ISSN: 2455 - 9172

EIJO Journal of Engineering, Technology And Innovative Research (EIJO–JETIR)

Einstein International Journal Organization (EIJO)

Available Online at: www.eijo.in

Volume – 5, Issue –1, January – February - 2020 , Page No. : 01 – 07

Web Scraping Technology

[1]Ms. Akansha Agarwal, [2]Ms. Mahima Saini, [3]Dr. Himanshu Arora and [4]Ms. Shilpi Mishra

[1,2]B.Tech Student, Department of Computer Science & Engineering, Arya College of Engineering & Research Centre, Jaipur

[3]Professor, Department of Computer Science & Engineering, Arya College of Engineering & Research Centre, Jaipur

[4]Assistant Professor, Department of Computer Science & Engineering, Arya College of Engineering & Research Centre, Jaipur

Abstract

Internet has the vastest information and the data sources ever built by mankind. For the evolution of World Wide Web, the scenario of internet user and data exchange is fastly changes. Due to all changes, large number of users joined the internet and use the facilities. By the daily use of internet, large amount of data is available on internet and data plays an important role in every field. Researchers, market analyzer or academicians, Businessman all are share their advertisements, information on internet so that they can connect the people easily. To share and store data on internet, a new problem arises that how to handle such overloaded data and how the user will get or access the web information in least efforts. To solve this is problem, a new technique is used i.e. web scraping. In this paper, our main focus is on recovering the web information using python script. The greater part of the web data shows in unstructured configuration. Thus, web scraping is utilized for separating the unstructured information from the sites and transformed into organized way.

Keywords: Web scraping, Selenium, Beautiful Soup, HTML Tags.

Introduction

Data plays an important role in the field of marketing, researchers, business etc. Many users gather data from different websites for their better improvement. Some of the them uses the traditional method i.e. same part copying, Text grabbing and regular expression matching to extract the data from the website. Many users uses the copy-paste technique for gathering and analyzing data on the internet.[7] Copying of data on the website to user local storage in forbidden by most of the website authority. So that the user wants to manually copying the data from website to local computer file storage. But this technique is very tedious and time consuming.

Ordinarily information isn't effectively open through website in spite of the fact that it exist. As much as we wish everything was accessible in CSV format or our preferred organization generally used information which is distributed in various formats on the web. Here, copy-paste technique is not utilized. Because of such restriction of this method, web scraping is used. As compared with copy-paste method, web scraping is most effortless web scraping procedure. [5]

The mechanized assembling of information from the internet is almost as old as the web itself. In spite of the fact that web scraping is not the another term in years past the training has been all the more ordinarily known as screen scrapping, information mining, web harvesting or comparative variation. It is the act of processing to retrieve the information through any methods other than of program intersecting with Application Programming Interface (API). The essential and significant aim of the web scraping process is to mine data from an alternate and unstructured web-site and change it into an comprehensible structured like spread sheet, data-base or a comma-separated values (CSV) file. For example- Information like item pricing, stock evaluating, various reports, market analysis, online value correlations and product detail can be assemble through web scraping. Extracting specific form of data from site contribution to take compelling choice in business process.[11]

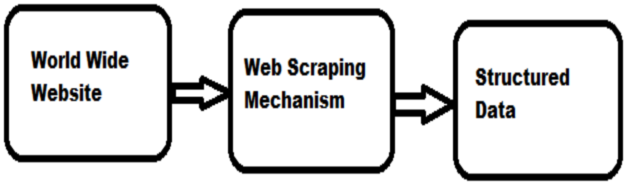


Figure 1- Basic Architecture of web scraping

As we probably aware, Web-scraping is a most popular program that work as a software to extract the data from the website. These projects can simmulate human web –surfing conduct by hyper text transfer protocol(HTTP) or an internet browser service. Web scraping concantrates on transforming unstructured information into organized information form that can be save and analyse in a spread sheet format or any database type. This procedure ends up being an invited change from the procedure of manually attempting together huge measure of information.. [10]

As we talk about the unsturucted information implies website pages are written in hyper text markup language(HTML) and all the more as late of XML type. Web document are represented by sequential structure this is known as object model, or simply the dom tree. The objective of html which is a markup language is to indicate the configuration of content displayed by the internet browser.

