بسمه تعالی

گزارش کار پروژه پایانی درس بازیابی اطلاعات

استاد : دکتر حسین امیر خانی

روح الله مظفری

۹۵۱۳۲۰۰۰۵۵

مقدمه :

در این گزارش شیوه جمع آوری اطلاعات از سابت دنیای اقتصاد و همجنین روش tf-idf مورد برسی قرار خواهذ گرفت.

سر فصل ها :

۱ - جمع آوری اطلاعات (Crawling)

برای جمع آوری دیتا نیاز به دانلود صفحات مختلفی از سایت مورد نظر به شکل html داریم. در این جا از کتابخانه requests زبان پایتون استفاده می کنیم. فرمت url سایت برای صفحات خبری به شکل زیر است

```
https://donya-e-eqtesad.com/%D8%A8%D8%AE%D8%B4-%D8%AE%D8%A8%D8%B1-64?np={news_n
{umber
```

که new_number شماره خبر مورد نظر است. تتها نیاز است که این لینک را با کتابخانه requests دانلود کنیم. پس از آن یک ابجکت از کلاس *BeautifulSoup*

می سازیم . BeautifulSoup به ما کمک میکند که صفحه دنلود شده با به متغییر های قابل دسترسی(tag های html) تغییر شکل دهیم. با کمک متد find_all تمام لینک های قرار گرفته در تگ های h2 که عناوین خبری ما هستند را استحراج میکنیم. این لینک ها ما را به سمت صفحه ی اصلی هر خبر هدایت می کند. برای دانلود هر صفحه خبر و استخراج داده از آن از کتاب خانه newspaper استفاده می شود. این کتاب خانه متن و عنوان خبر را پس از دانلود کردن صفحه خبر ذخیره میکند. این عملیات را تا پایان پیدا کردن اخرین خبر ادامه می دهیم. دیتای استخراج شده شامل لینک صفحه خبر و عنوان و متن آن می باشد. این دیتا را توسط کتاب خانه pandas به شکل یک ابجت DataFrame تغییر میدهیم. پس از آن این دیتا را در داخل یک فایل با فرمت csv ذخیره میکنم. این قطعه کد در داخل فایل scraper.py قرار دارد. پس از اجرای ان فایل دیتای جمع آوری شده در فایل csv با نام new_data.csv قابل مشاهده خواهد بود.

2-مقایسه داکیومنت های مشابه با روش tf-idf

برای مقایسه و پیدا کردن داکیومنت های مرتبط با کوییری کاربر از روش rf-idf استفاده میشود. ابتدا توسط کتاب خانه pandas فایل csv مورد نظر را read میکنیم و آن را به فرمت ارایه تغییر می دهیم. ابتدا یک شی ااز کلاس TfidfVectorizer می سازیم و تمام داکیومنت های مورد نظر را fit میکنیم. برای تمام کوییری های کاربر متد find_query صدا زده میشود. در این جا ابتدا کوییری توسط vectorizer زا به شکل tfidf در می اوریم. پس این مرحله ما vector های تمام داکیومنت و کوییری کاربر را در اختیار داریم کافیست که این دو را مقایسه کنیم. به دلیل محدود بودن جواب برگشتی که در این تست عدد ۱۰ ثبت شده این فرایند مقایسه به صورت خطی انجام میشود. مقایسه vector ها توسط متد cosie_similary انجام میشود. پس از مرتب کردن cosine های بدست امده ۱۰ داکبومنت با بیشترین مقدار شباهت را باز میگردانیم. در دایرکتوری samples تعدای کوییری همراه با داکیومنت ها و میزان مشابهت آنها به ترتیب نزدیک ترین داکیونت به کوییری ذخیره شده است.

کد های مربوطه در صفحه ی گیت هاب به آدرس

-https://github.com/ruhollahmozafari/tf-idf-method-for-information-retrieval

قابل مشاهده می باشد.