بسم الله الرحمن الرحیم



درس : جستجو و بازیابی وب

موضوع : شرح پروژه

ارائه دهنده : محمد خسروی

استاد : جناب آقای دکتر مهدی جبل عاملی

مقدمه

در این پروژه، یک سیستم بازیابی اطلاعات بر روی مجموعه داده **CISI**پیاده ‌سازی شده است. هدف اصلی، استخراج و تحلیل اطلاعات موجود در اسناد و ارزیابی عملکرد سیستم بر اساس معیارهای استاندارد بازیابی اطلاعات بوده است.

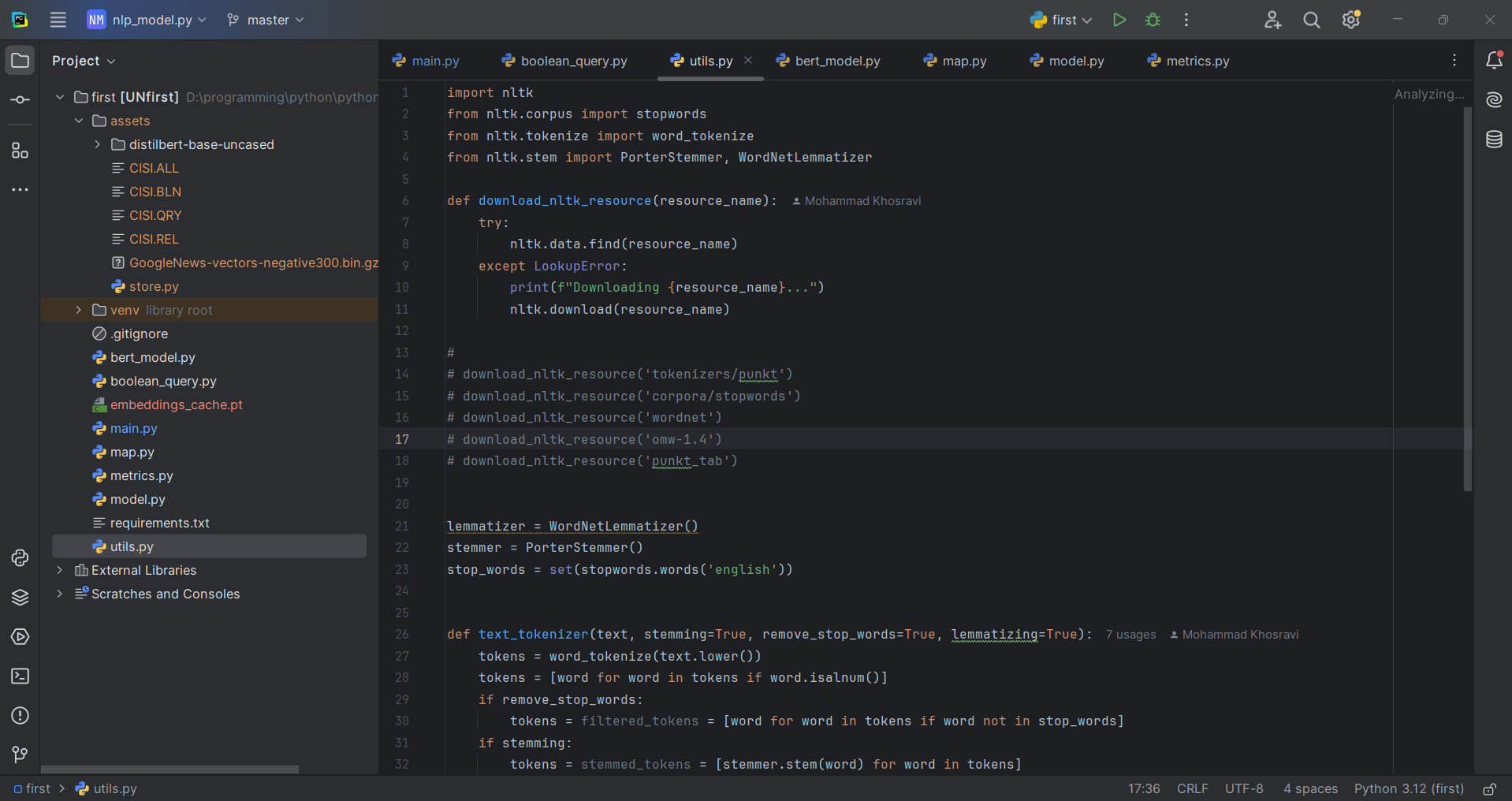
در این راستا، از ابزار**NLTK** برای پیش‌پردازش داده‌ها استفاده شده و روش‌های مختلفی برای بازیابی و رتبه‌بندی اسناد پیاده‌سازی شده است. همچنین عملکرد سیستم با استفاده از معیارهایی مانند Precision و MAP ارزیابی شده است.

