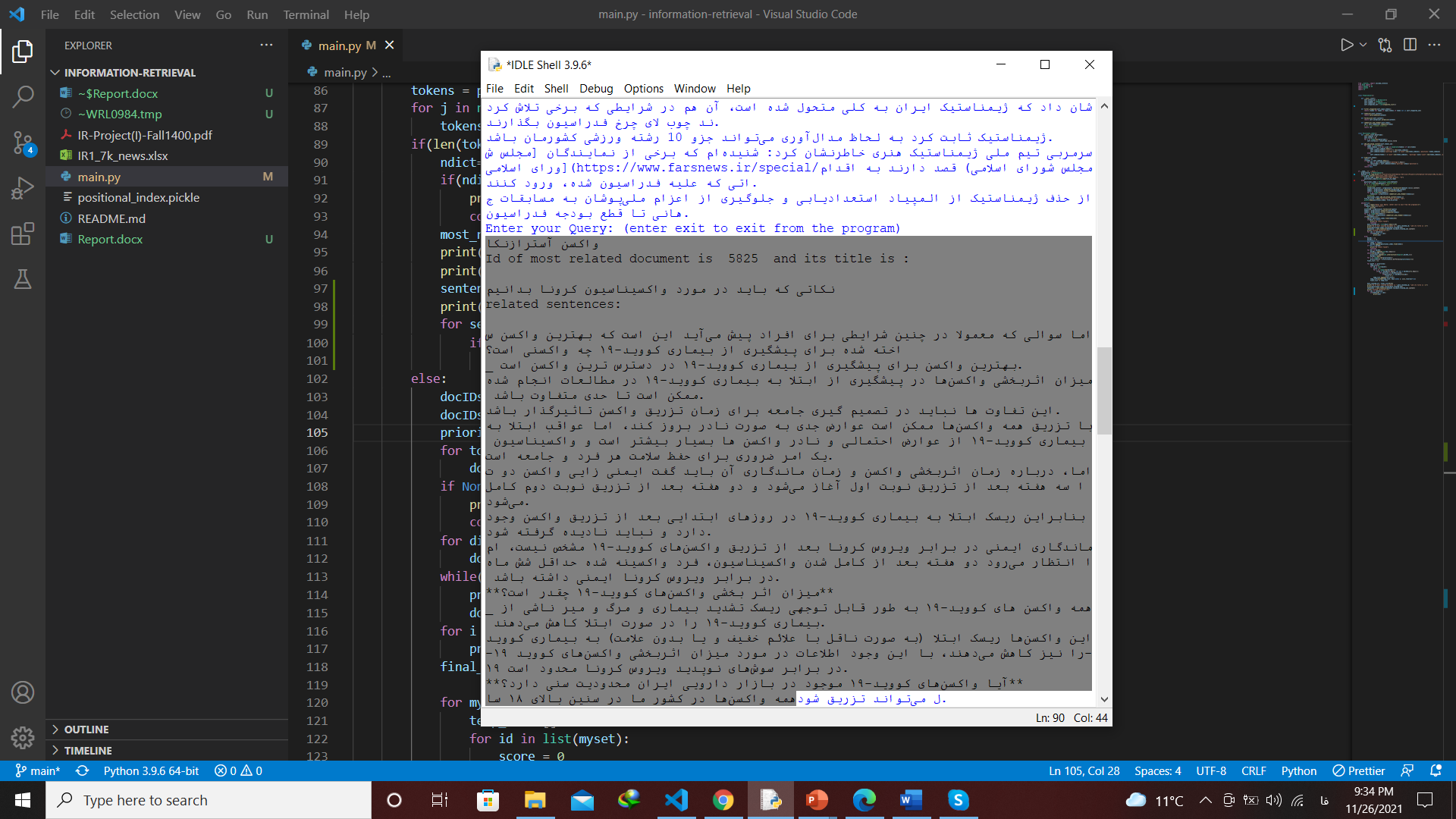


هر دو سند واکشی شده مرتبط می باشد. اما نکته حائز اهمیت این می باشد که در سند واکشی شده در موتورجستجو فاز 2، کلمه ی دانشگاه و امیرکبیر بدون توجه به مکان بسیار زیاد تکرار شده است و به همین علت رتبه این سند بالا آمده است. اما در فاز 1، کمتر این کلمات تکرار شده است اما جایگاهشان در بیشتر مواقع در کنار هم می باشد زیرا که در فاز 1، رتبه بندی براساس جایگاه کلمات نیز می باشد.

