ابتدا فایل های تمرین را دانلود و از حالت فشرده خارج کردیم. CISI.ALL را بصورت خط به خط خواندیم. اگر خط با .I شروع شده مشخص کننده شناسه داکیومنت است و اگر با .W شروع شده باشد متن داکیومنت را داریم که هر خط بعد از آن را نرمالسازی و توکن سازی کردیم و در همین حین پستینگ لیست را نیز ساختیم. کلمات توقف را حذف نکردیم که بعدا برای بررسی آن را حذف کنیم همچنین ریشه یابی هم بعدا برای بررسی با وجود شرط انجام میدهیم.

در مرحله بعد CISI.QRY را خواندیم که همانند فایل قبلی شامل ایدی و متن کوئری میباشد. یک دیکشنری از ایدی کوئری ها و متن آن هم ساختیم.

سپس فایل CISI.REL را خواندیم که با توجه به کوئری ایدی ها، داک ایدی هایی که مرتبط با هر کوئری هستن را مشخص کرده. این را میخوانیم تا بعدا برای تست کنیم که ببینیم برنامه ای که نوشتیم تا چه اندازه توانسته اسناد را به درستی بازیابی کند.

سپس توابع برای معیارهای ارزیابی را تعریف کردیم که شامل دقت، بازخوانی، f-measure، map، dcg.

و بعد از آن مدل های وزن دهی شامل lnc-itc، lnc-inc، lnn-lnn، ltc-ltc، ltn-ltn و bnn-bnn.

و سپس با بررسی دو حالت حذف کلمات توقف یا نه و همچنین ریشه یابی یا نه، با انواع مدل های وزن دهی اسناد را بازیابی کردیم و برای هریک معیارهای ارزیابی را محسابه کردیم و همه اطلاعات را در result ذخیره کردیم. نتیجه چاپ و در یک فایل output.csv نیز ذخیره کردیم که در کنار این فایل قرار دارد.