گزارش حاضر به تشریح نحوه پیاده‌سازی، چالش‌ها و نتایج به دست آمده می‌پردازد.

**پیش‌پردازش متن و ساخت شاخص معکوس در سیستم بازیابی اطلاعات**

ابتدا عملیات پیش‌پردازش متن انجام شد. این فرآیند شامل توکن‌سازی بود که متن ورودی را به کلمات مجزا تقسیم می‌کند. سپس کلمات غیرضروری یا به اصطلاح stopwordsمانند "the"، "is" و سایر کلمات پرتکرار و کم‌ارزش حذف شدند. علاوه بر این، با استفاده از روش‌های ریشه ‌یابی (Stemming) و لماتیزه کردن (Lemmatization)، کلمات به فرم پایه یا ریشه‌ای تبدیل شدند. این کار به استانداردسازی متن و کاهش پیچیدگی داده‌ها کمک می‌کند.

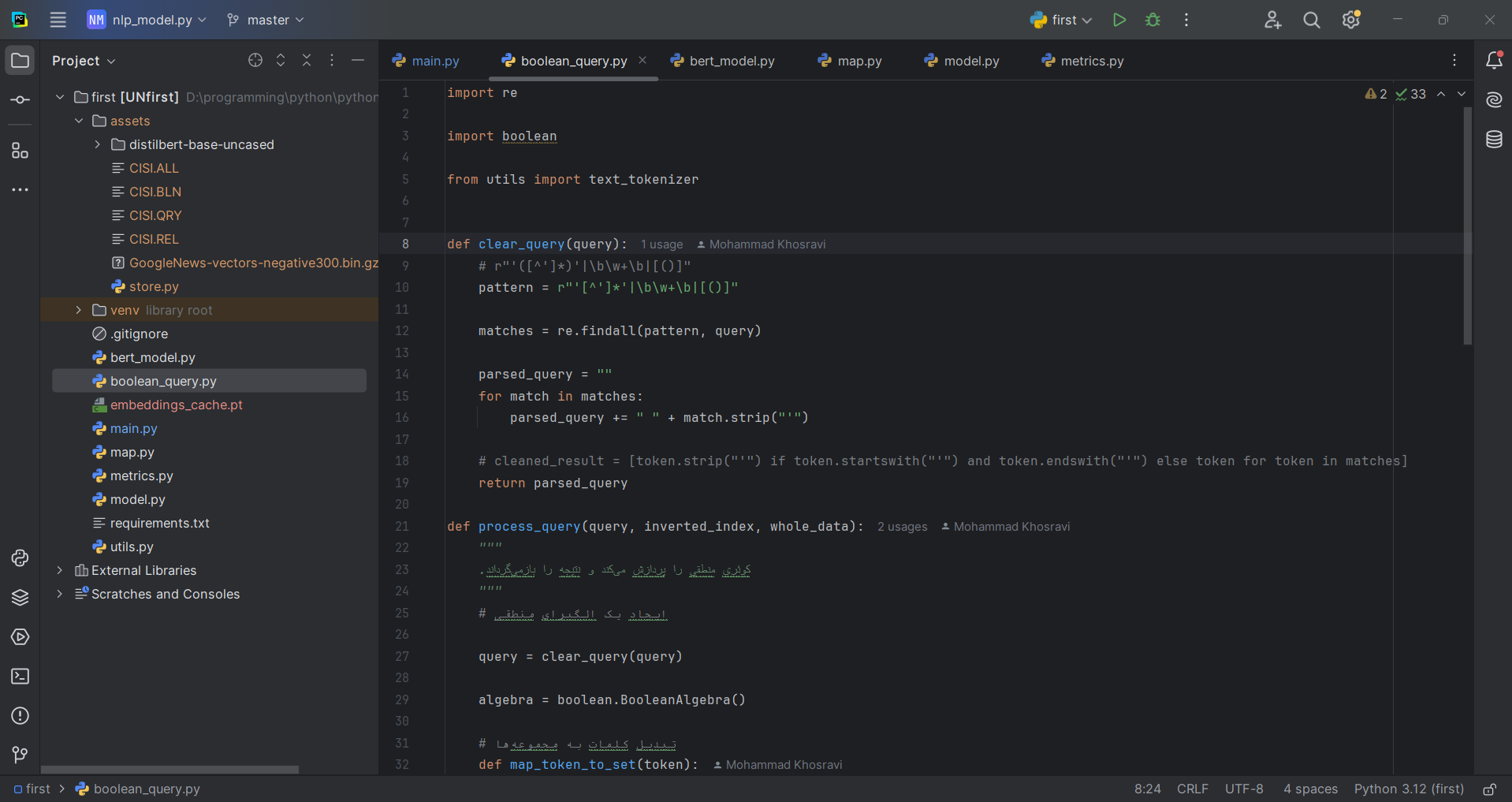
در مرحله بعد، شاخص معکوس ساخته شد. شاخص معکوس ابزاری است که برای هر کلمه کلیدی، لیستی از اسناد حاوی آن کلمه را ذخیره می‌کند. این شاخص به سیستم امکان می‌دهد اسناد مرتبط با یک کوئری خاص را به سرعت و دقت بازیابی کند.



