به نام **خدا**

""گزارش فاز دوم پروژه بازیابی اطلاعات""

محمدرضا قادری 9627057

در این فاز ما به جای استفاده از Posting list مربوط به هر کلمه برای ارتباط درخواست‌ها استفاده می‌کردیم ولی در اینجا می‌خواهیم با استفاده از tfidf برای شاخص استفاده کنیم.

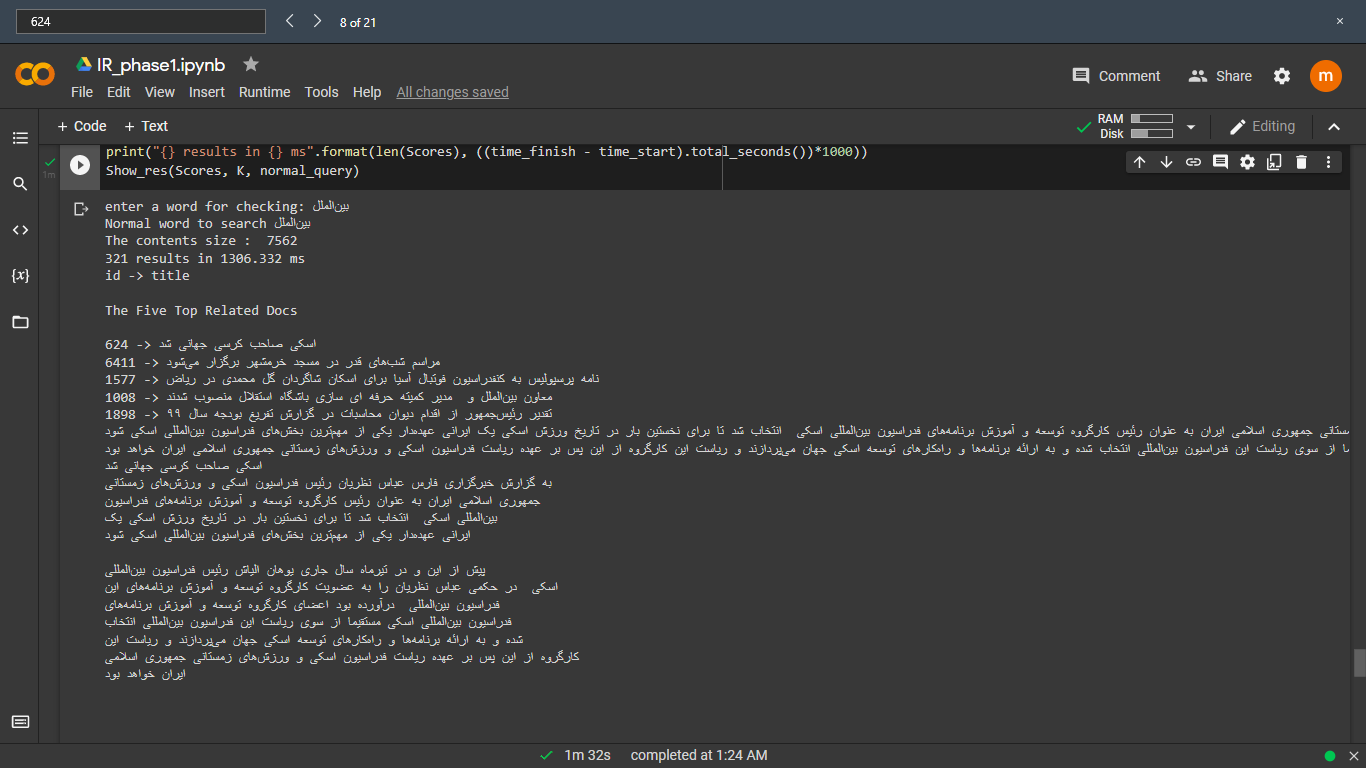
ما برای هر کلمه برای هر سند tf , idf , tfidf را محاسبه می‌کنیم و برای محاسبه شباهت نیز از روش کسینوسی استفاده خواهیم کرد پس همین جا که tfidf حساب می‌کنیم طولش را نیز برای هر سند استفاده می‌کنیم. حالا ما میتوانیم با استفاده از همین tfidfها میتوانیم Champion List تشکیل دهیم ، در اینجا ما Index Elimination لازم نداریم چرا که از روش TaaT استفاده می‌کنیم خودش این قضیه رو برای ما حل می‌کند. برای Champion list نیز هم چندین روش وجود دارد یکی از روش‌ها استفاده از یک ترشهولد برای اینکه در این لیست از اون کلمات استفاده کنیم و روش دیگه نیز می‌توانیم با سورت کردن همه لیست‌ها بر حسب tfidf در ابتدا و سپس بر حسب شماره داک می‌تونیم میتواند برای آن استفاده کنیم ازش برای همین که حساب کردن ترشهولد نیازمند این است که کلیه مقادیر رو بشناسیم اعم از میانگین و ماکسیمم و شاید هم مینیمم که بتوانیم یک محدوده‌ی تیjdsdasd‌‌ای برای اینکه tfidf مشخص کنیم پس ما از روش اول استفاده می‌کنیم. گزارش از ما خواسته که با این کار ابتدا Query خواسته شده را با این روش با doc ها مقایسه کنیم و شباهت رو بدست بیاریم و سپس با روش اول که در فاز اول داشتیم مقایسه انجام دهیم.یبییس

