Web Scrapping: Data Extraction from Websites

Iqtibas Salim Hilal Almaqbali MIDDLE EAST COLLEGE
Fatmah Mohammed Ali Al Khufairi Middle East College
Mohamed Samiulla Khan Middle East College
Anjum Zameer Bhat Middle East College
Imran Ahmed Middle East College

Data is very important nowadays for almost all organizations for their existence as well as for their growth. The Internet has become the major source of data for individuals and almost all organizations. Authentic Websites are a major source of reliable data for many individuals and organizations. Extracting Data from websites is commonly referred as Web Scrapping, which refers to both manual and automated process. Extracting large amount of meaning full data from the websites manually is very difficult, tedious and redundant task. Automated Scrapping is done by writing specific programs to extract the required data from the websites. These programs are usually called as web scrappers. Web scrappers are written using many programming languages like Python, Node.js, Ruby, C++, PHP etc. Each language has its own unique features and built in libraries for performing data extraction. There are many web scrapping tools like Beautiful Soup, Octoparse, Parsehub etc. In this article we are going to analyses few recent Web Scrappers tools used in scrapping the Web.

Introduction

Data is the backbone of any organization. Nowadays organizations are creating data management roles to acquire, access, validate, store, protect and process data. Tim Berners-Lee describes the World Wide Web is web of data(Tim Berners-Lee,1999).This data is used for taking strategic business decisions required for sustenance and continual growth of the organization. Huge data is available because of internet. As internet is growing day by day .Its services are becoming cheaper and it’s used in unimaginably in large scale by business organizations to run their business efficiently. World Wide Web is such one of the services offered by internet. The ability of Uniform resource locators to uniquely identify information system consisting of documents and other web resources, which can be accessed through internet world wide using hyperlink has revolutionized the world of data. Like this these Websites have become a major source of data. This huge amount of data present in various websites having different formats and heading requires some kind of processing if we want to make sense of it. One way of solving this problem is to manually save copy of data from a single or multiple websites, even though this work is very hard ,time consuming and prone to error. Further Many websites do not permit copy of data to be saved to your system for further processing (Penman et al., 2009). Attempt was made to solve this problem of extracting data from different websites using a technique called Web Scraping, which allows us to extract data from multiple websites for further processing. We are going to illustrate the working the data extraction using Python Programming Language. Also we are going to analyses few of the latest tools available for data extraction.

Why Web Scrapping

1. Comparing prices of items from online websites and displaying in one place with our own format.
2. Email Address are gathered by companies those who use emails for marketing
3. Social media tools are scrapped to find the topics which are trending so as to use the same for marketing purposes.
4. Listings –any kind data is collected is collected and listed from different websites having information like jobs, scholarships , etc
5. Predicting future trends by collecting and analyzing data through web scrapping techniques.
6. Weather monitoring.
7. Website change detection.

Basic steps used for web scrapping using Programming Languages

1. Find and Examine the Web Page you want to scrape.
2. Identify the required data you want to extract.
3. Write appropriate code using python language.
4. Execute the code to extract required data and stored in the required format.

Illustration of Web Scrapping using python Language

Step 1:

We are interested in extracting information related with jobs which are listed in the following Webpage