**پردازش کوئری منطقی و بازیابی اسناد**

در این بخش، سیستمی برای پردازش کوئری‌های منطقی و بازیابی اسناد مرتبط پیاده‌سازی شد. ابتدا کوئری با استفاده از تابع clear\_query پیش‌پردازش شده و به قالبی قابل ‌پردازش تبدیل می‌شود. سپس با بهره‌گیری از کتابخانه boolean، کوئری به یک ساختار منطقی تبدیل می‌شود که شامل عملیات AND ، OR، و NOT است.

با استفاده از تابع بازگشتی evaluate\_expression، گره‌های این ساختار ارزیابی شده و بر اساس شاخص معکوس، اسناد مرتبط بازیابی می‌شوند. این روش امکان جستجوی پیشرفته و دقیق را در مجموعه داده فراهم می‌کند.

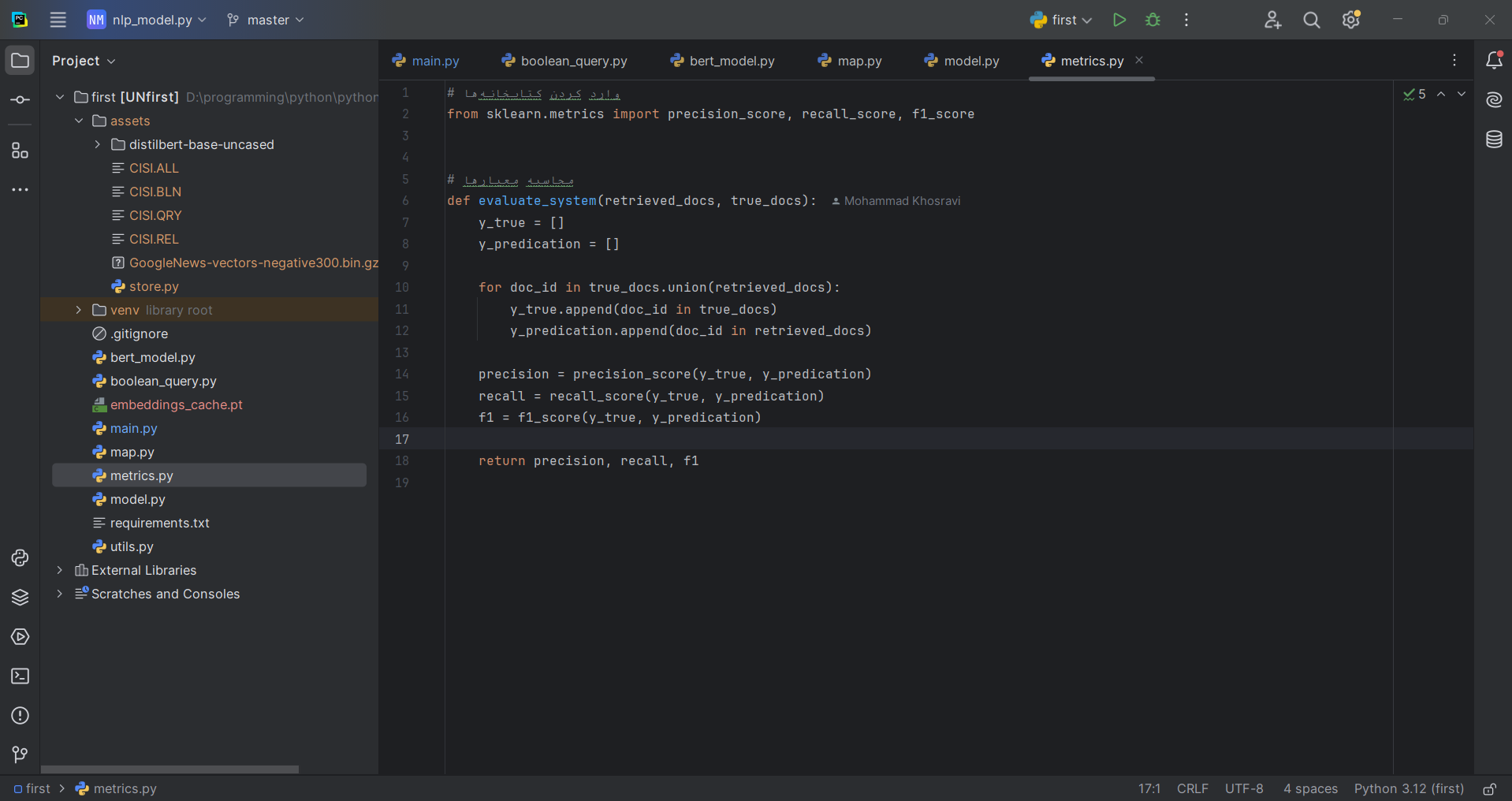


**محاسبه معیارهای ارزیابی عملکرد سیستم بازیابی اطلاعات**

این بخش به محاسبه معیارهای ارزیابی سیستم بازیابی اطلاعات اختصاص دارد سه معیار اصلی شامل Precision، Recall و F1-Score محاسبه شده‌اند

