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.

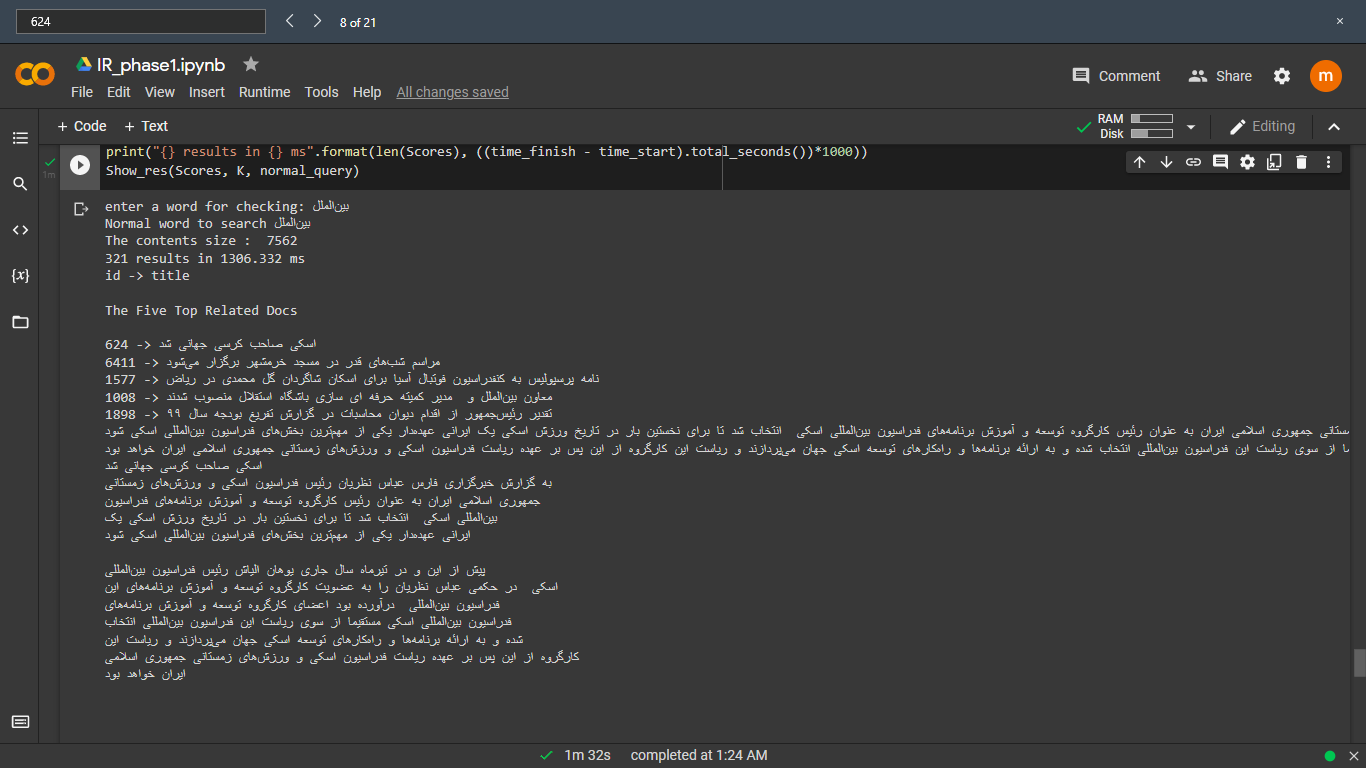
ما برای هر کلمه برای هر سند tf , idf , tfidf را محاسبه می‌کنیم و برای محاسبه شباهت نیز از روش کسینوسی استفاده خواهیم کرد پس همین جا که tfidf حساب می‌کنیم طولش را نیز برای هر سند استفاده می‌کنیم. حالا ما میتوانیم با استفاده از همین tfidfها میتوانیم Champion List تشکیل دهیم ، در اینجا ما Index Elimination 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*