از همان کلمات که در فاز اول استفاده شده بود برای مقایسه استفاده شد.

1) پرسمان‌های برای tfidf

*الف)* یک پرسمان از کلمات ساده و متداول تک ‌کلمه‌ای

*بین الملل*

**

*با توجه به اینکه برای ما موثر هست اندازه متن می‌بینیم در بین متن‌های بدست آمده تاثییر خودش را گذاشته هست و بله این سند مرتبط هست و تعداد کلمات منحصر به فرد زیادی به کار رفته در کوئری ما هست به نسبت اندازه کل سند.*

*ب)یک پرسمان از عبارات ساده و متداول چندکلمه‌ای*

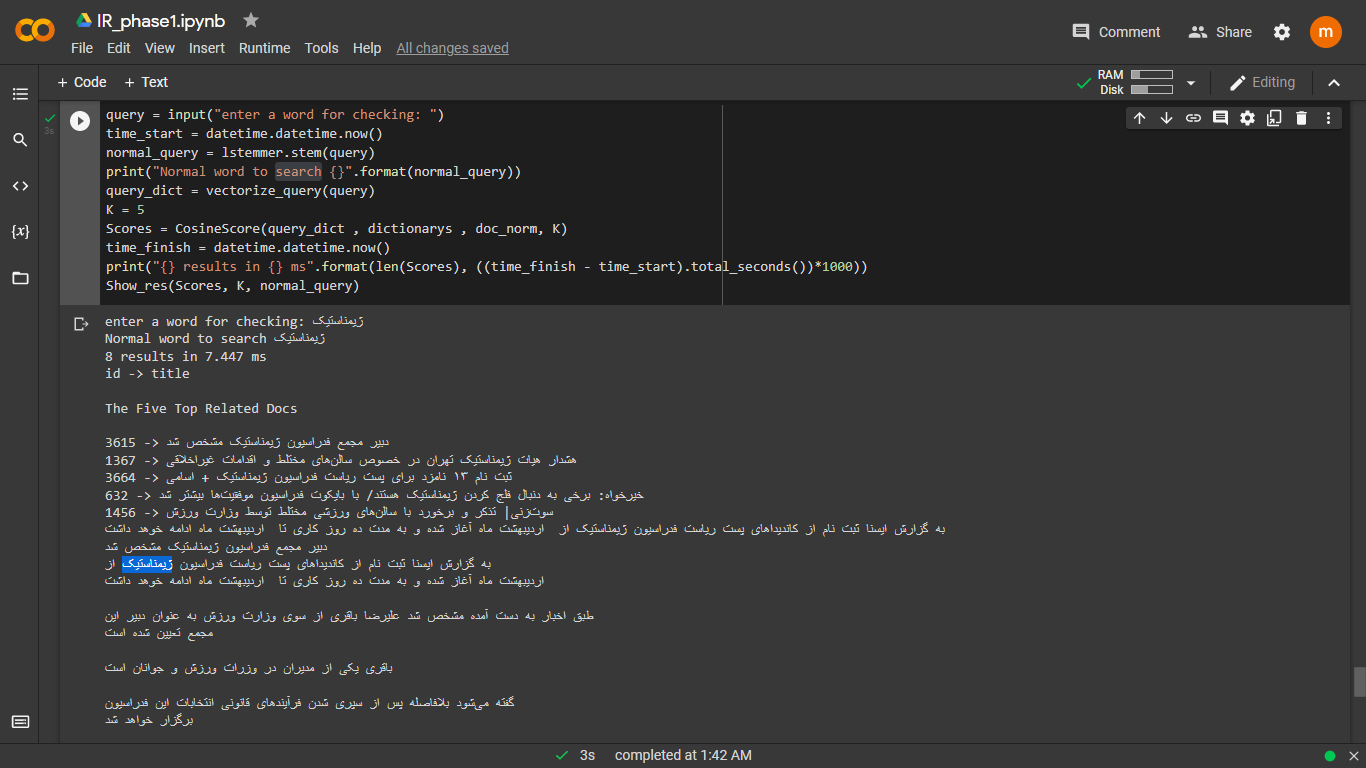
*دانشگاه امیرکبیر*

**

*بله این نیز مرتبط هست خبر حاوی تعداد خوبی کلمه دانشگاه امیرکبیر هست و سند نیز مربوط به سخنان رییس دانشگاه می‌باشد.*

*پ) یک پرسمان دشوار و کم تکرار تک‌کلمه‌ای*

*ژیمناستیک*

**

*باز هم مرتبط است خبر منتشر شده درباره فدراسیون ژیمناستیک که در باره تغییر دبیر فدارسیون هست در اینجا باز هم تاثییر idf رو مشاهده می‌کنیم که به خاطر کم بودن متن و تکرار یک باره این متن مرتبط شده است.*