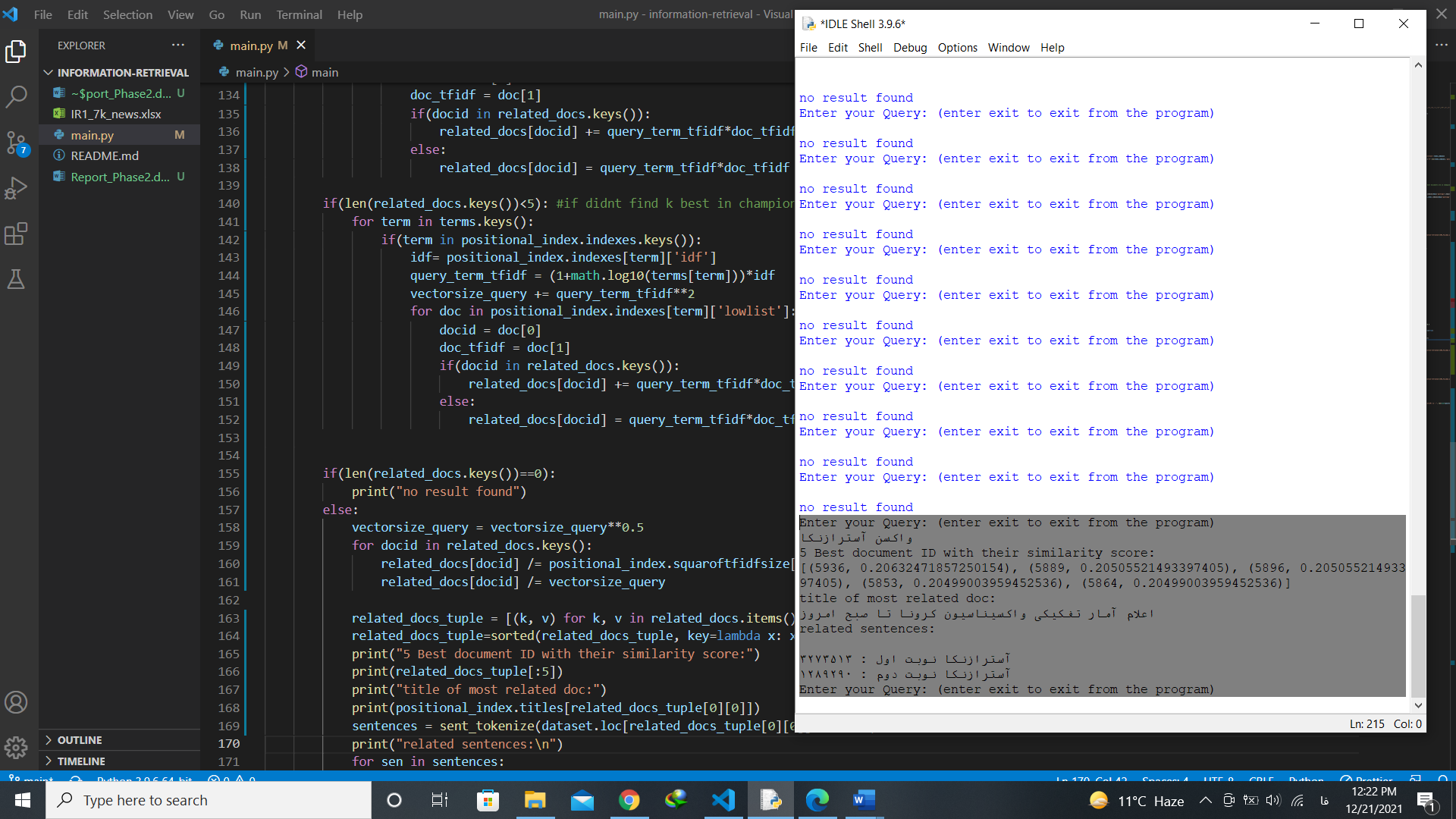
به همین علت سند های واکشی شده متفاوت می باشد اما هر دو سند مرتبط است.

ت)

فاز 1:



فاز 2:



هر دو سند مرتبط است. اما سند واکشی شده در فاز 2 مرتبط تر می باشد. در فاز 1 سندی در رابطه با واکسیناسون کرونا واکشی شده که به واکسن آسترازنکا نیز اشاره ای شده است. اما در فاز دوم به کلمه آسترازنکا که دارای idf بیشتری می باشد اهمیت زیادی داده شده و خبرنوبت دهی این واکسن خاص واکشی شده است زیرا که در فاز دوم رتبه بندی به tfidf توجه بیشتری دارد گرچه کلمه آسترازنکا دقیقا بعد از واکسن نیامده است.