

The Benefit and Risks for Scraping Based on Python

Zhenhao Xie *

Tunghua Wenzel international school, Dongguan, China

* Corresponding Author Email: zxie@dgtwis.com

Abstract. As society's demand for the web has become greater, web crawler technology has been introduced and widely applied in various applications. Web crawling technology has a very controversial nature, with many people loving to use it, but many resisting. This study will discuss the definition of web crawling technology and analyze the reason why it is popular but also resisted by the public, and demonstrate the detail approach to implement web crawlers on Python. This leads to the conclusion that web crawling is a web technology that can save time and labor and can be applied to data investigation. The reason for its popularity is that it saves time and can automatically visit the desired web pages and retrieve data from them. However, it is resisted for many reasons, one of which is that it can cause damage to the interests of others, because web crawlers can collect paid data by special means. Python is the best language for running web crawlers because Python is the programming language that most resembles human language.

Keywords: Web crawler; Python; Data; Collect.

1. Introduction

With the progress of human to digital society, there were huge numbers of data or information published on the internet. People are constantly looking for a technology that can make search more efficient in order to reduce the time and increase the accuracy of search. Therefore, people created the first web scraping in December 1993, name as JumpStation, the first online robot-based search engine on the World Wide Web (WWW). At that time, search engines relied on human administrators to gather and format links because there were fewer websites available on the internet [1]. Since, the needs of people using WWW was rapidly increasing, the JumpStation cannot satisfy people's needs. Therefore, the programmer created the Applications Programming Interface (API), can enable the download of enormous collections of heterogeneous information and their connectivity in a much more robust manner [2]. Web scraping is technique that can be used to accurate and rapid discover the information from the internet [3]. Now, there are a lot of programming languages available to write Web scraping code, for example Python, Java, JavaScript. However, Python Web scraping is the most versatile and efficient because Python is a relatively simple language that can accomplish what other languages need more code to accomplish with less code, so Python has become one of the most used languages by Web scraping [4].

Web scraping has four main applications, which can be used to enhance purchasing power, to understand consumer behavior, to enhance marketing strategies, and to investigate useful information on the Internet [3]. For enhancing purchasing power, Web scraping can quickly and accurately search for people needs, which can stimulate people's desire to buy at the first time. It can also find a lot of resources that about the need of people, that people will be more likely to buy. For understanding consumer behavior, the merchants can use Web scraping to analysis and know the consumer's action in order to do the new marketing or business plan. For enhancing marketing strategies, the market researcher can use it to contact the customer's information, e.g., recent demand, phone number and social media accounts. Therefore, the researcher can specify the right program for the potential population [5]. To investigate useful information on the Internet, it is the most popular function of the Web scraping, when the data collector collects the data, it takes just a few lines of code to get the desired data. Hence, more and more people are using Web scraping. On the other hand, Web scraping are visiting the site a large number of times to get the data they need, and too many times can crash the site. In this case, some "negative user" might use it to attack the website. This has led to Web scraping also being boycotted by a large number of website developer and governments.

The overall purpose of this thesis is to discuss and analysis the advantages and disadvantages of Web scraping based on python with specific cases that will enable data collectors to know and understand how to use Web scraping technology properly and legally and how web developers should protect against illegal Web scraping. Moreover, it will discuss and analyze whether web crawler technology should continue to be used. The structure of this discussion is to introduce the background of Web scraping first, which divided into background information and applications method, some of which are better explained by photographs. Second, the paper will analysis the benefit and risks with some specific cases and data, that can be described more authoritatively. Afterwards, this study will argue the current limitations and prospects of Web scraping and make some recommendations based on the case studies. Last, the content of the whole article will be summarized.

2. Basic Descriptions

Web scraping (Web crawler) is the tool or program to automatically and numerous obtain the information of the website by downloading the web document [6]. It allows the user's code to better access the information the user wants and exclude the information that is not useful to the user. Web scraping is the practice of collecting data via "scraping" data from websites. Although it is theoretically feasible to scrape more than just websites, i.e., things like papers and games may make good candidates, the majority of scraping is often directed at websites [7]. It is a tool used by search engines and other information searchers to collect material for indexing and allow them to maintain current databases. All search engines utilize internal web crawlers to maintain current copies of the material. The search engine is divided into many components. The search engine depends most heavily on the crawler module out of all of these modules since it helps the search engine provide the best results. Little programs called crawlers "explore" the web on behalf of search engines in a manner akin to how a human user would click links to access various pages [8].