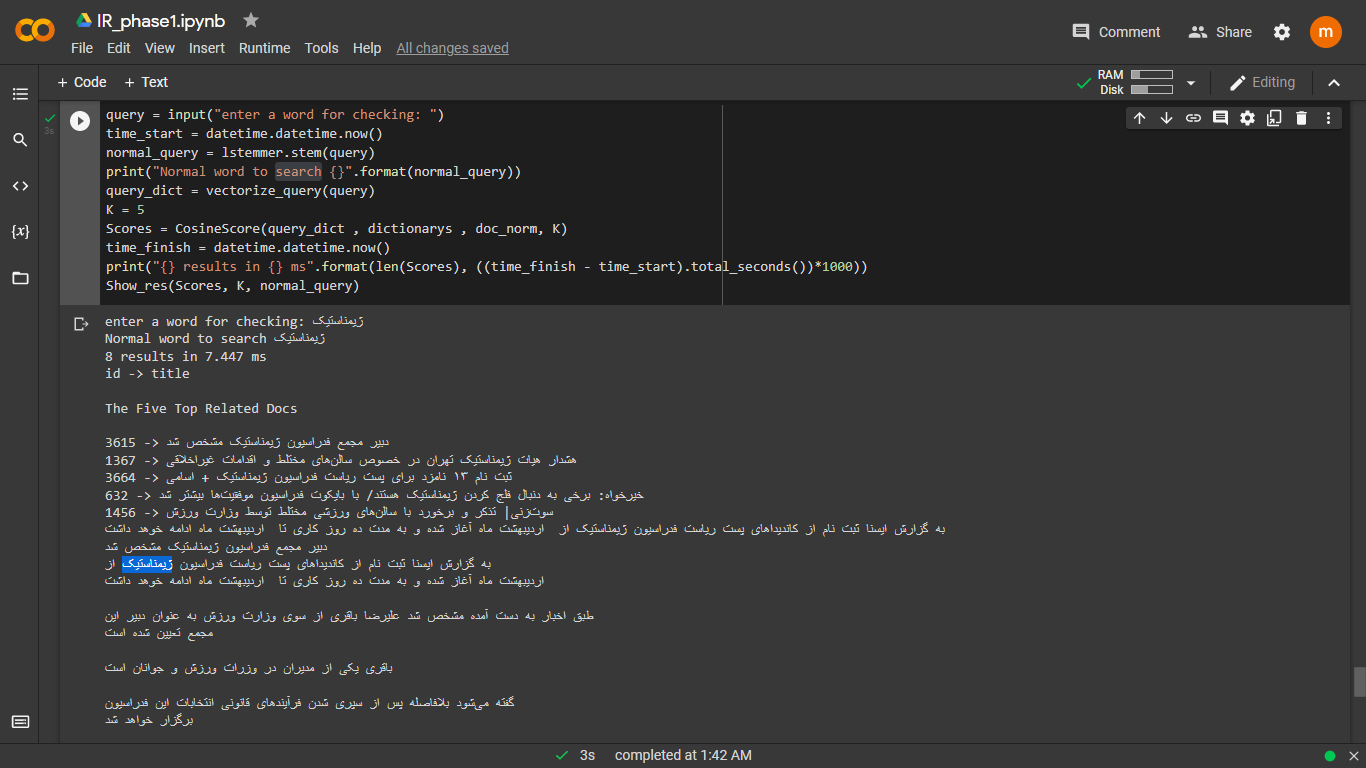
*دانشگاه امیرکبیر*

**

*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.*

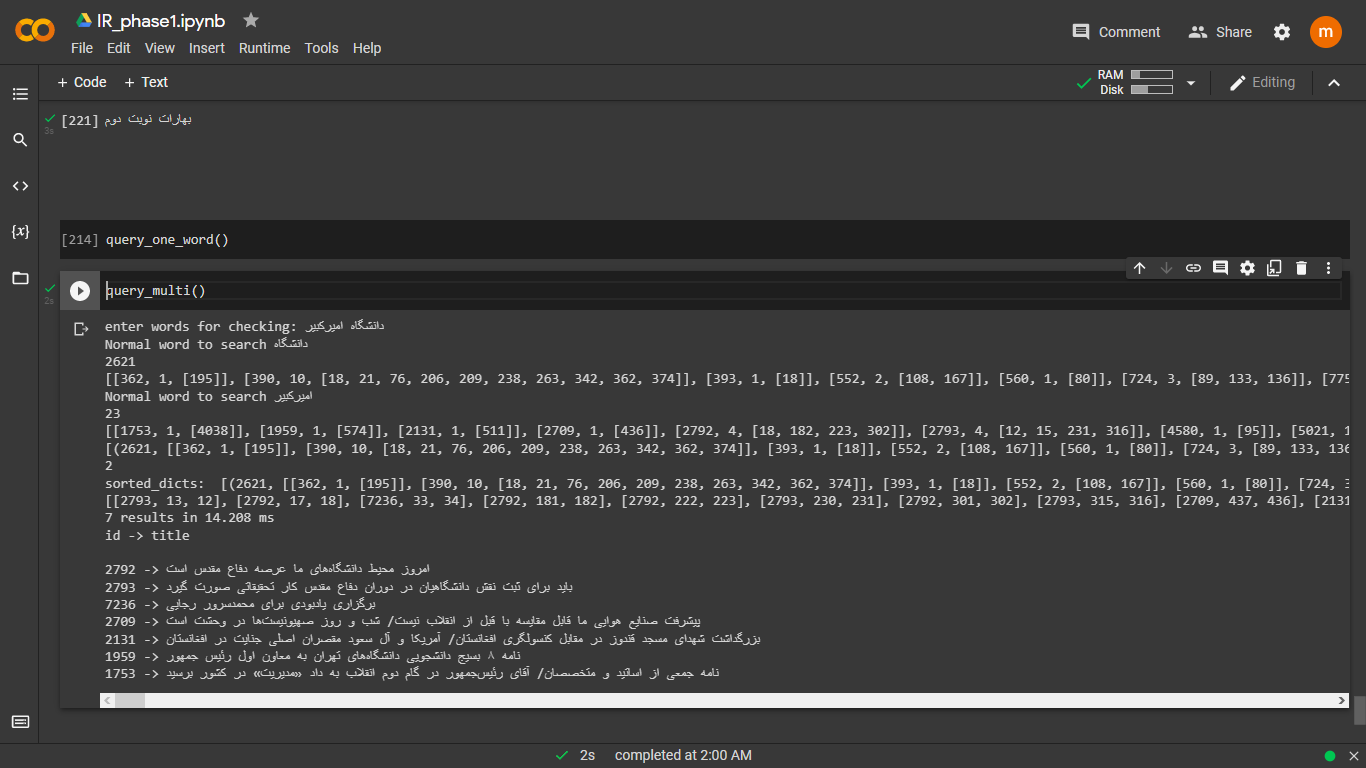
*برای مقایسه فاز 1 و 2*

*حالت ب)*

*فاز 2*

**

*فاز 1*

**

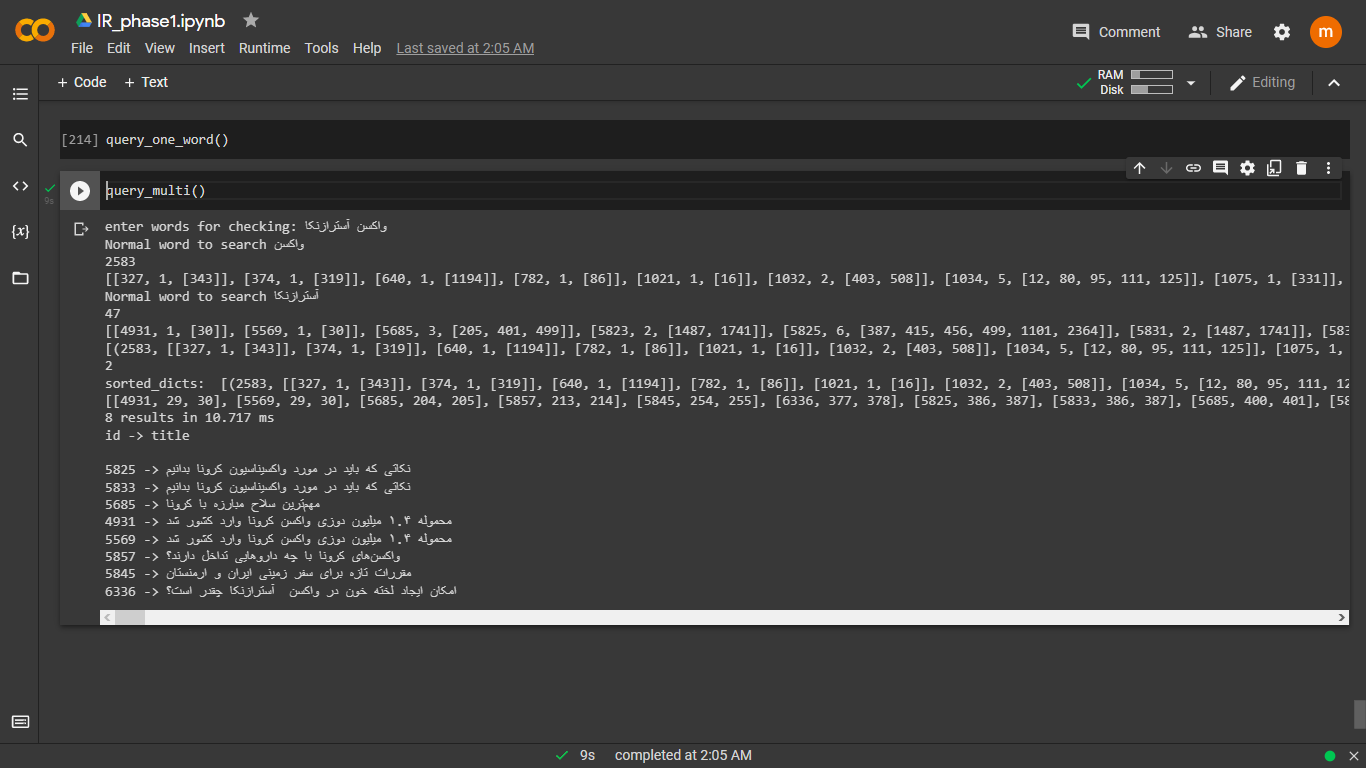
*همون طور که مشخص هست هر دو سند واکشی شده در صدر یکی و مرتبط هستند ولی در فاز دوم ما کلمه مورد نظر را بدون توجه به مکان این دو و فقط برحسب تکرار زیاد در صدر قرار داره در حالی که در فاز یک به دلیل اینکه این دو خیلی زیاد کنار هم اومدن مدنظر ما بوده.*