url = https://boston.craigslist.org/search/npo

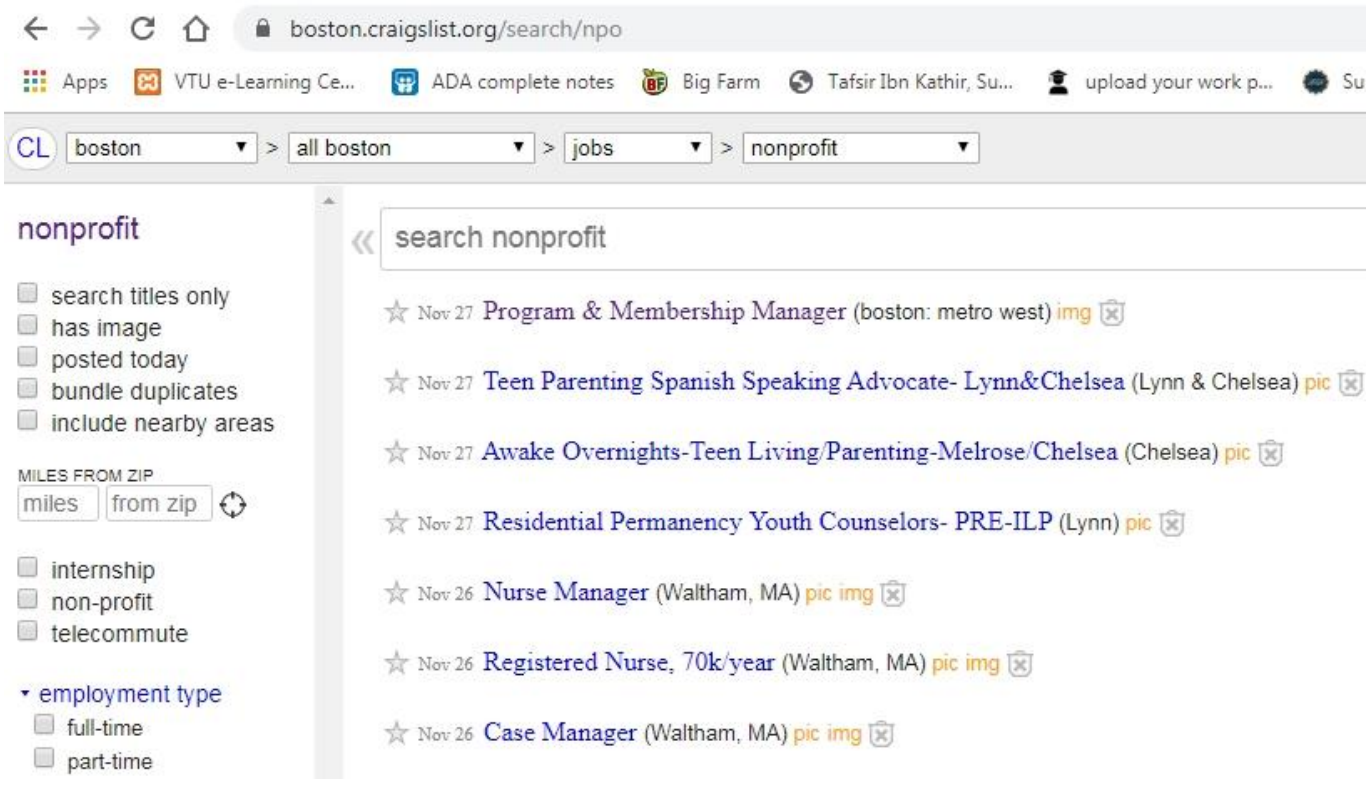

Figure 1. *Screen shot of selected url to extract.*

Step 2:

Job Title, Location ,Date ,Link ,Job Attributes, Job Description are the fields we are interested to extract from the given web page.

Step 3:

Appropriate code((GoTrained Academy, 2019)) is written by importing built in libraries available like BeautifulSoup , pandas, requests

```
from bs4 import BeautifulSoup

import requests

import pandas as pd

code…
```

Step 4: Once code is executed, required data is extracted and can be stored as CSV(comma separated values) or any other format desired by the programmer.

Figure 2. *Snapshot from variable table from Python Editor*

Figure 3. *Snapshot of page (corresponding to first entry)*

Brief Survey of Scrapping tools

Visual Web Ripper

Visual Web Ripper is a great web page scraper developed by Sequentum , which can be used to extract data from website by clicking on the content data elements you are interested to extract like catalogs of product, classifieds, financial web sites ,etc.

ScrapeSimple

ScrapeSimple is very easy to use tool for Web Scrappers .Using tool is just like filling out a form with commands for what type of data you wish to extract. It can be customized to extract data periodically (weekly, monthly, etc) from the websites and send you through email in CSV format.

Octoparse

It is a user friendly yet powerful tool developed by Octopus Data Inc. which can be used to scrape website to get data in different formats like CSV, Excel. Data can also be directly saved to databases.

ParseHub

Parsehub is an amazingly influential tool for constructing web scrapers without coding developed by a company based in Canada by name ParseHub. Data can be extracted into Excel, CSV and JSO format. Images can be scrapped and downloaded using this. ParseHub can be used by anyone interested in extracting data.It can be used by analysts and consultants, sales leads, developers, journalists and data scientists, etc.

Scrapy

Scrapy is an open source code ,which can be used by Python Programmers to develop Scalable Web Crawlers. It runs on Windows, Mac, Linux, and BSD. It has very good documentation and lots of tutorials available to start using and developing web scrappers. It is supported by healthy social media platforms like GitHub,Twitter,StackOverFlow.

Conclusion

In this paper an attempt is made to introduce the concept of Web Scrappers and their importance in today's world. A Data Extraction Process was demonstrated using Python programming language. Few latest web scrappers tools are analyzed.

References

GoTrained Academy (2019). Learn Web Scrapping with Python from Scratch. [video] Available at: https://www.udemy.com/course/web-scraping-python-tutorial/learn/lecture/12934390#overview [Accessed 30 Nov. 2019].

Daniel Ni, (2019). "The 10 Best Data Scraping Tools and Web Scraping Tools."

T. Baldwin, D. Martinez, and R. Penman. (2007). "Automatic thread classi_cation for linux user forum information access." ADCS 2007, 72–9.

T. Berners-Lee and M. Fischetti (1999). ". Weaving the Web" HarperOne, San Francisco, USA, 1999.

Penman, R.B., Baldwin, T., Martinez. (2009). "Web scraping made simple with site scraper."