The idea for Python originated in the late 1980s, and Guido van Rossum of CWI in the Netherlands built it in December 1989 as a replacement for the ABC language (which had been influenced by SETL) that could handle errors and communicate with the Amoeba operating system. The nickname given to van Rossum by the Python community, the "benevolent dictator of life," reflects the fact that he is the main author of Python and that he still plays a crucial part in shaping Python's trajectory [9]. Moreover, Python has grown to be one of the most widely used programming languages, and as of November 2020, it was rated second only to C in terms of popularity [7]. It is easy to read, write and understand because it just like human languages [4]. For example, while the Python executes the "Print" code, it only uses as "print ('')". On the other hand, the other programming languages need to use the more cumbersome "print" function.

If the data analysts need to run a web crawler on python, they need to install the package requests in python first, because the requests library is a third-party library for python and cannot be used directly by python. Next, they need import requests in Python, that it can be used in the python. Then, they need copy the Uniform Resource Locator (URL) of the website and store it in a variable. Next, one can pressure the Ctrl+Shift+I buttons to open inspect elements window to find and click the Network buttons. After that, one refreshes the page, then clicks on the first one that appears inside the network. Next it is necessary to make sure the request Method is "Get". After that, one pulls the scroll to the bottom of the side, that can see the "user-agent". To copy it and storage it in a variable. This way the required conditions are collected. Finally, just use "res=requests.get (url= , headers=)" to fill in all the variables, and then use "print (res.text)" to print out the network's Source code.

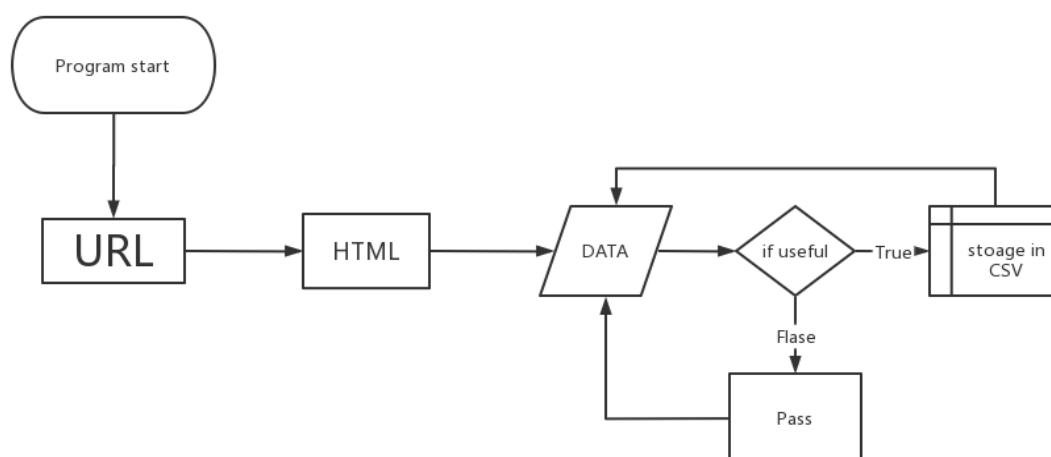


Fig 1. Process of how Python Web Scraping work.