تابع evaluate\_system عملکرد سیستم را بر اساس اسناد بازیابی‌شده و اسناد واقعی ارزیابی می‌کند  
**Precision** : درصد اسناد بازیابی‌شده که واقعاً مرتبط هستند  
**Recall** : درصد اسناد مرتبط که توسط سیستم بازیابی شده‌اند  
**F1-Score** : میانگین موزون Precision و Recall برای ارائه یک ارزیابی کلی

این تابع ابتدا اسناد بازیابی‌شده و اسناد واقعی را باهم مقایسه کرده و مقادیر مربوط به هر معیار را محاسبه و بازمی‌گرداند مورد بررسی قرار میدهیم. سپس براساس آن ها اسناد را بررسی می کنیم.



**مقایسه اسناد با استفاده از مدل Word2Vec**

در این بخش، یک سیستم مقایسه سند با استفاده از مدل Word2Vec پیاده‌ سازی شده است. مراحل اصلی این کار به شرح زیر است:

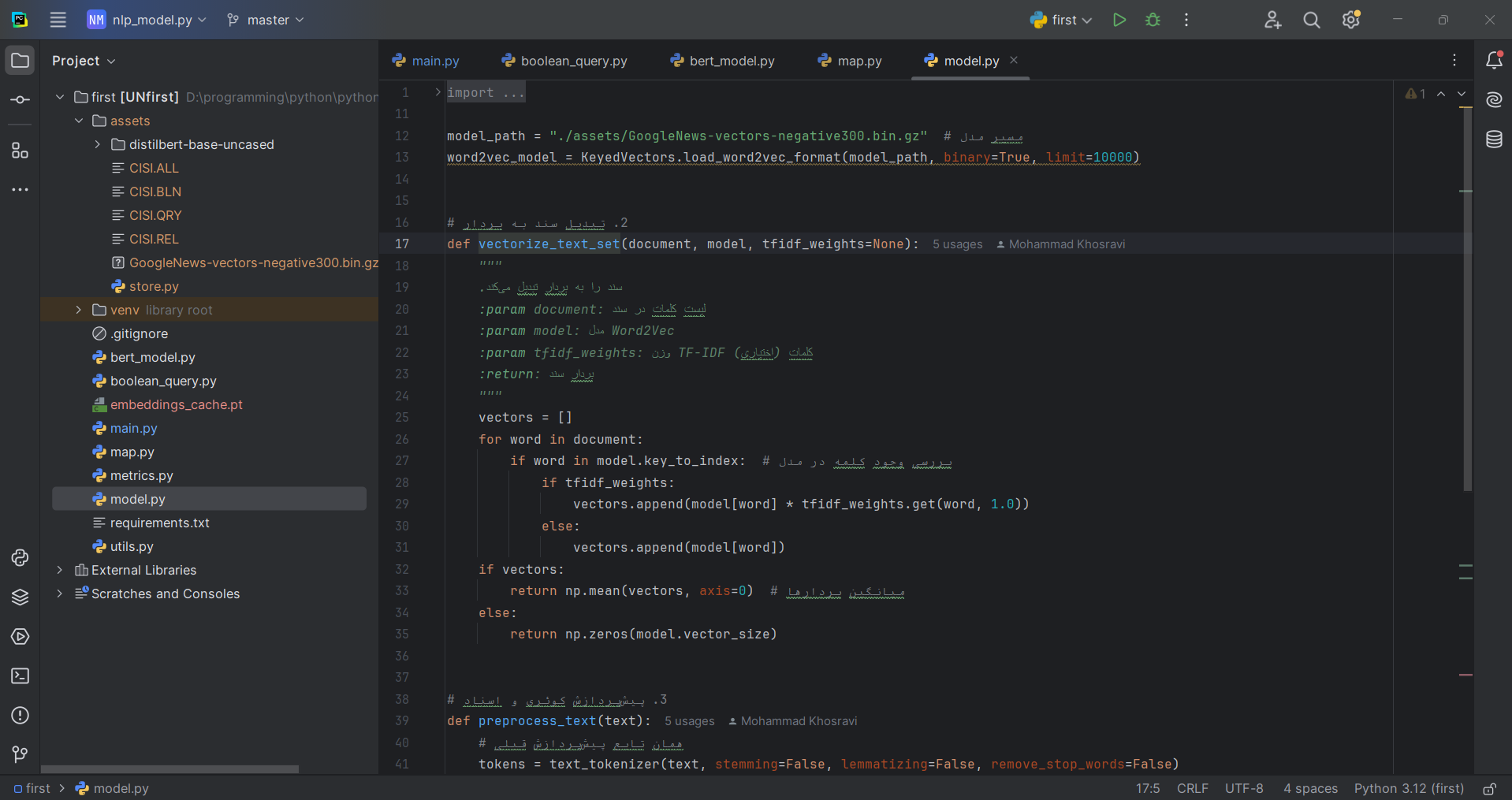
**مدل Word2Vec**: از مدل Word2Vec از پیش‌آموزش‌داده‌شده (مدل Google News) برای تبدیل کلمات به بردارهای عددی استفاده شده است. این مدل توانایی درک روابط معنایی بین کلمات را دارد.

