

دانشکده مهندسی کامپیوتر

دانشگاه اصفهان

**تکلیف اول بازیابی اطلاعات**

استاد: دکترمحمد مهدی رضاپور

مهروالسادات نوحی

993613061

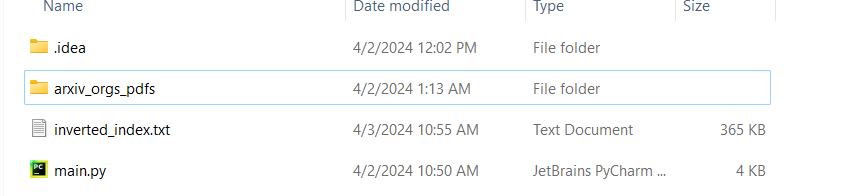
بهار 1403

ایندکس معکوس:

صورت تمرین: شما می‌بایست یک ایندکس معکوس روی 20 مقاله از پایگاه arxiv.org که به صورت رندوم و البته در موضوعات متنوع دانلود کرده‌اید ایجاد کنید. سپس می‌بایست برنامه‌ای بنویسید که کوئری‌های بولینی کاربر را توسط این ایندکس پاسخ دهد. کدها می‌بایست با زبان پایتون نوشته شود. برنامه می‌بایست قابل تست باشد وگرنه نمره‌ای به تکلیف تعلق نمیگیرد.

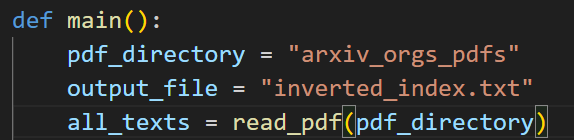
گام اول:

در این گام ابتدا با مراجعه به [arxiv.org](https://arxiv.org/) شروع به انتخاب و دانلود 20 مقاله کردم و در پوشه arxiv\_orgs\_pdfs فایل‌ها را قرار دادم. و در گام بعد شروع به کدنویسی می‌کنیم.



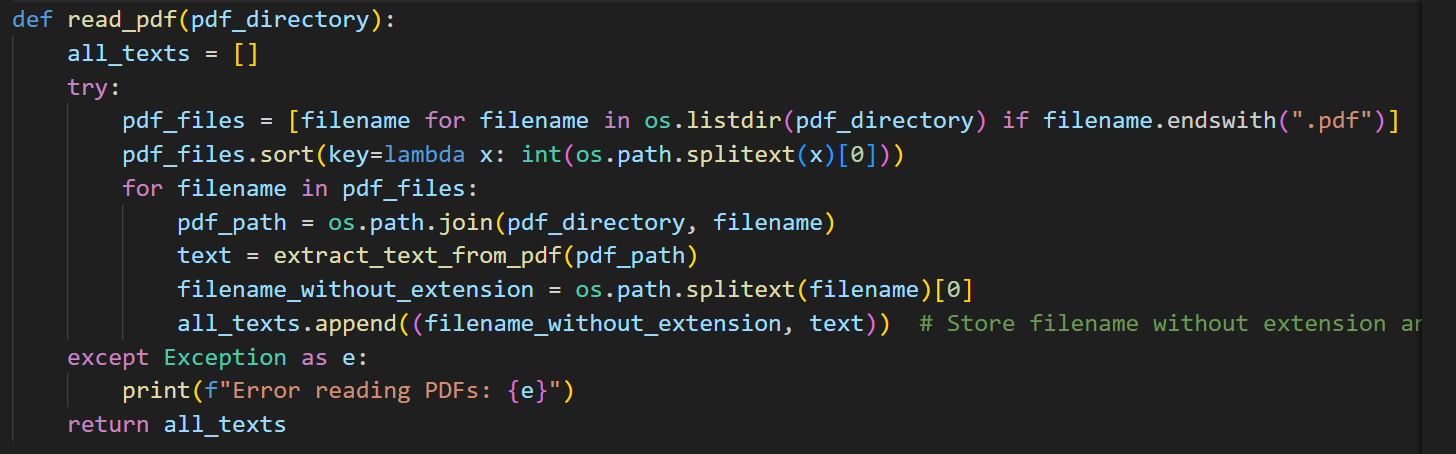
گام دوم:

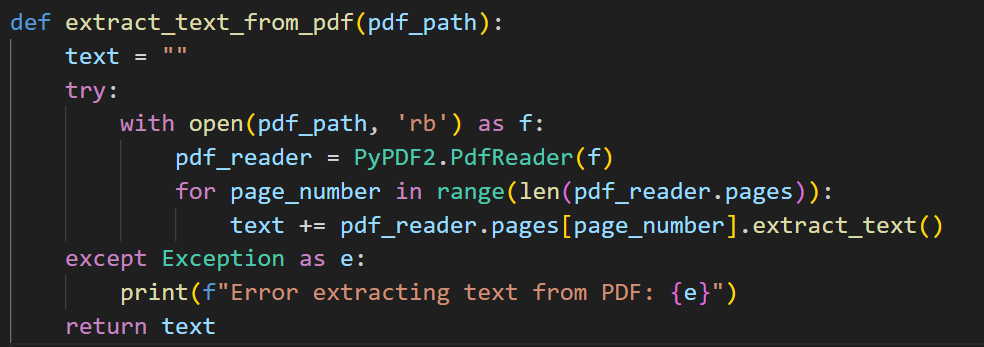
در این گام باید متن‌های داخل فایل‌های .pdf را استخراج کنیم که بتونیم روی آن‌ها کار کنیم. برای این کار دو کتابخانه معروف پایتون هست.1-pdfpy2 و2-pdfplumber که من از اولی استفاده کردم. و خروجی برنامه را در inverted\_index.txt ریخته می‌شود. با فراخوانی تابع read\_pdf و دادن آدرس مکان pdfs ها شروع به استخراج متن می‌کنیم.



با تعریف یک دیکشنری که کلید آن شماره داکیومنت و مقدار آن متن داخل فایل است .

تذکر: من اسم فایل‌ها را برای راحتی کار به مثلا 1.pdf و.. تغییر دادم وبرای استخراج فقط شماره فایل داریم.





در واقع به شکل زیرشده است:

{

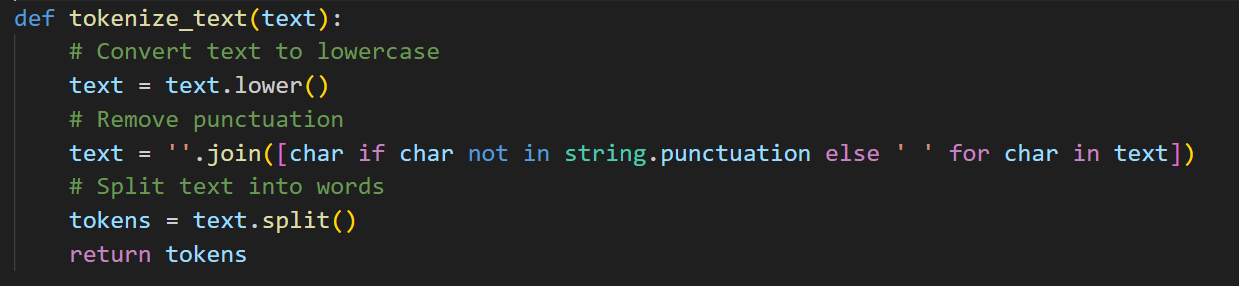
“Filename1”: text1,

“Filename2”:text2,

…{

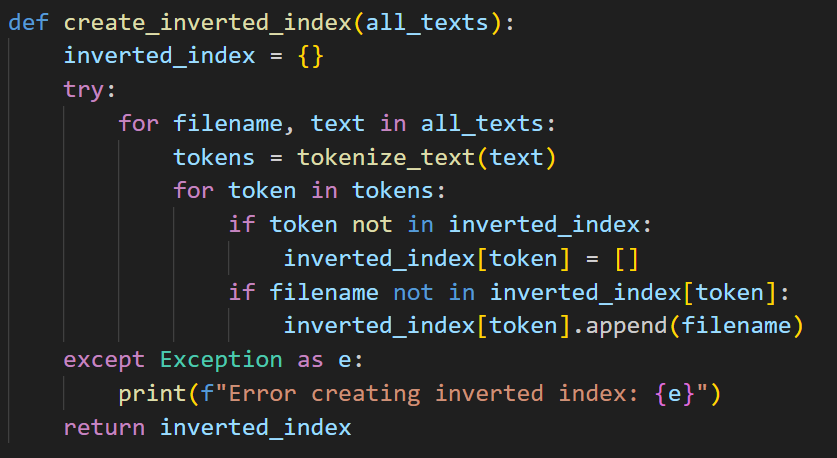
گام سوم:

بعد از اینکه متن‌ها استخراج شده است حال باید توکنایز شود و به اصطلاح هر توکن یک کلمه در نظر بگیریم برای درخواست‌های کاربر.