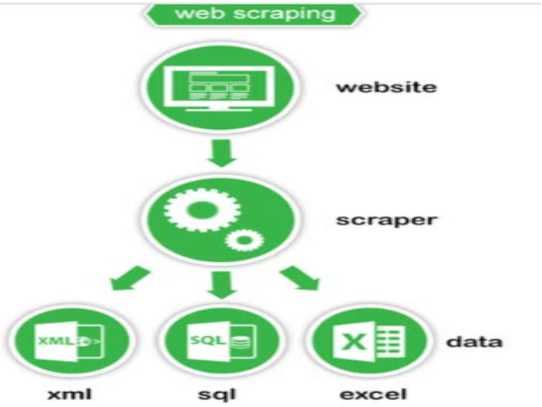


Figure 2 - Web Scraping Steps

Techniques of web scraping

(A) Simple Copy and Paste: The human manual assessment and copy-paste technique is the best and the useful web scraping method. But it is an tending to implement or cause errors. When user want to examine and save the lots of data sets.

(B) Content Pattern Matching: A easy and strong approach to retrieve the data from website pages which is basically based on the unix GREP command or regular expression matching facilities of programming language.

(C) Hyper Text Transfer Protocol (HTTP): By using these method user can be extract data from static and dynamic web pages. Information can be recovered by putting the HTTP request on the remote web server using socket programming.

(D) Hyper Text Markup Language (HTML Parsing): Structured data query language, for example X-query language and the Hyper-Text query language(HTQL) can be used to parse the HTML pages and to extract and change page content.

(E) Document Object Model (DOM) Parsing: By using the embedding a fully-arranged web browser such as a Internet Explorer or the Mozilla browser controlled, programs can retrieved the dynamic content produced by customer side scripts.[2] These browser controls additionally, parse web pages into a DOM tree dependent on which programs can retrieve parts of the pages.

(F) Web Scraping Software: Now a days, numerous tools are available that can be used to alter web scraping solutions. This software may attempt to automatically recognize the data structure of a page or give a recording interface that eliminates the necessity to manually write web scraping code, or some scripting functions that can be used to extract and change content, and database interfaces that can store the scrapped data in local database.

Web Scraping Related Work

Digital world is developing with a stride that exceed the speed of any artificial quickest prime movers. Here the term developing used in setting to estimate of information. If the worlds quickly using digital content were printed and bound into books. It will frame a stack that would starch from earth to Pluto multiple times. The principle supporters to this advanced distribution center are social media, government observation cameras and plane of other independent website which are refreshed in the day by day bases. Right now, this web information is the most fundamental asset for any business. [10]The foremost concentration of this paper to gather information though scraping as API are not available for every single information source. [6]

Scraping is the most prominent regions in the domain of huge information and sentiment analysis. A number of software product and tools are available in the market technology which are used in the system infrastructure.

Python used in web scraping

Many programming languages are used to implementing web scraping technology. But, mostly Python programming language is used. Because Python is a popular tool for implementing web scarping. Python programming language is also used for other helpful task related to cyber security, advanced legal applications. Utilizing the base programming of Python web scraping can be performed without utilizing some other outsider apparatus. Reasons for using python for web scraping-

(A) Syntax Simplicity: Python has a simplest structure when compared to other programming languages. This feature of Python make the testing easier and a developer focus on a program.

(B) In-Built Modules: Another reason for using Python for web scraping is the in-built as well as external useful libraries which can be perform many implementation related to web scraping by using python.

(C) Open Source: Python has huge support for the community due to this it is an open source programming language.

(D) Wide Range of Application: Python can be used for various programming task ranges from small shell script to enterprise the web applications.[4]

Python libraries used in Web scraping

A portion of the Python libraries are utilized in web scraping are as follows-

(A) Requests: Requests is a most significant python library utilized in web scraping. It is designed to simplify the process of making HTTP request. [9]This is highly valuable for web scraping because the first step in any web scraping work process is to send a HTTP request to the site server to recover the information shown on the web page.[13]

(B) Beautiful-Soup: Beautiful-Soup is a library intended to parse information that is to retrieve information from HTML or XML documents. Beautiful soup can only parse the data however, can not retrieve the site pages. A few highlights of beautiful soup are- [1]. This library give the some basic strategy and pythonic figures for navigating, searching and altering the parse tree. It does not take more code to write an application.[6]

(C) Selenium: Selenium Python is an open source web based automation tool which provides a simple API to write functional or acceptance test using selenium web driver. It is basically a set of different software tool each with a different approach to supporting test automation. With the help of selenium Python API a user can access all functionalities of selenium web driver in an intuitive way.[12]

(D) URLLIB2: urllib2 is an refreshed version of urllib, this library has a rich arrangement of capacity contained it into open the urls. The primary link of the web page, valuable bits are covered up under hyperlink on the page. This module characterize then following function.[9]

urllib2.urlopen(url[data[,timeout[,cafile[,capath[,cadefault**[,context]]]]**

[1]. url: url can be string as well as the request object. [2] data: this portion is utilize if there is an extra information to be sent to the server. Timeout If there must be break after certain number of effort, timeout, which is an optional parameter is used. cafile: They indicates set of trusted ca-authentication for HTTP request.[3]. Context this parameter manage with SSL (secure socket layer).

