محمدامین پوردرخشان

9713200012



در مرحله اول کتابخانه های فوق را اضافه میکنیم

requests: برای ریکویست زدن به سایت مورد نظر استفاده میشود.

Beautiful soap: برای استخراج تگ های HTML مورد نظر و همچنین استخراج href .

Pandas: برای نوشتن دیتا روی فایل CSV



متغیر Page مشخص کننده url اول است .

متغیر scrapted data یک آرایه از دیتا های استخراج شده است .

متغیر page num تعداد صفحه هایی که میخواهیم پیمایش کنیم است.

یک whileکه همیشه trueاست که در آن صفحه ها پیمایش میشوند و هربار به متغیر pageیک عدد اضافه میشود و زمانی که متغیر page به page num میرسد breakاتفاق می افتد.

با استفاده ازکتابخانه request ما html صفحه مورد نظر را get میکنیم .

با استفاده از کتابخانه beautiful soup یک ابجکت soupدرست کرده و html را به آن پاس میدهیم.

با متد find all از ابجکت آن تگ های مورد نظر را استخراج میکنیم و سپس hrefرا استخراج میکنیم و urlمد نظرمان را میسازیم.سپس با یک حلقه تمام url های مد نظر را با استفاده از ابجکت article دانلود کرده و سپس parseمیکنیم و دیتای بدست آمده را با استفاده از pandasدر یک فایل اکسل ذخیره میکنیم.



کتابخانه feature extraction برای برداری کردن داده ها

کتابخانه pairwise برای cosine کردن داده ها .

در وهله اول فایل را با فرمت utf\_8 باز میکنیم. سپس با csv.readerفایل را میخوانیم

متغیر docsرا برای ذخیره کردن متن استفاده میشود.

حلقه for برای اضافه کردن متن از فایل به متغیر docs

یک متغیر vectorizer از ابجکت tfidfvictorizer میسازیم و دیتای docsرا با آن فیت میکنیم .

بعد تابع fit-transform را صدا میکنیم

Query preprocessing : در مرحله اول کوئری مد نظر را با استفاده از تابع voctorizer transform مقدار tfidf را بدست میاوریم.

متغیر cosines برای بدست اوردن نزدیکی کویری ها با داکیومنت ها .

یک حلقه برای اضافه کردن cosine های مد نظر به متغیر cosines

با استفاده از np متغیر cosinesرا سورت کرده و k تای اول را برمیگردانیم.