Indeed.com Scraper Handbook

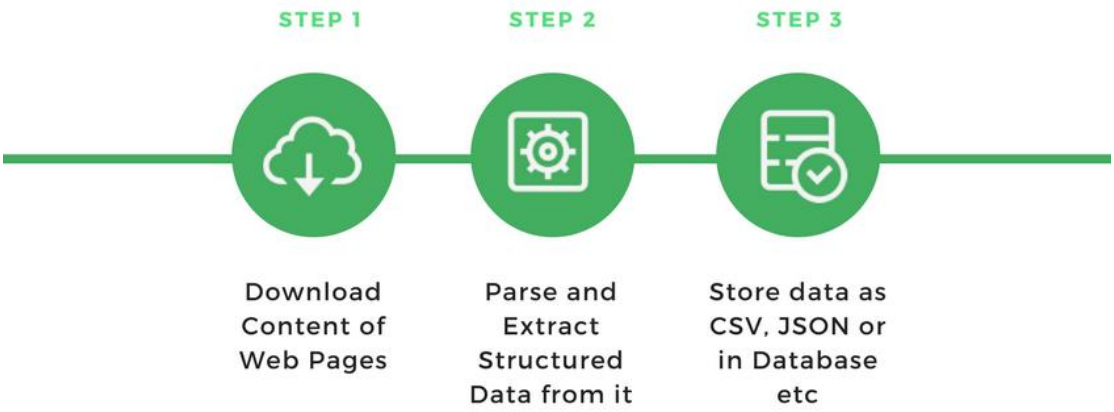
By Xiaoxuan Ma

What is web scraping

Web scraping is used to extract or "scrape" data from any web page on the Internet.
A Web scraper is built specifically to handle the structure of a particular website. The scraper then uses this site-specific structure to extract individual data elements from the website.

How does a web scraper work?

A web scraper is a software program or script that is used to download the contents (usually text-based and formatted as HTML) of multiple web pages and then extract data from it.



Libraries used for Web Scraping

As we know, Python is used for various applications and there are different libraries for different purposes.
Selenium: Selenium is a web testing library. It is used to automate browser activities.
BeautifulSoup: Beautiful Soup is a Python package for parsing HTML and XML documents. It creates parse trees that is helpful to extract the data easily.
Pandas: Pandas is a library used for data manipulation and analysis. It is used to extract the data and store it in the desired format.

HTML Structures

Scrapy Framework

Operation of Indeed Scraper

Step 1 Finding the matched URLs

We need have company names stored in a CSV file or XLSX file. Then change the path and file name in the codes.

Step 2 Scraping

Environment: Anaconda Prompt

Copy the entire file named indeed in the same location of your Anaconda. Do not change any name of the file.

Install the Scrapy framework:

Anaconda Prompt

```
(base) C:\Users\MXX>pip install scrapy
```

Change several key words:

For pipeline.py

pipelines.py

```
28  class WriteItemPipeline(object):
29
30      def __init__(self):
31          self.filename = 'reviews_CN.csv' #change the file name
32
```

For indeed_s.py

```
6   urls = pd.read_csv("C:/Users/MXX/Desktop/cleaned_company_list.csv") #change the path and file name
```

```
14      def parse(self, response):
15          for url in urls:
16              yield scrapy.Request("https://www.indeed.com" + url + "/reviews?fcountry=CN",callback = self.parse_single)
17              #change the country code
```

Then run the scraper in Anaconda Prompt:

Anaconda Prompt

```
(base) C:\Users\MXX>cd indeed

(base) C:\Users\MXX\indeed>cd indeed

(base) C:\Users\MXX\indeed\indeed>cd spiders

(base) C:\Users\MXX\indeed\indeed\spiders>scrapy crawl indeed_s
```

Step 3 Rearrange the data

In this step, we'll order the columns, drop duplicates and make a list of companies and numbers of reviews for each company.