*حالت ت)*

*فاز 2*

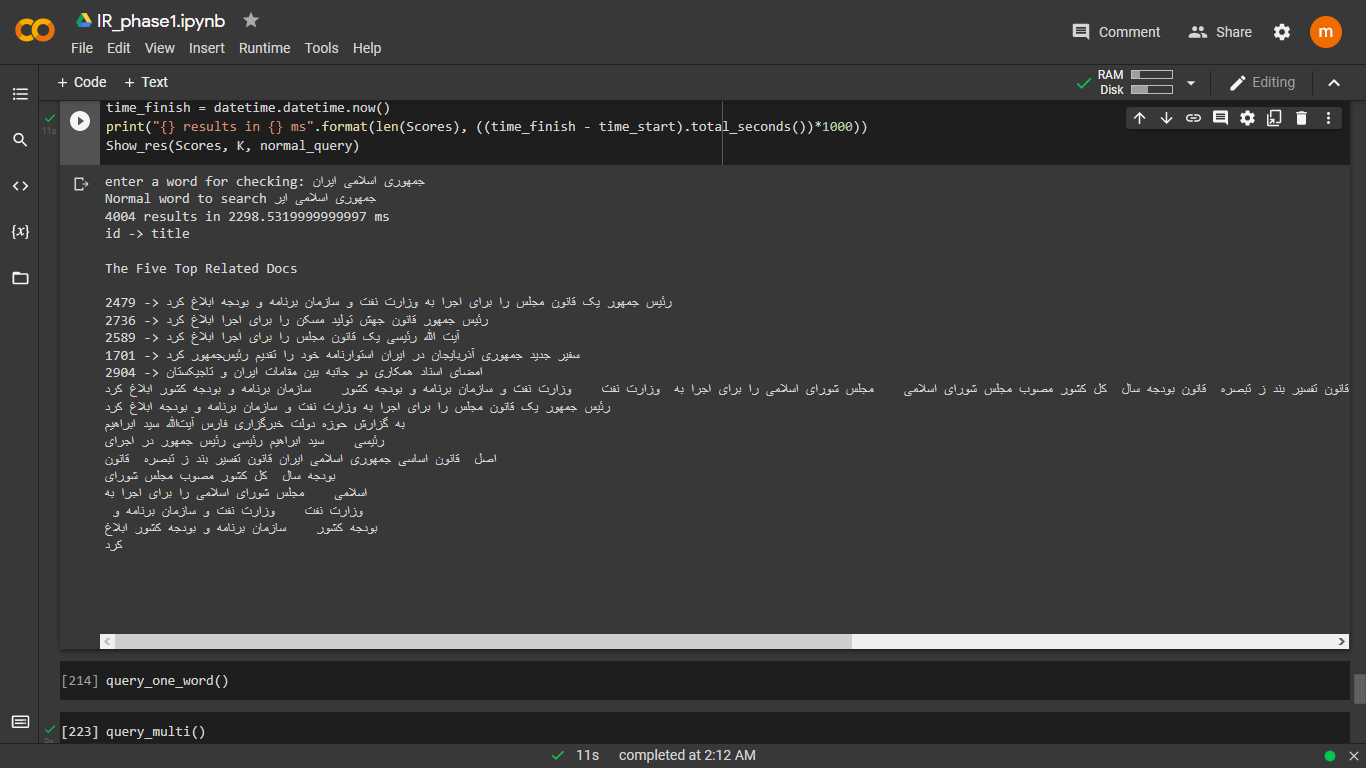
**

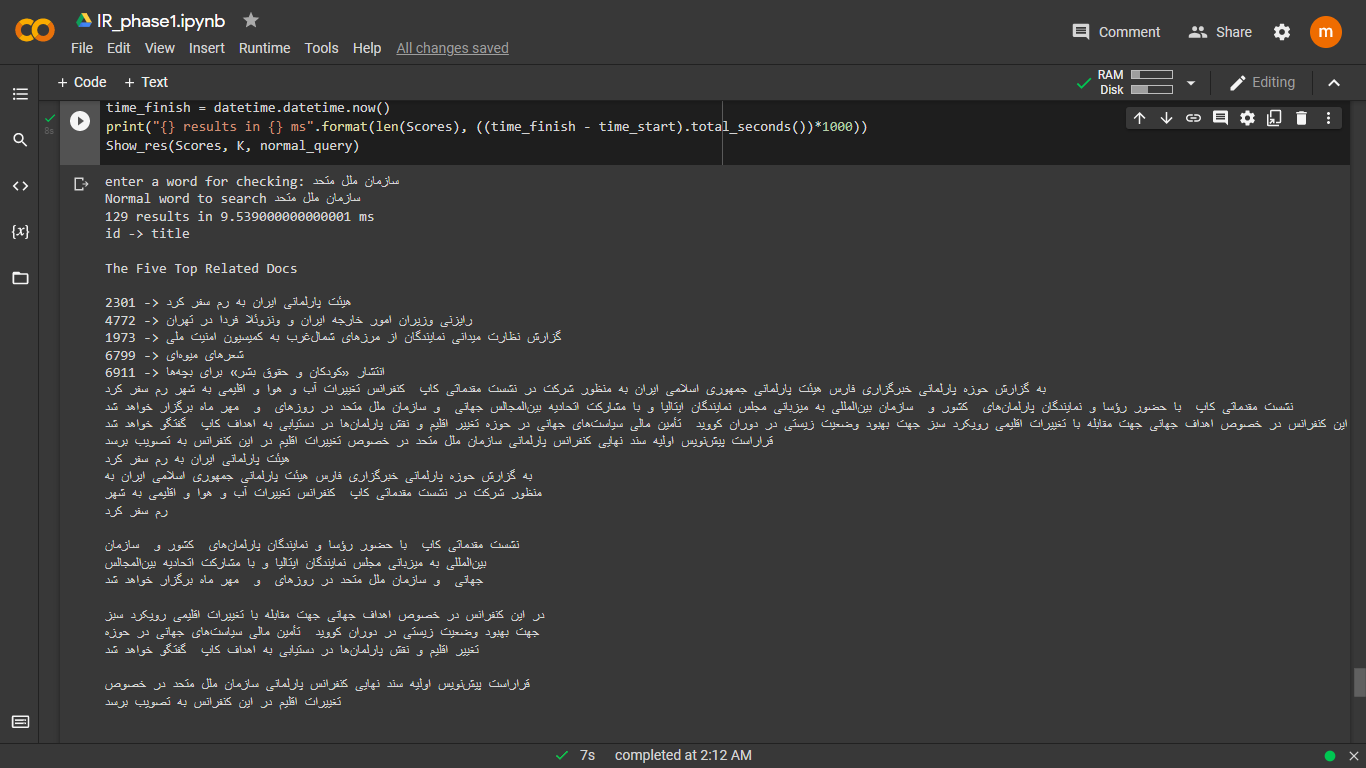
*فاز 1*

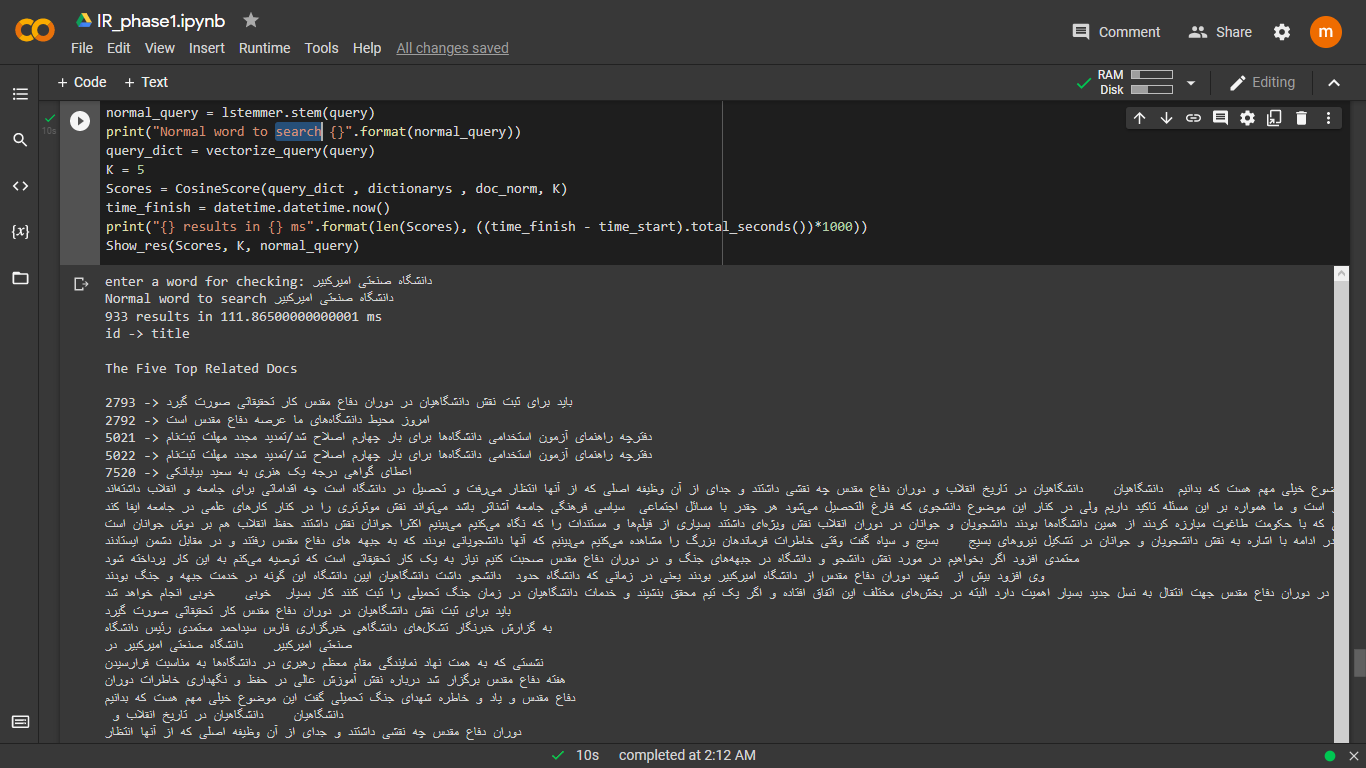
**

*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.*

*بخش امتیازی*

*در این بخش ابتدا ما دیتا‌های خود را مورد پردازش قرار می‌دهیم به این صورت که تمامی کلمات رو می‌دهیم تا* word vector *انها رو بدست بیاریم حال برای وزن دهی آنها از همان مقادیر tfidf خودمان که در همین فاز بدست آورده‌ایم استفاده می‌کنیم.*