**تبدیل سند به بردار**: با استفاده از مدل Word2Vec، اسناد به بردارهایی تبدیل می‌شوند که نمایانگر محتوای معنایی آن‌ها هستند. در این فرآیند، کلمات داخل سند به بردار تبدیل شده و میانگین آن‌ها به‌عنوان نمایه سند محاسبه می‌شود. در صورت استفاده از وزن‌های TF-IDF، این وزن‌ها نیز به بردارها اعمال می‌شود.

**مقایسه اسناد و کوئری‌ها**: برای مقایسه کوئری‌ها با اسناد، ابتدا کوئری توکنایز شده و به بردار تبدیل شده و میانگین ان بدست می اوریم. سپس شباهت کسینوسی بین بردار سند و بردار کوئری محاسبه می‌شود. این شباهت معیاری برای سنجش میزان ارتباط سند با کوئری است.

**یافتن اسناد مشابه‌ تر**: تابعی برای مقایسه چندین سند با کوئری نوشته شده که بر اساس شباهت کسینوسی، اسنادی که بیشترین مشابهت را دارند پیدا کرده و به‌عنوان نتایج باز می‌گرداند.

این کد به شناسایی اسناد مشابه‌تر نسبت به یک کوئری خاص کمک می‌کند.



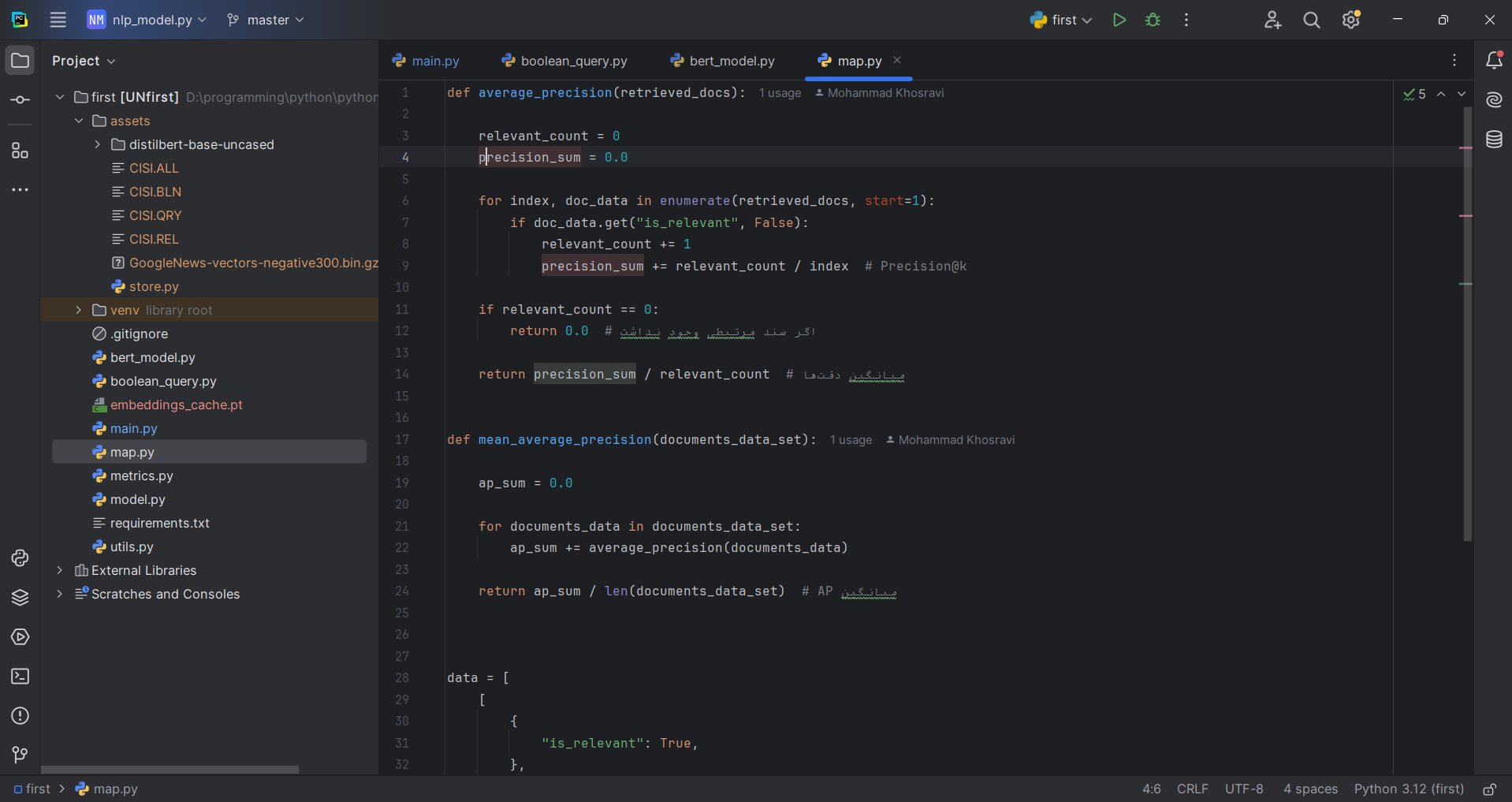
