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 می-دهد.

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