3. Benefits for Web Scraping Based on Python

On the one hand, there are a large numbers of the advantages for Web Scraping based on Python. It offer a tool for data analyst, which might have batter understanding on the information and data, that they cost less time, energetic and workforce. In many cases, web crawler technology requires only a few lines of code to obtain the information contained in a web page or HyperText Markup Language (HTML) [1]. On top of that the analysts can import a third-party library for Python - RegularExpression (RE) can improve the accuracy of the information to be retrieved. This allows analysts and investigators to write just a few lines of code to get the important information they need to analyze it. Thao has an Airbnb property in Kamppi, Helsinki, called Abbey Corner. The listing became active in April 2019 and aims to provide accommodation for travelers. As the author is one of the creators of the listing, Abbey Corner was chosen as a case study for this research. Through this study, the Airbnb market in Helsinki was investigated using web scraping techniques and it is expected that the findings will be useful for Abby Corner and other listings with similar ideas. Eventually the data they needed was obtained through web crawling techniques and this acquisition was accurate [3]. In a similar vein, the author uses RE and web Crawling to analyze Douban, one of the largest movie sites in China. The main data obtained are movie name, director, actors, rating, number of viewers, etc. These are the more important information for movie viewers. Finally, it is saved by Comma-separated values (CSV). In this investigation, the authors first looked at the HTML code of Douban through the URL of Douban, then found the general content in the HTML code, and then saved the required data through RE parsing (seen from Fig. 1). The final saved data is completely needed and correct. Those Case study shows that Web crawler technology allows data investigators to have better and more accurate access to data, making it easier for them to understand the data.

In many cases web crawler technology also helps data investigators to reduce a lot of time and effort. Previously, investigators need to keep looking on the web maybe to find information, or need a lot of time to save the data needed, while the web crawler technology on Python only needs a few lines of code to crawl the data needed to save in the storage space. In other words, twice the work with half the effort. At the same time, using traditional survey methods for a company may require a lot of manual labor to increase the speed, while web crawler technology can replace the work of most people to carry out uninterrupted surveys, and efficiency increases. For example, the scholars collected price lists of vegetables through web crawling techniques, which include specifications, weekday prices, trends, and charts. The time from start to finish took less than 2 hours. This definitely shows the significant impact of web crawling technology on time, effort and labor control.

4. The risk of Web Scraping Based on Python

On the other hand, since Web scraping is to get the data by visiting the website or HTML for a short period of time in large quantities. Therefore, there are massive number of disadvantage for Web scraping. Especially Web crawling based on Python. It leaves a very large number of sites or HTML that cannot handle such a large number of visitors and eventually the site or HTML may be in danger of crashing. Since there are more and more websites and HTML on the web, and there are very many data investigators who have decided to use web crawler technology to help them accomplish this purpose in order to search for the information they need faster. As Python is the most human-like programming language and the easiest language to use for web crawling, many people are using Python for web crawling. But a small number of investigators will be illegal, copyright-free theft of data on the network to become their own data, or theft of other people's trade secrets for their own benefit. This has caused very many web developers to lose profits and be violated personal privacy [10]. Most of the cases can be called unethical use of technology, but there are a few cases that are called illegal use of technology, which is why web developers and individual governments are cracking down on web crawling technology.

In addition, some malicious users will attack websites through the characteristics of web crawlers visiting a large number of websites and websites that cannot withstand a large number of visits in a short period of time. Such attacks are generally organized and disciplined, and carry a strong sense of purpose. Specifically, going to a commercial company to damage another commercial company's website or stealing commercial information. This makes the damage to these victims very serious. Therefore, many web developers and companies have developed anti-crawler technologies to protect themselves from crawler attacks in order to protect their interests and prevent their websites from being attacked. In general, anti-crawler technology is to determine whether the same Internet Protocol (IP) has been visited a certain number of times in a short period of time. In this case, some restrictions will be imposed on this IP, and these restrictions vary according to different websites (as illustrated in Fig. 2) [11]. Web crawlers are not only a risk to the victim, but may also be a potential risk to the user. Because a large number of websites are visited on the Internet, they leave their IP on many websites, and if they are unfortunate enough to visit a phishing website, there is a high probability that the computer will have a virus or be hacked.