**محاسبه MAP**

این بخش برای ارزیابی عملکرد سیستم‌های بازیابی اطلاعات و سیستم‌های پیشنهاد ‌دهنده طراحی شده‌اند. معیارهایی که محاسبه می‌شوند، میانگین دقت برای هر جستجو و میانگین دقت سراسری برای مجموعه‌ای از جستجوها هستند.

عملکرد به این صورت ارزیابی می‌شود که دقت در هر موقعیت بر اساس تعداد اسناد مرتبط تا آن نقطه محاسبه شده و برای هر جستجو میانگین گرفته می‌شود. سپس میانگین این مقادیر برای تمام جستجوها به عنوان معیار کلی عملکرد سیستم ارائه می‌شود.

این روش ارزیابی به تحلیل و بهبود کیفیت بازیابی اطلاعات کمک می‌کند و در سیستم‌های جستجو و پیشنهاد‌دهنده کاربرد زیادی دارد. مزیت اصلی آن سادگی در تحلیل کیفیت بازیابی است، اما حساسیت به ترتیب اسناد و نیاز به داده‌های برچسب‌گذاری‌شده از محدودیت‌های آن محسوب می‌شوند.

در این پروژه، مقدار MAP برای ارزیابی مدل‌ها محاسبه شد. برای مدل Word2Vec، مقدار MAP حدود **0.578** و برای مدل BERT حدود **0.756** به دست آمد. این مقادیر با استفاده از تعداد زیادی کوئری آزمایشی و بررسی عملکرد بازیابی اطلاعات محاسبه شده‌اند.