*زمانی که می‌خواهیم doc embbeding رو بدست بیاریم می‌توانیم کلماتی که موجود نیست رو در نظر نگیریم(با try except)*

*حالا برای چند کوئری میایم ببینم مدلی که آموزش دادیم چه کرد*

*ژیمناستیک*

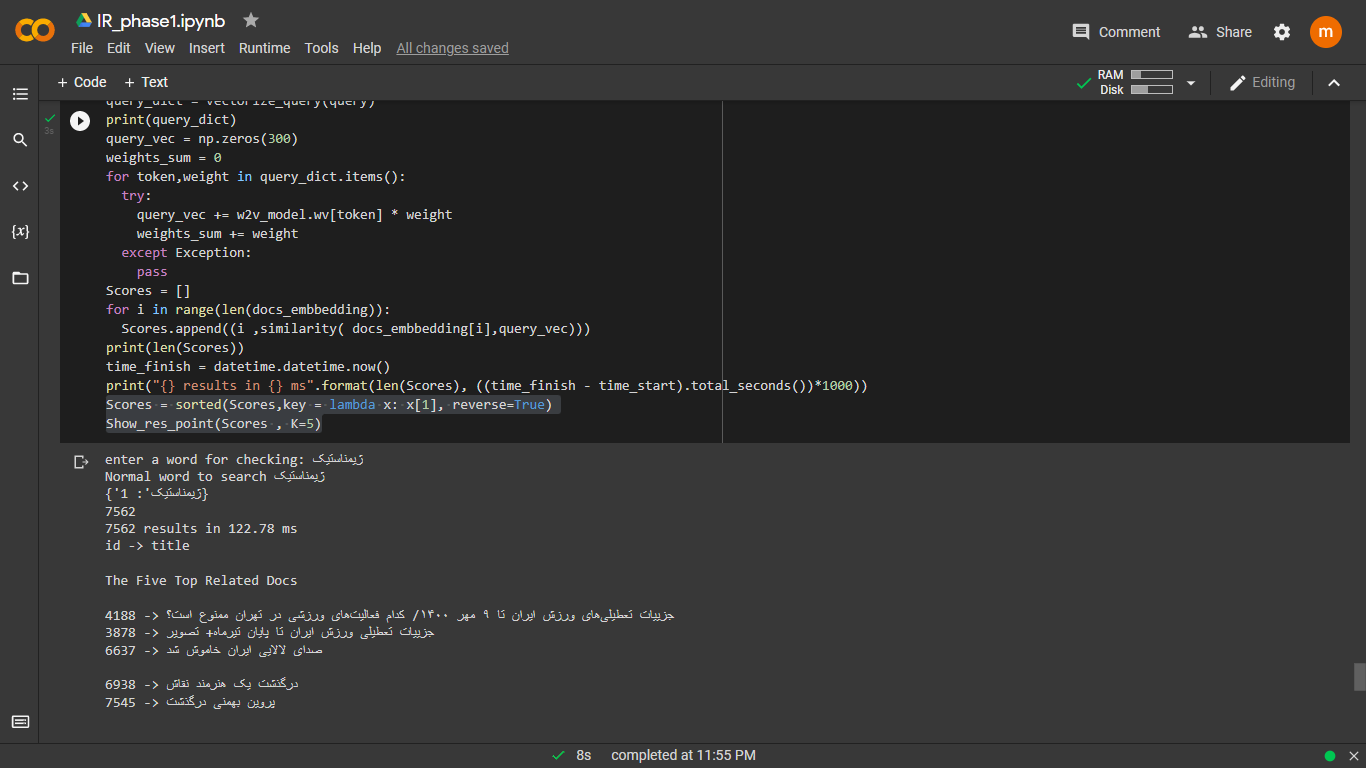
*در 5 تا داک باز گردانده شده در مدلی که ما آموزش دادیم به این صورت هست که مرتبط ترین داک در مورد تعطیلی رشته‌های ورزشی هست و سند بعدی هم به همین صورت ولی سند سوم به فوت هنرمند پیشکسوت بوده که اصلا مرتبط نبوده(تنها شباهتی که دارد وجود هنر بوده 😊) و دو داک بعدی هم به همین صورت بوده*

*برای*

*Precision@K = 1 (1 میشود برای این حالت زیرا اولینی که بازگشته مرتبط هست)*

*Precision@K = 5 (0.4 می‌شود برای این حالت زیرا که فقط 2 تا از 5 تا مرتبط بوده)*

*MAP (0.5(1+1/2))در اینجا دو تا اولی فقط مرتبط بود و مقدار 0.75 درصد می‌شود*

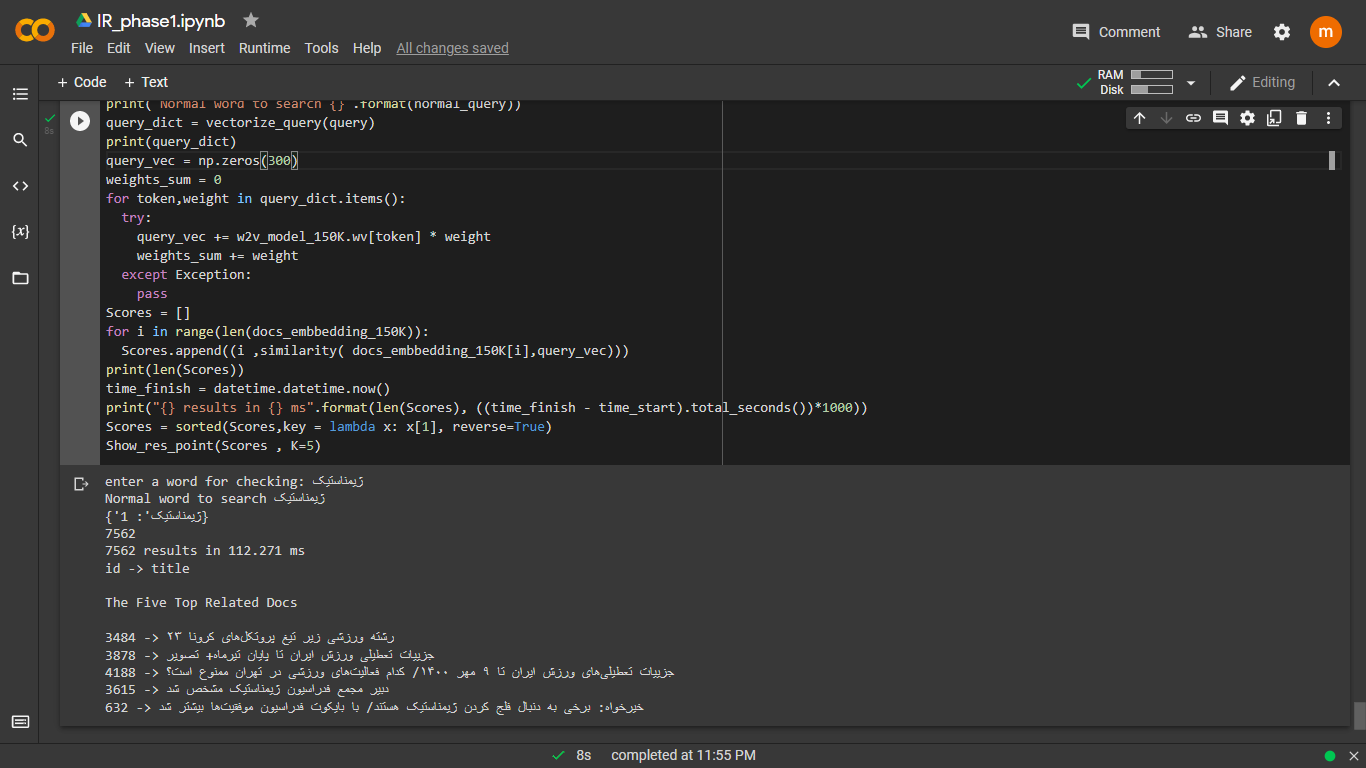
**

*در 5 داک اول که با مدل از پیش آموزش داده استفاده کردیم به این صورت هست که هر 5 داک به صورتی با کوئری مد نظر ما ارتباط داشتند و هیچ داکی بی ارتباط نبود*