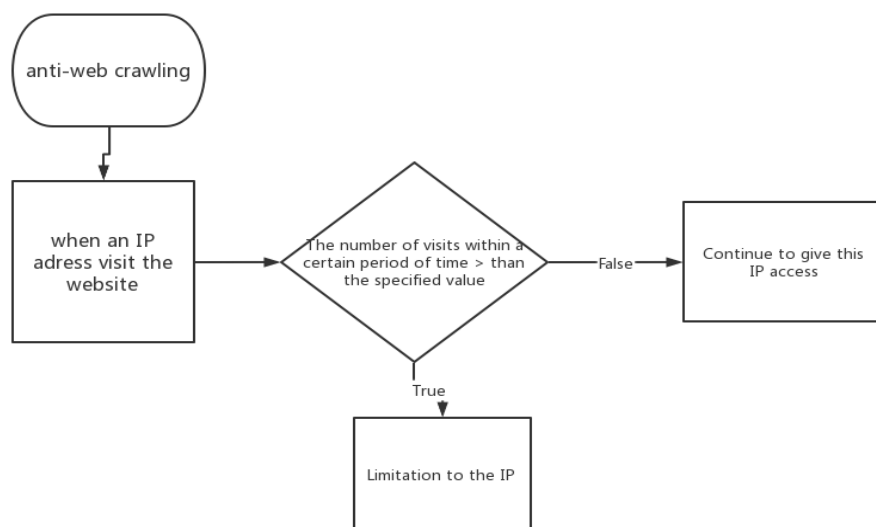


Fig 2. Process of how anti-web crawling work.

5. Limitations & Prospect

Web crawler technology currently has few limitations, one of the main ones being that most web crawler technology cannot parse the encrypted data or requires a lot of technical means to do so. In

most cases, the data on a website is not encrypted, but some entrepreneurs might choose to encrypt their commercial data in order to prevent their commercial works from being applied illegally. It results in data collectors being unable to use the data obtained even if they can download the web pages through web crawlers but suffer from the inability to decrypt or technical limitations.

Web crawling is likely to see more people using it in the future because the number of web pages on the web is increasing and the demand for web data is growing, so web crawlers will be a very useful technology in order to save time and find accurate information. There is a suggestion for Web scraping, which can try to reduce the difficulty of the decryption code, as the code difficulty affects the motivation of many data investigators for web crawling techniques and perhaps reduces the usage of web crawlers.

6. Conclusion

In summary, Web crawling technology is a very controversial technology, there are many people who love it, but there are also many who resist it. It has very many benefits, allowing web data collectors to gather the data they need accurately and quickly and analyze it, which greatly reduces the time and labor involved in searching for it. However, web crawling technology is used to access web pages by a large number of times, which makes some web pages instantly more accessible and causes some servers to crash, making the web pages temporarily unavailable, and this is just one of the risks of web crawling. But the existence of technology is reasonable, as long as the legal and compliant use of web crawler technology can bring a lot of convenience while not adding trouble to others. Web crawler technology is good, but in many cases it can't help, for example, when faced with data that requires complex parsing, which generally requires the design of complex and difficult-to-learn algorithms. As time goes on, society needs to investigate more and more data, which may lead to an increasing demand for web crawlers. The significance of this article is to make more people properly aware of, and properly use, web crawler technology.

References

- [1] Seymour T, Frantsvog D, Kumar S. History of search engines. *International Journal of Management & Information Systems (IJMIS)*, 2011, 15(4): 47-58.
- [2] Glez-Peña D, Lourenço A, López-Fernández H, Reboiro-Jato M, Fdez-Riverola F. Web scraping technologies in an API world. *Briefings in bioinformatics*, 2014, 15(5): 788-797.
- [3] Ho H P T. Leveraging web scraping for collecting competitive market data: Case: A case study of an Airbnb rental unit in Helsinki, 2020.
- [4] Python W. Python. Python Releases Wind, 2021, 24.
- [5] The Upwork team. (2022, September 20). Upwork. Web Scraping 101: Basics and Examples. Retrieved February 16, 2023, from <https://www.upwork.com/resources/web-scraping-basics#usage>.
- [6] Dhenakaran S S, Sambanthan K T. Web crawler-an overview. *International Journal of Computer Science and Communication*, 2011, 2(1): 265-267.
- [7] Andersson P. Developing a Python based web scraper: A study on the development of a web scraper for TimeEdit 2021.
- [8] Udupure T V, Kale R D, Dharmik R C. Study of web crawler and its different types. *IOSR Journal of Computer Engineering*, 2014, 16(1): 01-05.
- [9] Van Rossum G. Python Programming Language. In *USENIX annual technical conference*, 2007, 41(1): 1-36.
- [10] Krotov V, Johnson L, Silva L. Tutorial: Legality and ethics of web scraping, 2020.
- [11] Zhang L, Chen Z, Yang S. Research on the application of crawler technology in machine learning. In *Journal of Physics: Conference Series*, 2021, 1865(4), 042040.