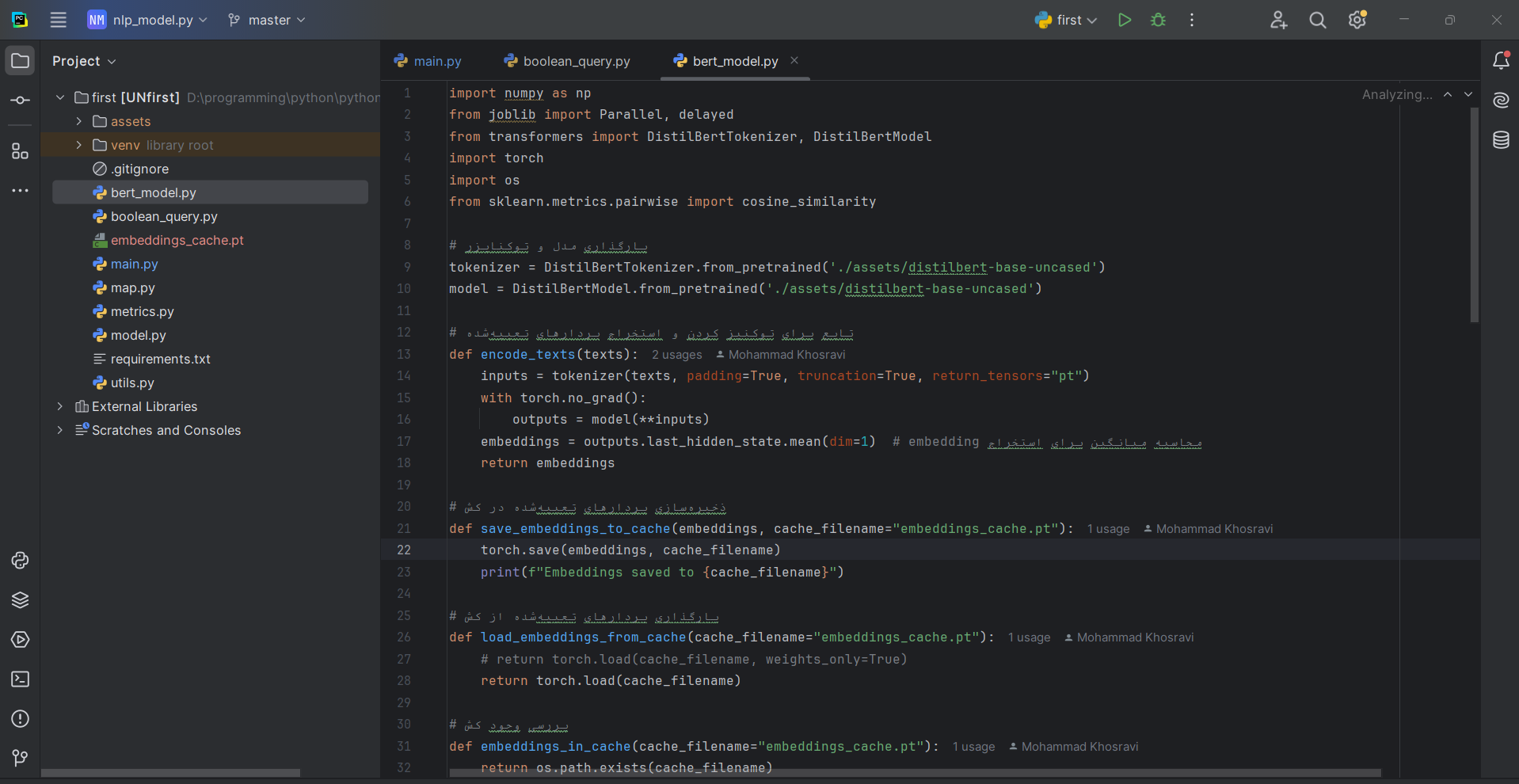
**مدل bert**

این کد با استفاده از مدل DistilBERT برای استخراج بردارهای تعبیه ‌شده از متون طراحی شده و هدف آن مقایسه شباهت معنایی بین یک کوئری و مجموعه‌ای از اسناد است.

ابتدا مدل و توکنایزر بارگذاری شده و متون به بردارهای عددی تبدیل می‌شوند. این بردارها با استفاده از مکانیزم Cachingذخیره می‌شوند تا در پردازش‌های بعدی نیازی به محاسبه مجدد نباشد و عملکرد بهینه ‌تری ارائه شود.

برای افزایش سرعت، متون به دسته‌های کوچک تقسیم شده و به ‌صورت موازی پردازش می‌شوند. در نهایت، شباهت معنایی با استفاده از معیار کسینوسی محاسبه شده و نزدیک‌ترین اسناد به کوئری شناسایی می‌شوند.

این کد در تحلیل متون، جستجوی معنایی و سیستم‌های بازیابی اطلاعات کاربرد دارد.