Scrapy: Scrapy is an open source and collaborative framework for extracting the information a client wants from the site. It is intended to scrape web content from sites that are made out of numerous pages of comparable semantic structure. The framework is executed as a fire-fox browser extension, and works in three key stages to scrape web information. Initial, a client explores to a page like to scrape and creates a prototype for the content. Next, the client chooses a set of links that point to pages matching the content template defined by the client. At last, the client chooses a yield information design and scrappy crawls the links indicated by the client and scrapes content related to the client template.[10]

Process of web scraping

We have a number of tools and libraries available for scraping a web data, some of them are briefly describe in above section. The act of scraping includes moving through the seed archives that is a web page and collective or assembling their necessary information in the most suitable format. One of the process of this technique is the structure of the document data is preserved. This seed document is parsed by passing it to the beautiful soup constructor. [1]

In dreadful terms it is refer to as creation of the soup. HTML parsing of any archive start with:

soup = Beautiful Soup(open(―URL))

The HTML document is chunked into python object by beautiful soup. These items are mainly: tag, navigable-string, beautiful-soup and remark.

TAG: These are normal tags found in HTML. Likewise, these labels contains attribute in many cases

<tag @attribute=‖att1‖>

</tag>

Navigable-String: The content found between the opening and the closing HTML tag of document is navigable string object.

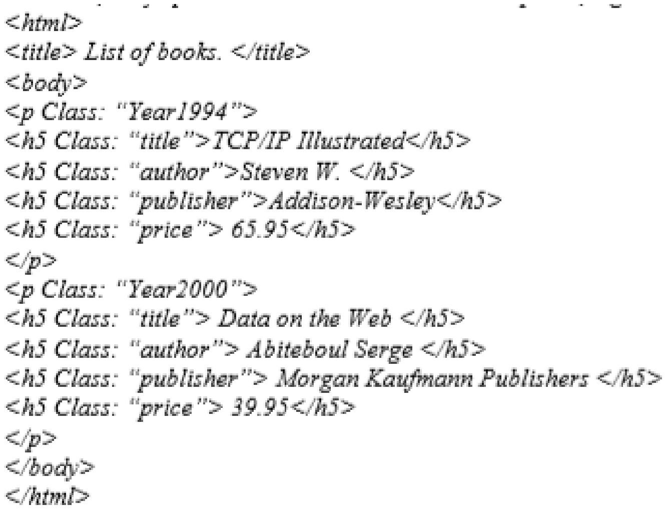
<tag @attribute=‖att1‖>

Navigable String

</tag>

BeautifulSoup: These object contrasts from normal HTML tag for it needs name and attribute that are available in tag object.

This shifting of information is clarify considering an example seed document. The document “Book.html” as shown in the figure3 contains the details of various books: year, title, author, publisher, price.[13] Through the whole detail is available, only specific details for the book are required. Example: Title of the book published in 2000[11].



```
<html>
<title> List of books. </title>
<body>
<p Class: “Year1994”>
<h5 Class: “title”>TCP/IP Illustrated</h5>
<h5 Class: “author”>Steven W. </h5>
<h5 Class: “publisher”>Addison-Wesley</h5>
<h5 Class: “price”> 65.95</h5>
</p>
<p Class: “Year2000”>
<h5 Class: “title”> Data on the Web </h5>
<h5 Class: “author”> Abiteboul Serge </h5>
<h5 Class: “publisher”> Morgan Kaufmann Publishers </h5>
<h5 Class: “price”> 39.95</h5>
</p>
</body>
</html>
```