*Precision@K = 1 (1 میشود برای این حالت زیرا اولینی که بازگشته مرتبط هست)*

*Precision@K = 5 (هر 5 داک مرتبط اند پس 1 می‌شود)*

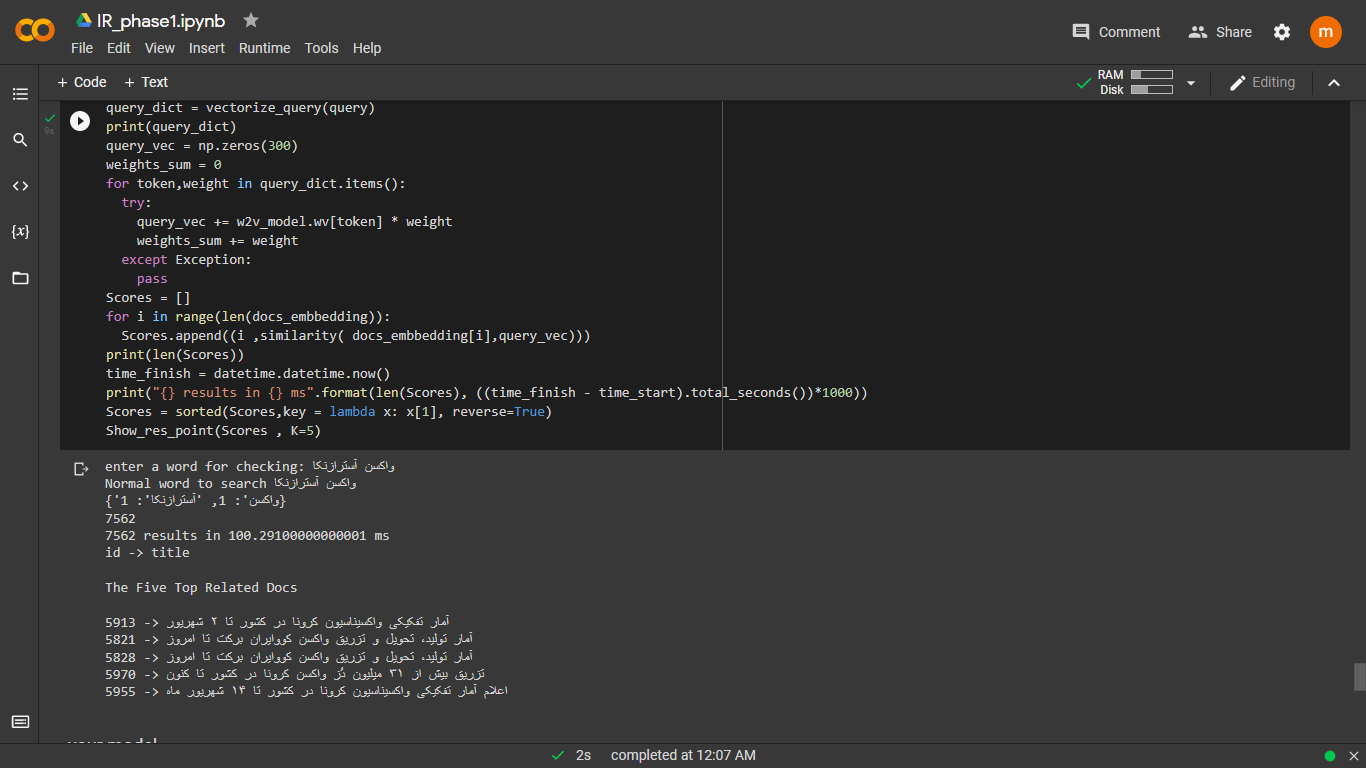
*MAP (0.2(1+1+1+1+1))هم 1 می‌شود*

**

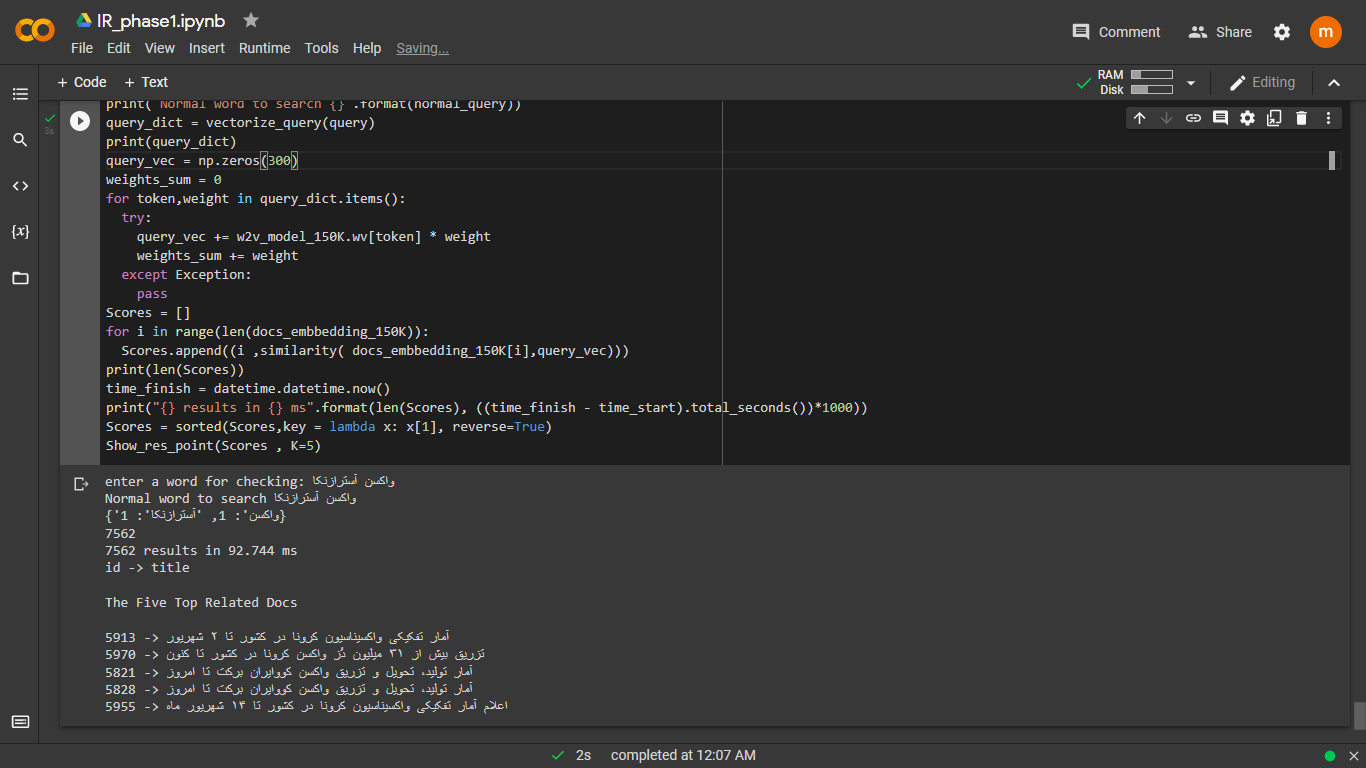
*در مورد کوئری بعد واکسن آسترازنکا در نظر گرفتیم*