*ت) یک پرسمان دشوار و کم تکرار چندکلمه‌ای*

*واکسن آسترازنکا*

**

*بله این نیز مرتبط هست که به نوبت واکسن آسترازنکا مربوط می‌باشد.*

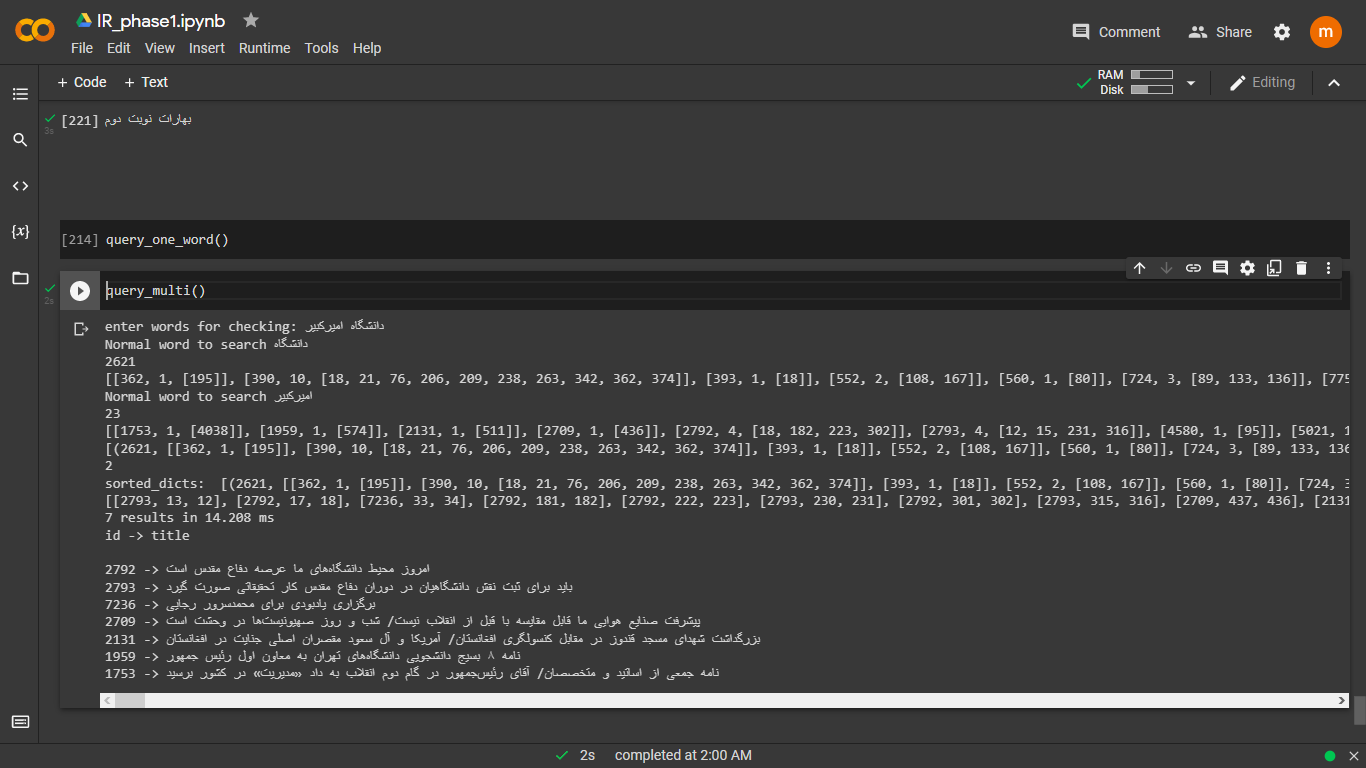
*برای مقایسه فاز 1 و 2*

*حالت ب)*

*فاز 2*

**

*فاز 1*

**

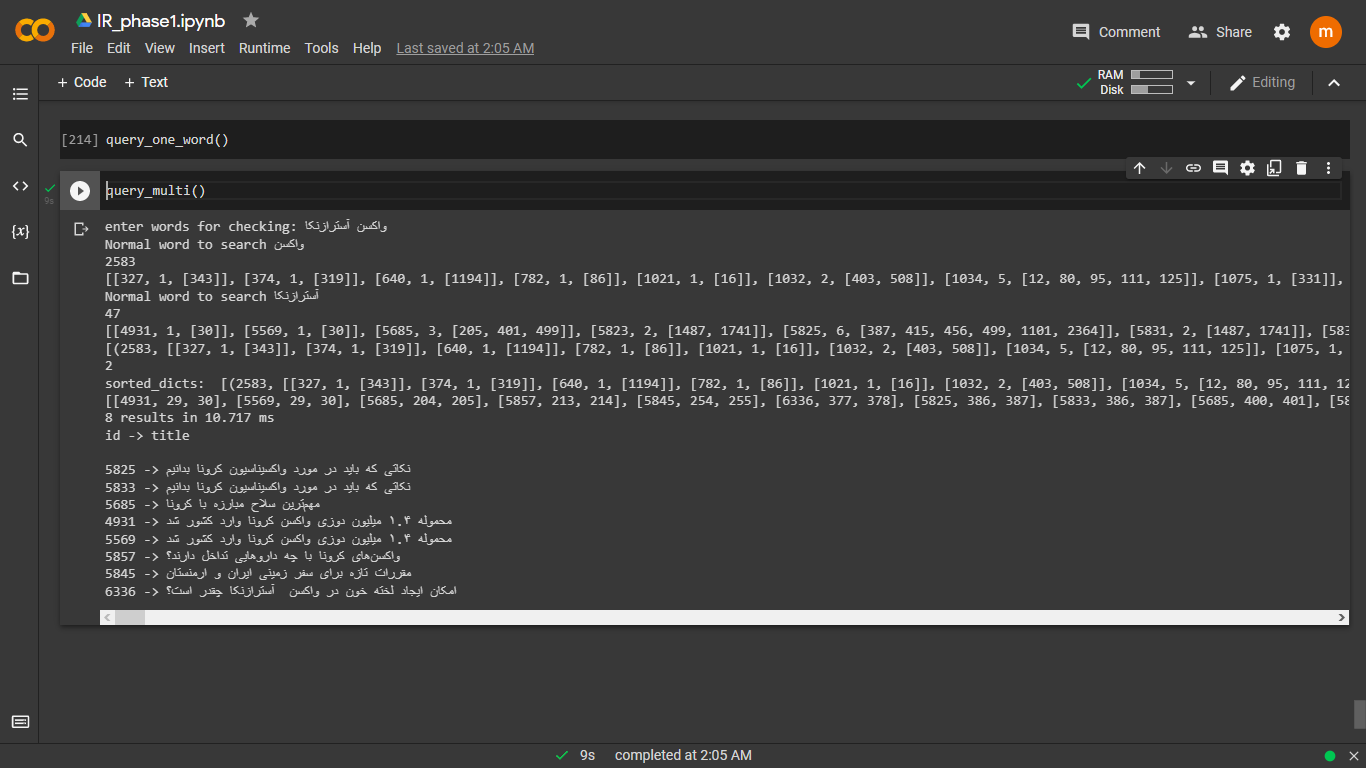
*همون طور که مشخص هست هر دو سند واکشی شده در صدر یکی و مرتبط هستند ولی در فاز دوم ما کلمه مورد نظر را بدون توجه به مکان این دو و فقط برحسب تکرار زیاد در صدر قرار داره در حالی که در فاز یک به دلیل اینکه این دو خیلی زیاد کنار هم اومدن مدنظر ما بوده.*