Figure 3- Book.html

Prior detailing the algorithm, lets first discusssed the outcome. Here the “soup” is made out of the url ‘Book.html’. The tag <body> contains detail of the two books inside <p> tag. The details are wrapped into <h5> tag along with appropriate attribute. The target information to be scrapped is contain inside the letter <p> tag. Over here the fundamental concern is over <h5> tag with the {class:title} attribute since its navigable string consist of the necessary title of the book. Parsing these lines should bring about information on the web as it it is distributed in the year 2000. Following algorithm shows python algorithm that scraps the required information as appeared as shown below:

```
frombs4import beautifulsoup
soup= beautifulsoup(open(“book.html”))
for all in soup:
book =all.findAall(“body”)
for IinsideBody in book:
AllParagraphTags=insideIBody.findAll(“P”,{“class”:”2000”})
```

```
For InsideParagraphTag in AllParagraphTag
 RequiredTitle=InsideParagraphTag.findAll(“h5”,{“class:”title}).string
PrintRequiredTitle
```

Figure 4- Needed data

Web scraping uses

The consistently improving innovation of scraping has come out into stars in many key territories of the universe. This field with dynamic advancements share a common objective with the semantic web vision, an aspiring activity that despite everything requires get through content handling, semantic understanding and human-pc association. A portion of the significant enhancements are as per the following:

(A) Freedom from local API: Scrapers are not relying upon local API for recovering the data. A scrapper is concern is to recover the human read able pieces of the website page and in the world web, html is utilized to represent the information. The scrapers, hence build to these HTML labels of the pages and acquire the data structure form. Therefore, the sites that doesn’t give API would now be able to scraped effectively and fastly.[1]

(B) Building repositories for web search engine: Search engine are fundamentally reliant on crawlers. Crawlers, beginning from the seed URL trends to every URL found on its way. Here, Scrappers can spare the content of every single URL gave the machine being utilized has astonishing existence unpredictability.

(C) Investigation purpose: Any business bigwigs worry with the client care or the national organisor like metrological office can utilized the gathered information to improve their working. For example, metro sensible office gathers the information for the future forecast and scrapper can be utilized to note down the information from almost any city of the worlds.[2] Also versatile organization can scrap audits for its new launch from various website and improve contingent upon example of fulfilment or disappointment obtained from purchasers survey.

(D) Classification or Arrangement: Content arrangement (Text categorization- otherwise called content order, or point spotting) is the task of undertaking of consequently arranging a set of document into classifications (or classes, or subjects) from a predefined set [9]. Scrappers are fit to carry out this responsibility for focusing on the key parts that shows class of the archive (or document).

(E) Web scrappers are also being used by online marketers to pool data privately from the competitors’s website such as high targeted keyword, valuable links, emails and traffic sources.

Conclusion

This paper gives in website webpage of extracting the data from website in case whether information isn’t given through API or in case where it is hard to get the data because of structure or arrangement of the website page. To get the automatic information from the website web scrapping is the most efficient techniques. Among all other techniques mention is this paper which are used to extract and store data , web scraping is more reliable and fast and automatic data retrival system By using web scraping technology users can easly extract unstructured data on single or multipal websites which can be stored into aspecific formate and analyse for the future used. The main aim of these technique is to gain information from web and aggregate into a new data set.

References

1. S.C.M de S Sirisuriya2015, A comparative study on web scraping. Proceeding of 8th international research conference, KDU.
2. List of Web harvester, data scraper, web scrapping, software and tools, n.d. web data scraping. URL https://web data-scraping.com/web scraping-software/.
3. Felipe Jorbao Almeida Prado Mattosinho.master thesis, mining product opinions and reviews on the web, TU Dresden.
4. Data toolbar.computersoftware.datatoolbar.2013. Web.
5. MCAFEE, Andrew and Erik Brynjonfsson. "Bigdata the management evolution" Harvard business review. Hank Boye, 1October 2012 web 8 April 2016.
6. Text categorization by Fabrizio Sebastiani bipartimento Dmentematica Pura E applicate universta Di padova 35131 padova, Italy.
7. Beautiful soup documentation_ www.crummy.com.
8. Screen scraping: Techopedia.
9. Urllib2_extensible library for opening URLS: https://docs.python.org.
10. Scherenk, M.Web boats, Spyders, and screen scraper: aguide to developing internet isn't with PHP/curl no starch press, 2007.
11. Deepak kumar mahto, lisha singh, A drive into web scraper world, 2016 International Conference on computing for Sustainable global development (Indiacom) 978-9-3805-4421-2/16/$31.00 C,2016 IEEE.
12. Amer Jahazvinian, sean Holbert, Nikil,Viswanathan, scrappy, simple, webscraping, department of bio-medical informatics, department of computer science , Stanford university .
13. ELOISA Vargiu, Mirko,URRU 1,2013 exploiting web scraping in a collaborative filtering/ based approach to web advertising, artificial intelligence research. 2013, Vol.2, No. 1, https://dx.doi.org/10.5430/air.v2n1p44.