*در هر دو مدل 5 داک ابتدایی بازگردانده شده یکسان هست ولی تفاوت در رنک بندی موجود هست و این می‌تواند به خاطر خاص بودن کوئری ما بوده*

*در این حالت مدل که ما آموزش دادیم*

**

*و مدل از پیش آموزش داده شده*

**

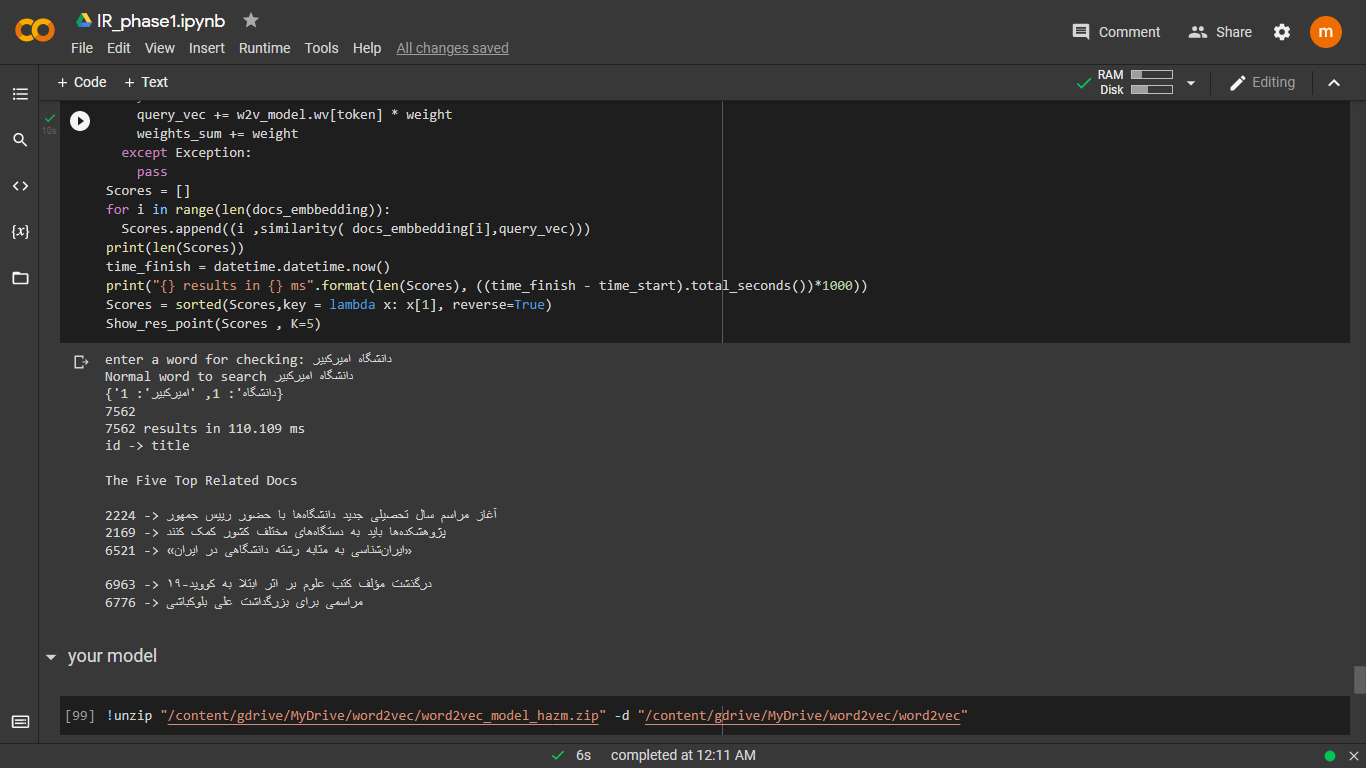
*Precision@K = 1 (1 میشود برای این حالت زیرا اولینی که بازگشته مرتبط هست)*