*حالت ت)*

*فاز 2*

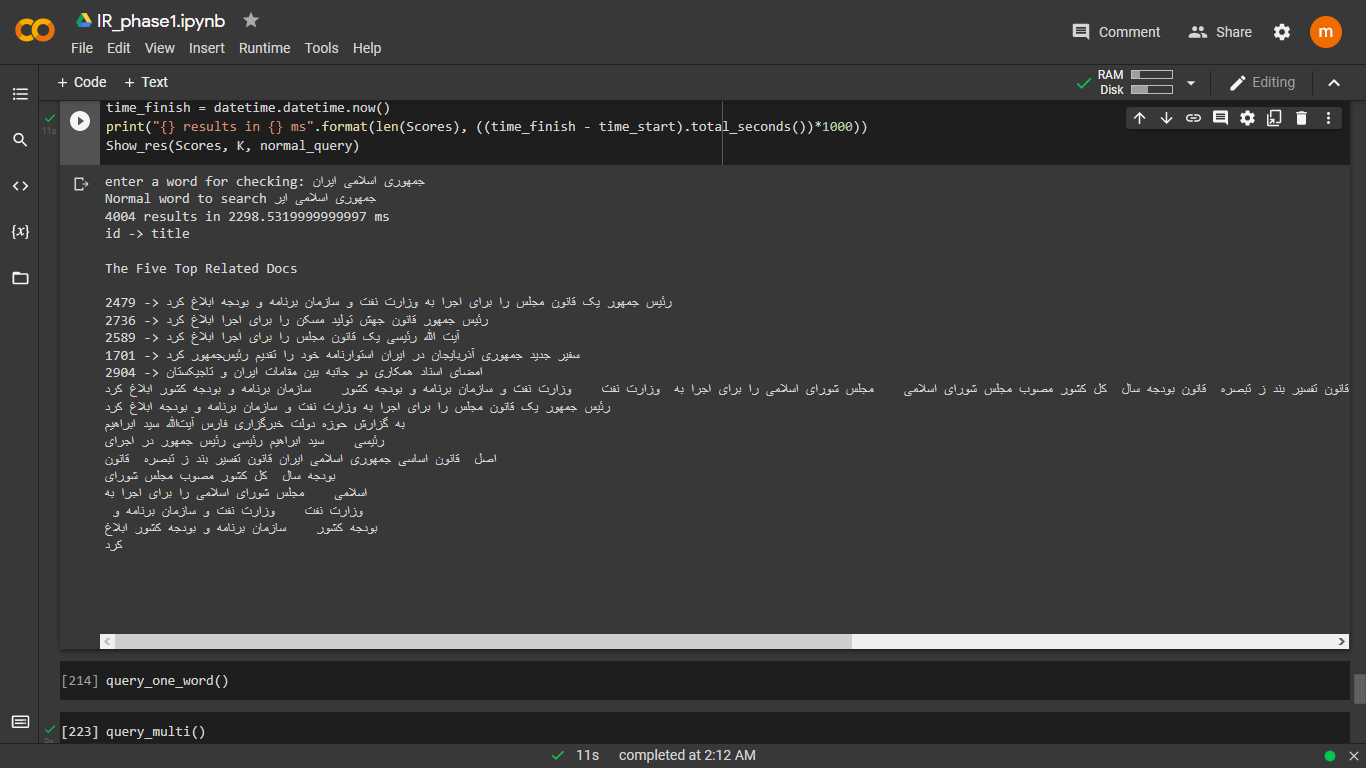
**

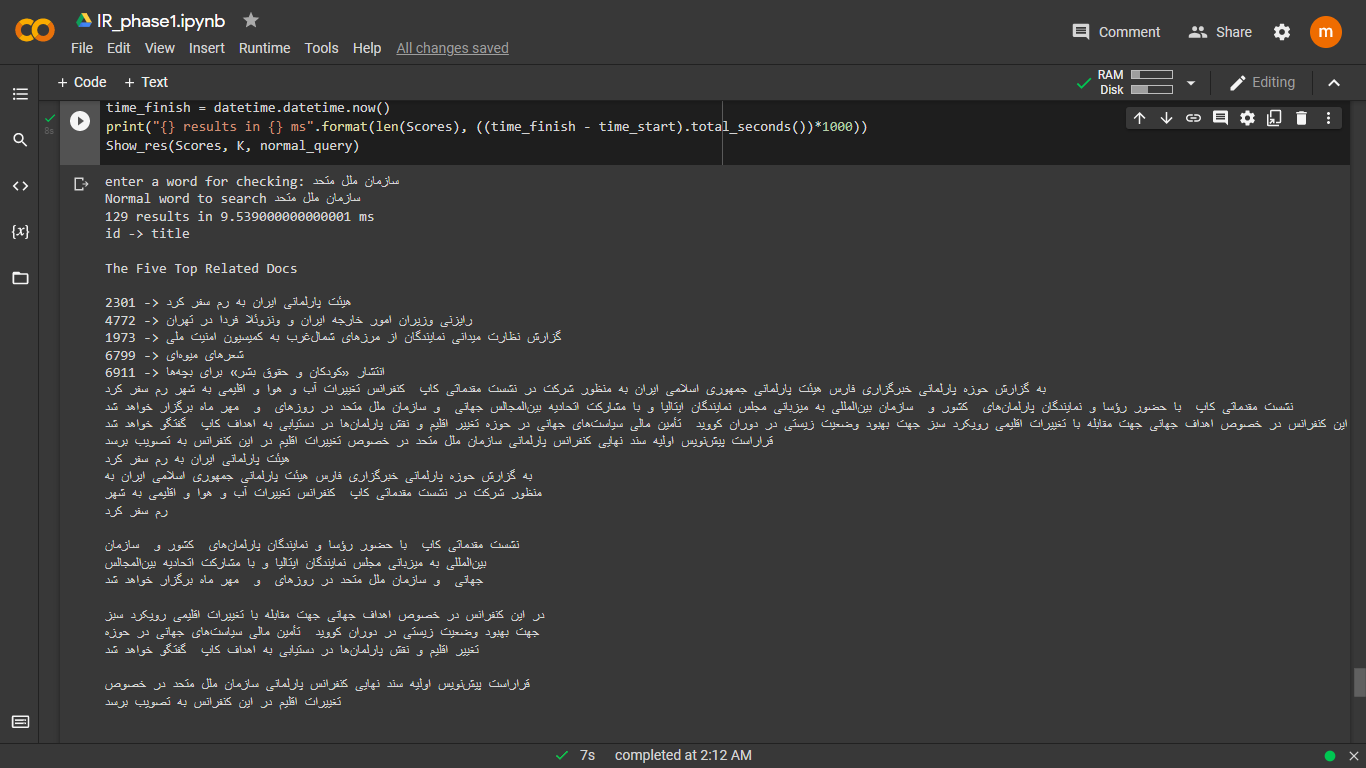
*فاز 1*

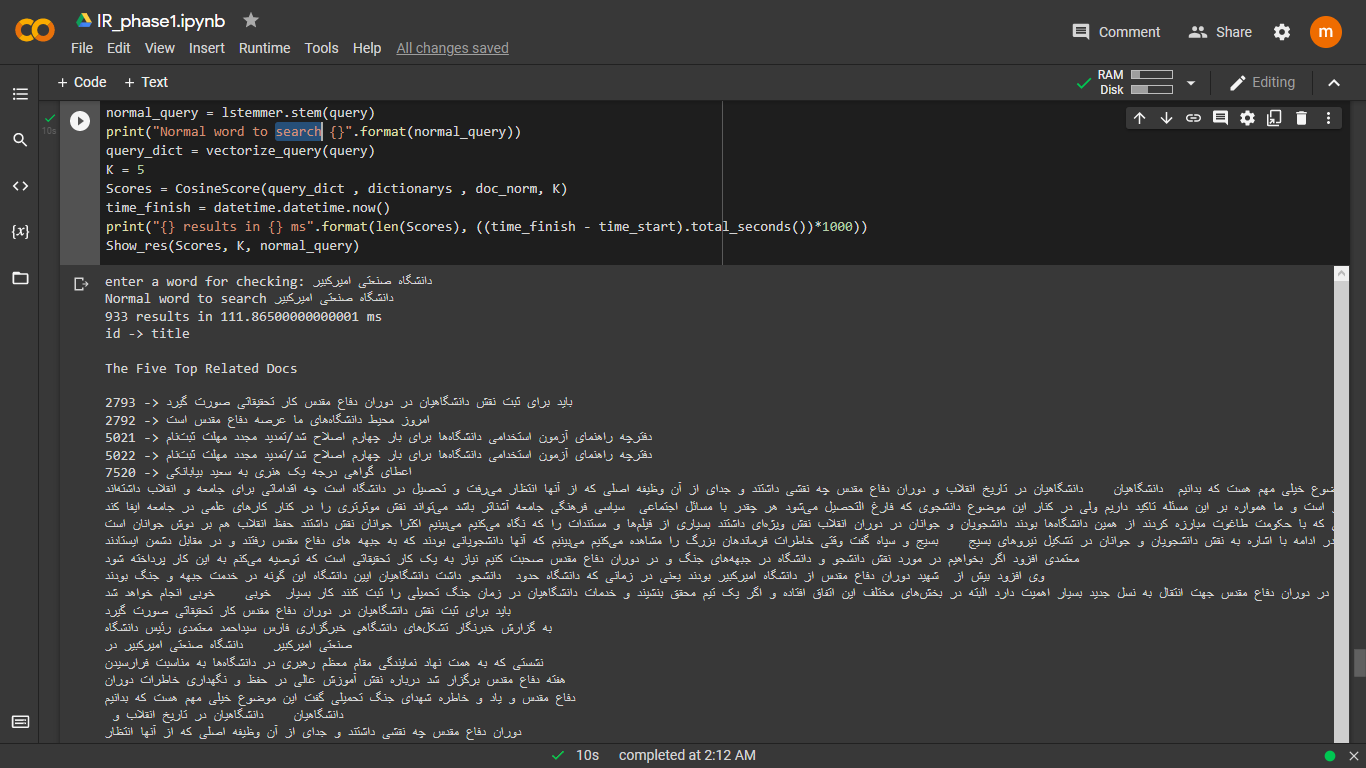
**

*هر دو سند مرتبط هستند اما نتیجه در حالت 2 مرتبط تر است در فاز 1 بیشتر به بودن دو کلمه توجه شده در حالی که در حالت دوم به این توجه شده است که کلمه آسترازنکا با توجه به این که خاص تر هست idf بیشتری دارد و تاثیر زیادی دارد و خبر در مورد نوبت دهی این واکسن آورده شده است و رتبه بالاتر نسبت به قبلی دارد.*

*کلمات دیگر که در فاز یک آمده بود نیز بررسی شدن که به همانگونه برای رتبه دهی بیشتر به مرتبط بودن آن سند دارد*

**

**

**

*که با توجه به content ها میتوان تشخیص داد واکشی ها بهتر و مرتبط تر شده است در کل.*