گام چهارم:

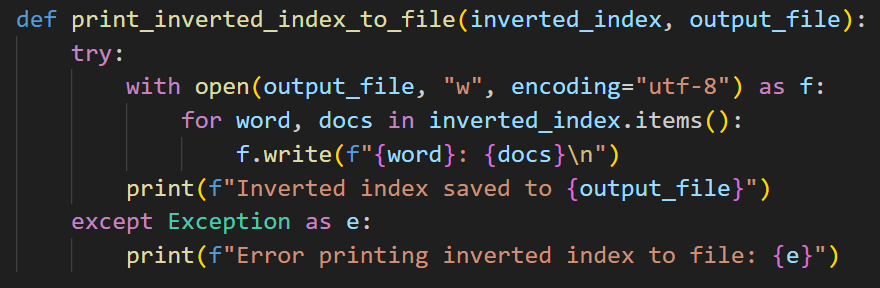
در این گام شروع به ایجاد جدول معکوس می‌کنیم:



حلقه بیرونی برای پیمایش هر داکیومنت و حلقه داخلی به ازای هر توکن در متن چک می‌کند اگرتوکن در جدول که به صورت دیکشنری تعریف شده است که کلید آن توکن یا کلمه هست و مقدار آن لیستی از داکیومنت یا داکیومنت‌هایی که آن کلمه در فایل‌ها موجود است اگر وجود نداشته باشد به جدول اضافه می‌شود و روبرو آن یک لیست خالی تعریف شده واگر شماره فایل نبود آن فایل به لیست اضافه می‌شود.

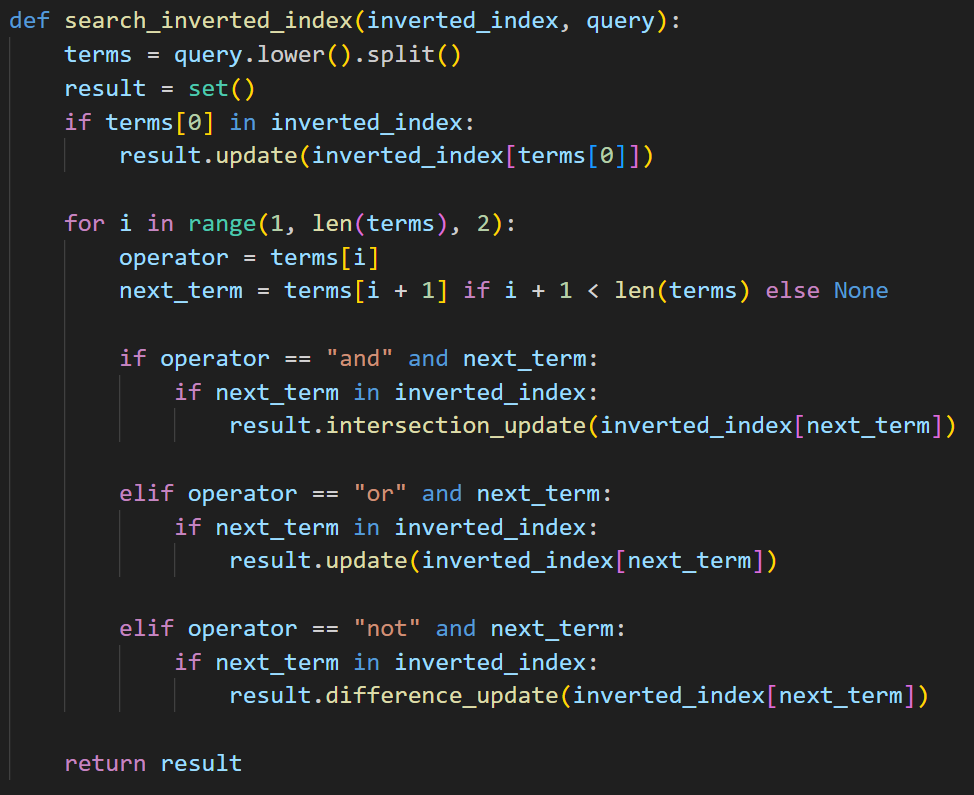
گام پنجم:

برای نوشتن جدول معکوس در فایلی که درابتدا گفتم این تابع را تعریف می‌کنیم:



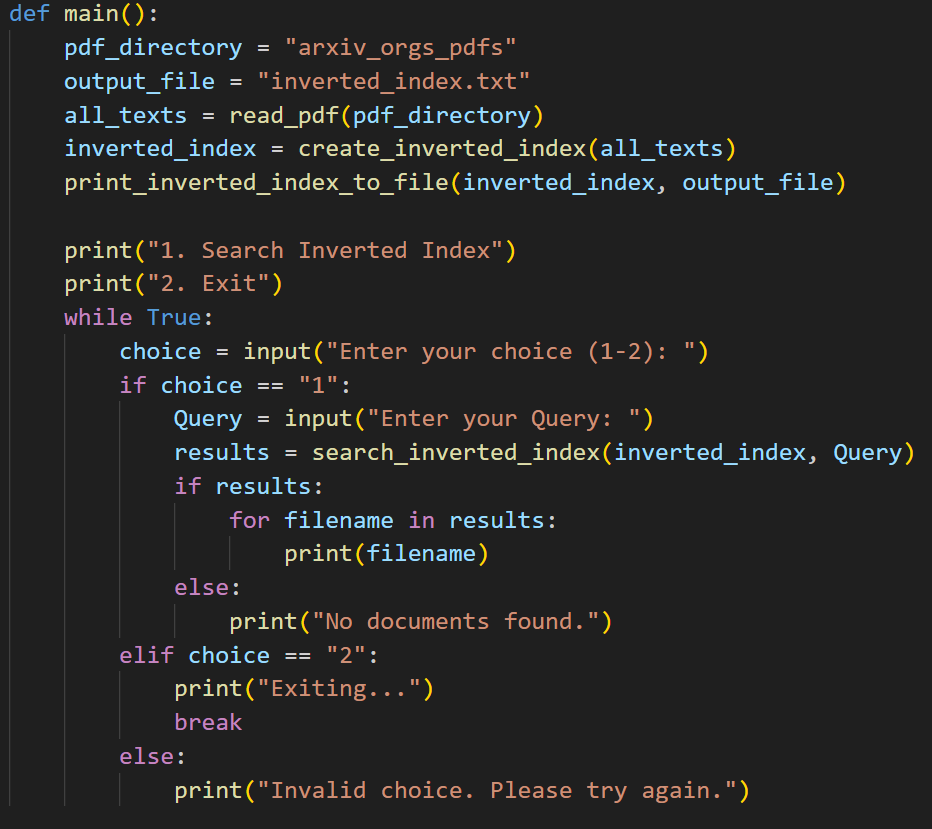
گام ششم:

جدول ایجاد و آماده سرچ کاربر می‌باشد (بستگی دارد صرفا سرچ کاربر یک کلمه ساده یا کوعری بولینی باشد تصمیم می‌گیریم چی در خروجی به او نشان دهیم.)

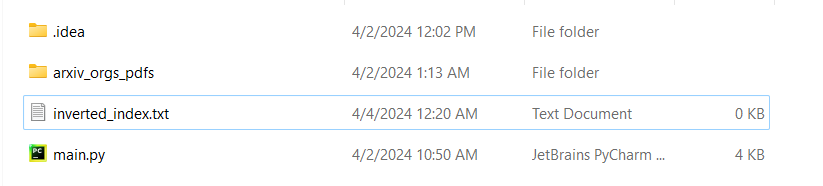


بعد از اینکه کاربر کوعری خود را وارد کرد کلمه به کلمه پردازش می‌کنیم. و اول کلمه اول را سرچ کرده که در جدول معکوس هست یا خیر اگر بود در result اضافه می‌کنیم در For میبینیم اگر کوعری تک کلمه ای بوده باشد که خروجی برگردانده می‌شود و اگر کوعری بولینی باشد شروع به پیدا کردن اپراتور و کلمه بعدی کرده و مطابق با اپراتور نتیجه اپدیت می‌شود.

و شروع برنامه main است که به صورت یک حلقه همیشه درست برای کوعری کاربر نوشته شده است.