*Precision@K = 5 (هر 5 داک مرتبط اند پس 1 می‌شود)*

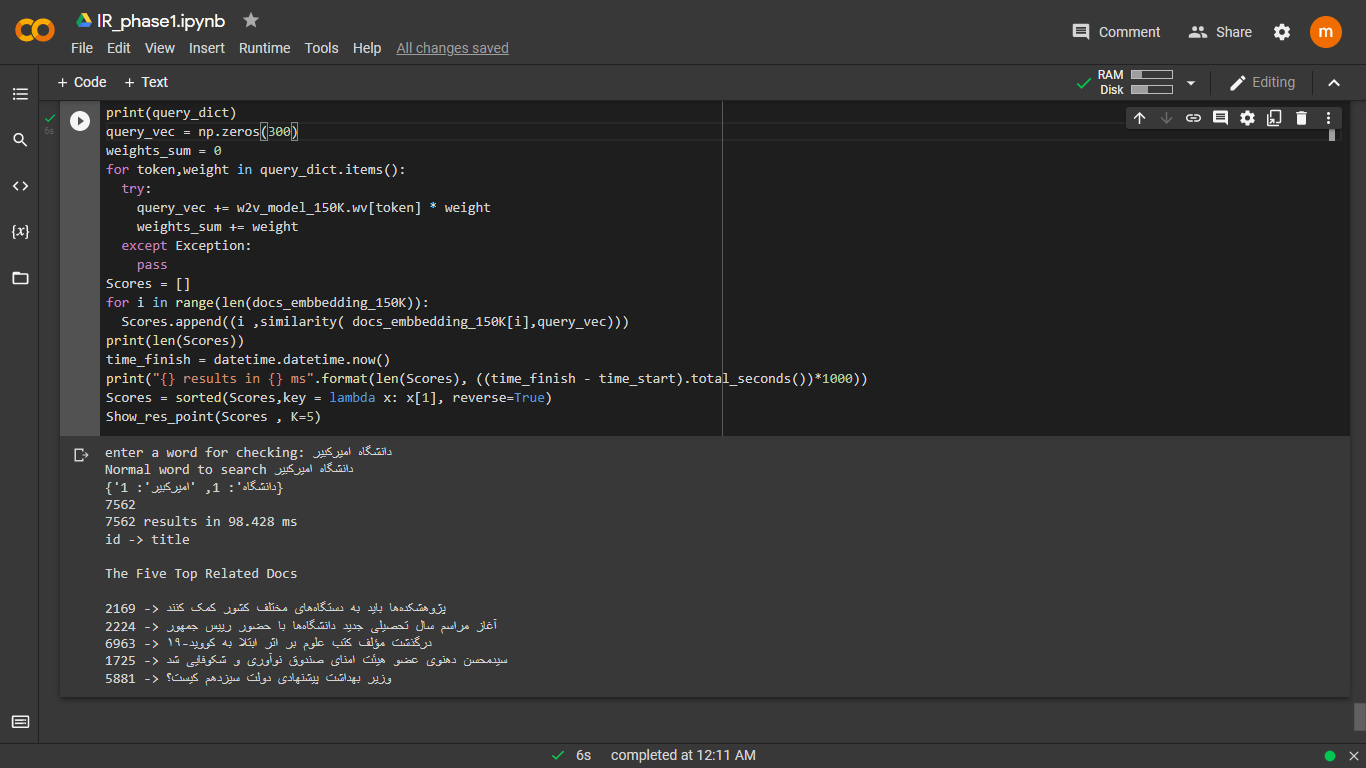
*MAP (0.2(1+1+1+1+1))هم 1 می‌شود*

*برای دانشگاه امیرکبیر که کمی عام تر از قبلی هست به این صورت هست*

*در مدلی که ما آزمایش انجام دادیم*

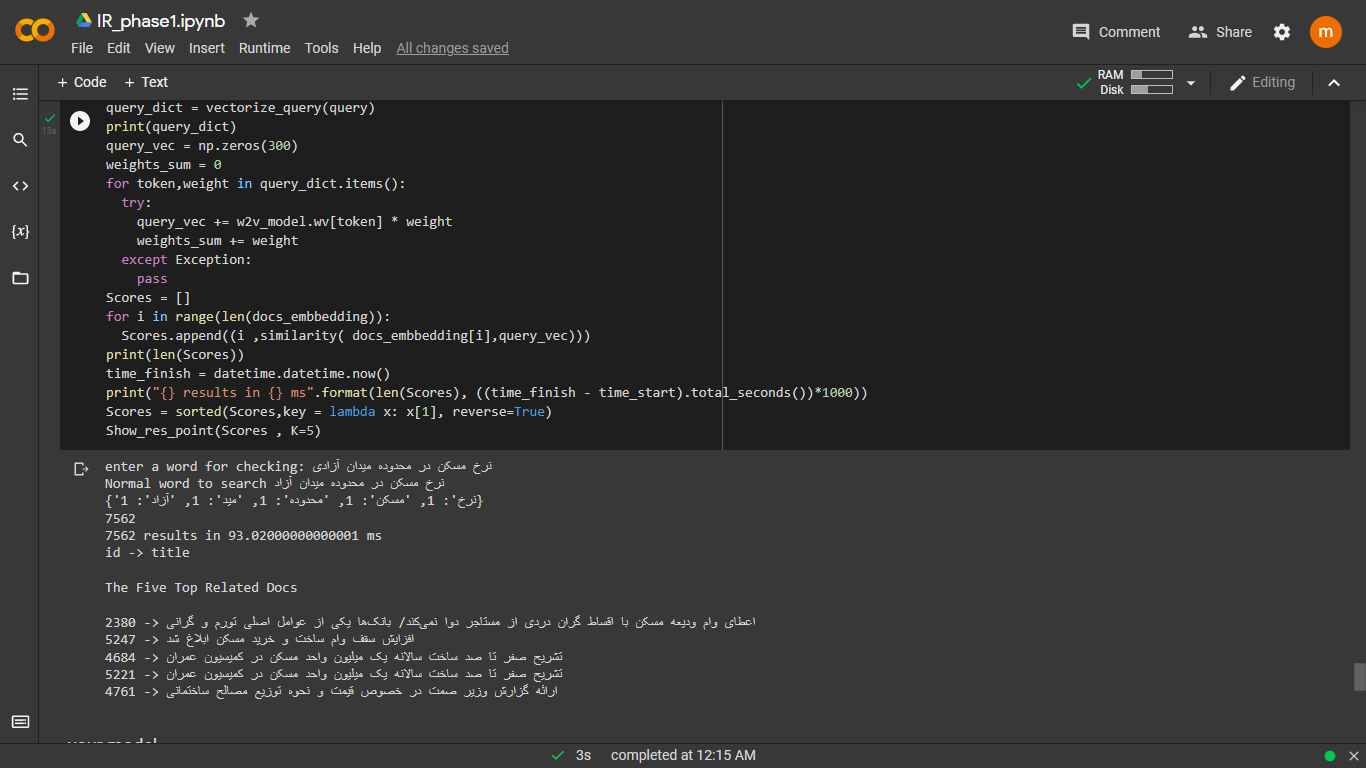
**

*و در مدل آموزش شده*

**

*برای نرخ مسکن در محدوده‌ی میدان آزادی*

*در مدلی که ما آموزش دادیم مهمترین سند واکشی شده اعطای وام ودیعه مسکن می‌باشد که به صورت کلی مربوط هست ولی نیاز ما رو برطرف نمی‌کند در دیگر حالت ها هم همین گونه است.*

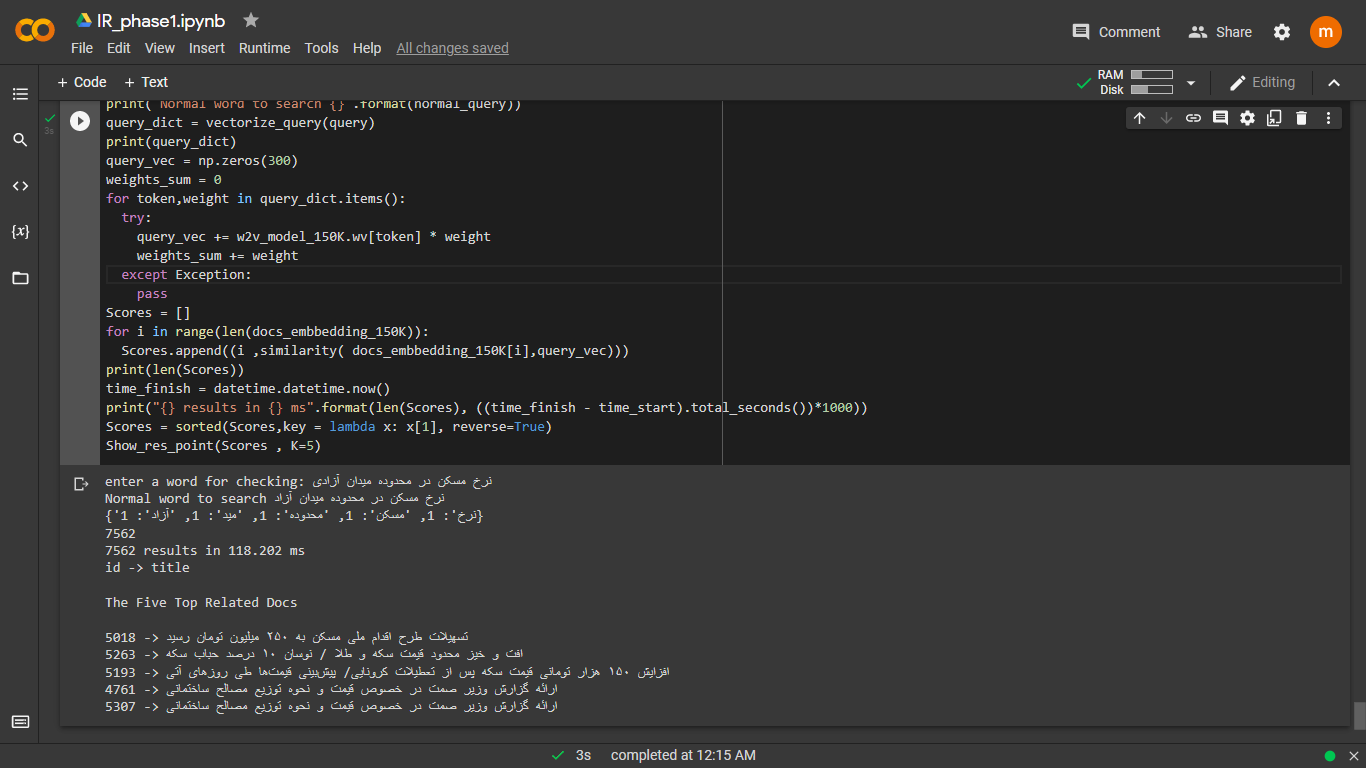
**

*Precision@K = 0(اولین داک باز گردانده شده کاملا مرتبط نیست)*

*Precision@K = 0 (هر 5 داک غیر مرتبط اند پس 0 می‌شود)*

*MAP (0 (0))هم 0 می‌شود*

*در مدل از قبل آموزش شده خیلی به صورت قبلی هست با این فرق که سند دوم وسوم در مورد مسکن هم نیست و در مورد قیمیت طلا و سکه است*

