گزارش­کار پروژه Scrape

سید علی نبوی 9813200077

در این پروژه به crawl سایت کجارو پرداختیم. این سایت در رابطه با گردشگری است و اطلاعاتی در مورد سفر و مقاصد مهم گردشگری را در خود دارد.

در قدم اول تابعی به نام crawl می­سازیم و 3 متغیر به شرح زیر تعریف می­کنیم:

page: این متغیر نشانن دهنده­ی صفحه­ای از سایت است که درآن قرار داریم.

data: در این متغیر ما دیتای مورد نظر که از صفحات استخراج کرده­ایم را نگه می­داریم.

urlList: این متغیر شامل لیست تمام url های دارای اطلاعات ما است.

در قدم دوم داخل حلقه while که همیشه درست است( به جز زمانی که تابع repetitive برابر با true شود) 4 متغیر در آن تعریف می­کنیم.

main\_url: این متغیر برابر با URL سایت به علاوه شماره صفحه­ی آن است.

html: درون متغیر با استفاده از کتابخانه requests تگ­ها و کدهای html صفحه­­ای که در آن هستیم را ذخیره می­کنیم.

soup: از این متغیر برای خارج کردن و سپس ذخیره کردن دیتا و اطلاعات از میان تگ های html استفاده می­کنیم.

links: از این متغیر برای ذخیره کردن اطلاعات تگ مورد نظر(در این جا تگ h3) استفاده می­کینم.

سپس در قدم سوم شرط تکرار حلقه را چک می­کنیم. این شرط که درون تابع repetitive آنرا نوشته­ایم به این شرح است: این تابع به عنوان ورودی دو متغیر urlList و links را می­گیرد. سپس چک می­کند که آیا url جدیدی از طریق تگ h3 به داخل متغیر ریخته­ایم قبلا چک شده است یا خیر. اگر چک شده بود شرط توقف صدا زده می­شود و این یعنی به آخرین صفحه رسیده­ایم. اما اگر چک نشده بود تابع به کار خود و استخراج دیتا ادامه می­دهد.

در قدم چهارم لینک صفحاتی که می­خواهیم از آنها دیتا استخراج کنیم را در متغیر sub\_url ذخیره می­کنیم. همچنین این لینک ها را برای چک کردن شرط در urlList نیز ذخیره می­کنیم. سپس از طریق لینک صفحه و دادن آن به کتابخانه Article صفحه­ی مورد نظر را دانلود می­کنیم و دیتای مورد نیاز خودرا از آن صفحه استخراج می­کنیم و داخل متغیر data ذخیره می­کنیم.

در انتها و در قدم پنجم دیتایی که ذخیره کرده­ایم را تبدیل به فایل اکسل کرده و آنرا ذخیره می­کنیم.

[لینک github](https://github.com/sali1290/Kojaro-scrape-uni-project-)