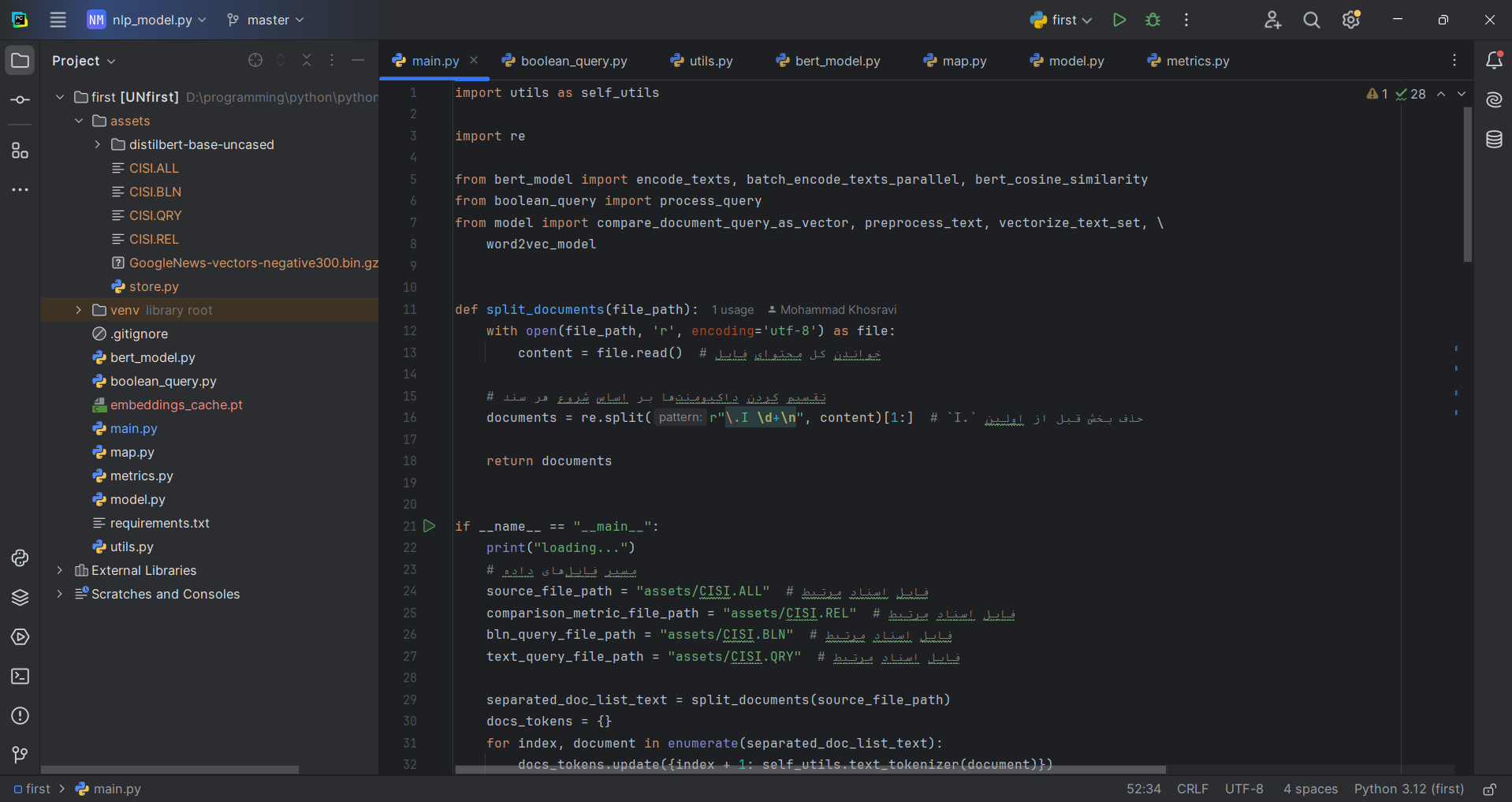
**اجرا و اینترفیس**

این فایل main یک اینترفیس ساده برای تعامل کاربر با پروژه فراهم می‌کند و نقطه شروع اجرای سیستم است. ابتدا فایل‌های داده خوانده و اسناد به داکیومنت‌های جداگانه تقسیم می‌شوند. سپس با پردازش متن، ایندکس معکوس و بردارهای Word2Vec و BERT برای جستجو ایجاد می‌شوند.

کاربر می‌تواند بین دو نوع جستجو انتخاب کند:

1. **جستجوی باینری**: بر اساس ترکیب منطقی کلمات و ایندکس معکوس.
2. **جستجوی برداری**: با استفاده از مدل‌های Word2Vec یا BERT برای یافتن اسناد مشابه.

در نهایت، نتایج جستجو به کاربر نمایش داده می‌شود.



"با تلاشی که ز دل برخاست،  
این پروژه به ثمر بنشست."

امیدوارم توضیحان جنابعالی را جمع کرده باشد. ممنون از وقتی که گذاشتید.