**

*Precision@K = 0(اولین داک باز گردانده شده کاملا مرتبط نیست)*

*Precision@K = 0 (هر 5 داک غیر مرتبط اند پس 0 می‌شود)*

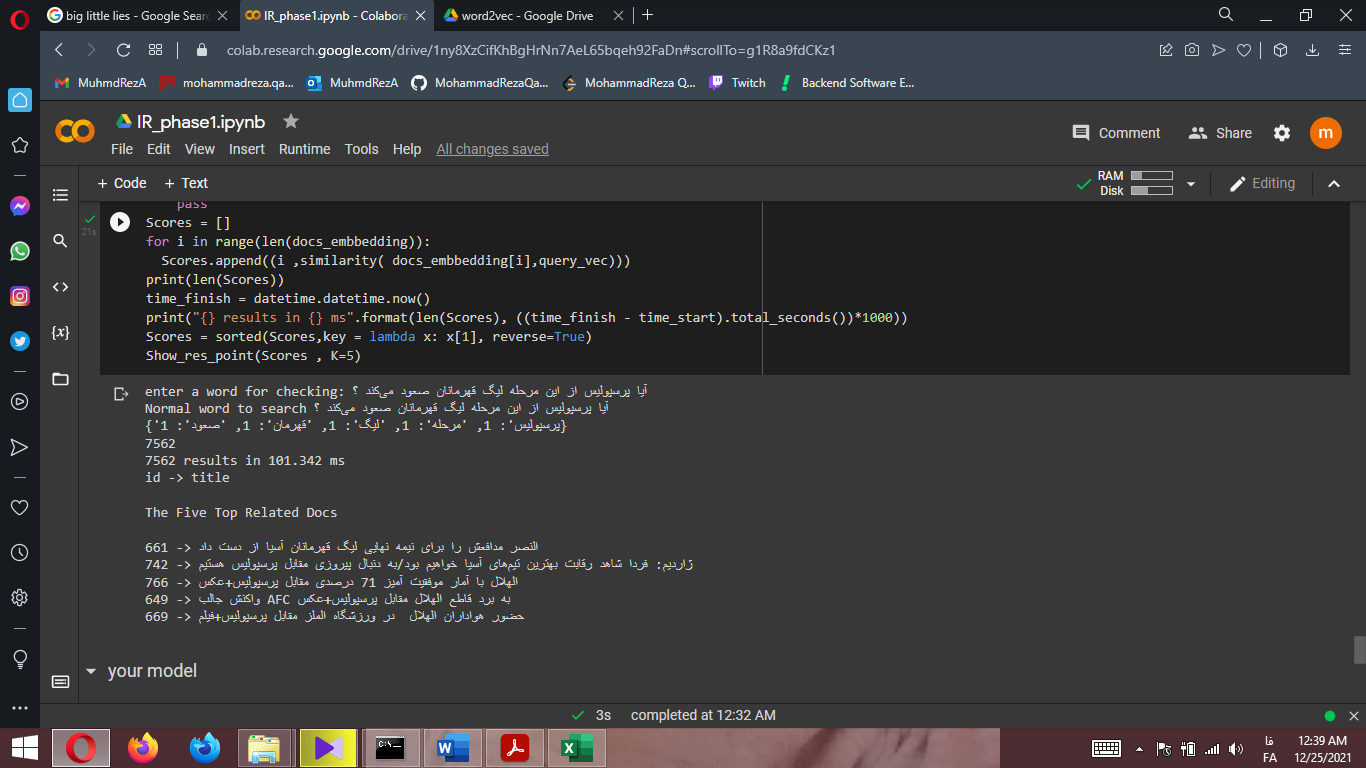
*MAP (0 (0))هم 0 می‌شود*

*اگه مرتبط بودن رو همین در نظر بگیریم که در مورد مسکن صحبت شده مقادیر متفاوت اند.*

*برای کوئری پرسشی از این کوئری استفاده کردیم*

*آیا پرسپولیس از این مرحله لیگ قهرمانان صعود می‌کند؟*

*در مدل آموزش ما همه سندها به بازی دو تیم پرسپولیس و الهلال اشاره دارند*

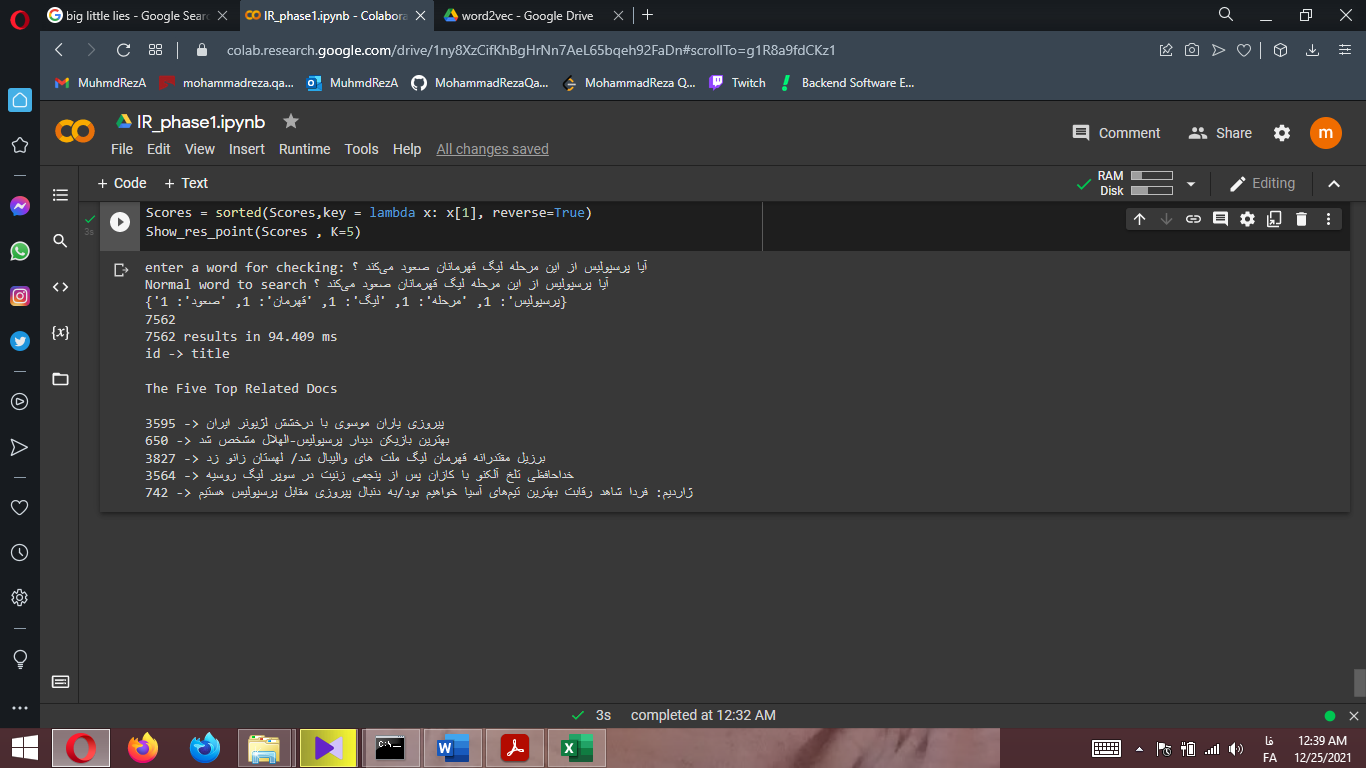
**

*Precision@K = 1(اولین داک باز گردانده شده مرتبط است)*

*Precision@K = 1 (هر 5 داک مرتبط اند پس 1 می‌شود)*

*MAP (0.2(5))هم 1 می‌شود*

*ولی در مدل آموزشی کلی سه تا از اسناد آمده مربوط به لیگ قهرمانان والیبال مربوط می‌شود*

**

*به دلیل اینکه مقادیر بیشتری رو دیده حالت کلی تری در نظر می‌گیرد*

*Precision@K = 0(اولین داک باز گردانده شده کاملا مرتبط نیست)*

*Precision@K = 0.4 (سه تا از 5 داک غیر مرتبط اند پس 0.4 می‌شود)*

*MAP (0.5(0.5 + 0.4))هم 0.45 می‌شود*

*حالا برای MRR*

*برای مدل که خودمون آموزش دادیم*

*سه تا از کوئری ها جواب درست را در اولین داک برگرداندن از چهارتا پس mrr که میانگین اینکه در چه رنکی بازگشته سند مرتبط هست می‌شود 0.75*

*برای مدل آموزش داده شده*

*یکی از کوئری ها پاسخ درست در مکان دوم و دوتا در مکان اول هست و میانگین در این حالت 0.625 می‌شود ولی باید این مورد را در نظر گرفت که این تعداد برای مقایسه این دو کافی نیست*

*برای برچسب گذاری می‌توانیم از میزان شباهت استفاده کنیم و ترشهولدی در نظر بگیریم.*