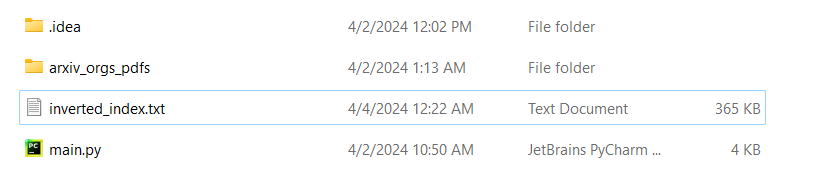


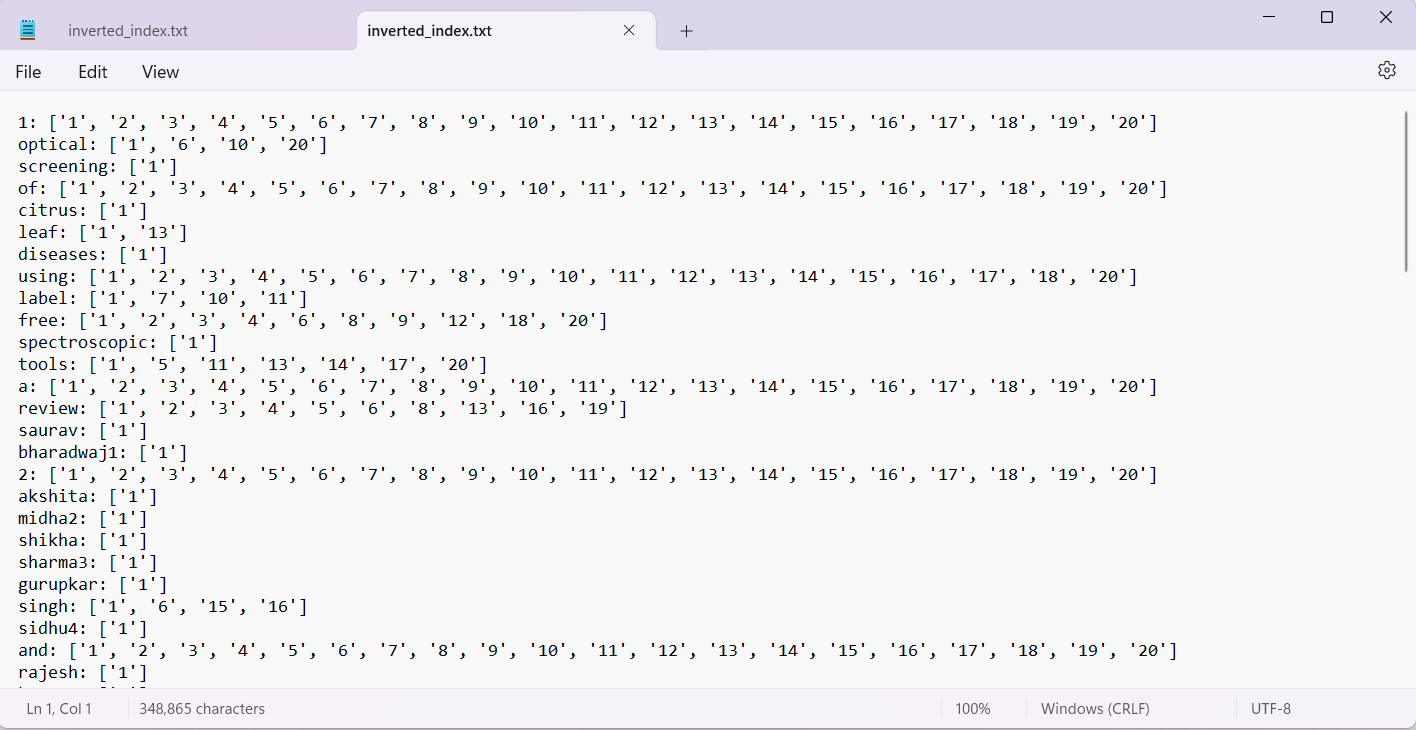
حالا شروع به اجرا کرده در ابتدا فایل خالی است. بعد از اجرای برنامه جدول معکوس را خواهیم دید.



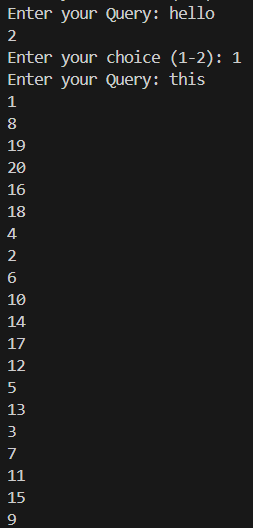
بعد از اجرا برنامه :



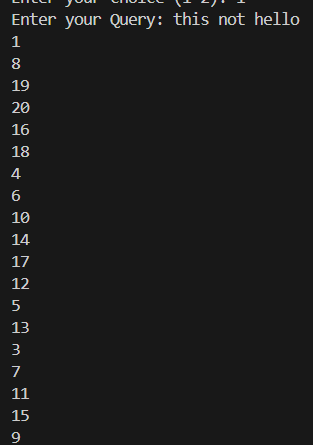




حالا بریم چند کوعری ببینیم:



کلمه hello در فایل 2 فقط هست و کلمه this در هر20 فایل من بوده است حالا بریم کوعری بولینی هم بزنیم:



الان شماره فایل 2 دیگر نیست چون از not استفاده شده است. من تست‌های دیگری هم زدم و اوکی بوده است.

