An Undergraduate Internship on Data Scraping

Mahmudul Islam Akhund

Student ID: 1620645

Autumn, 2021

Supervisor:

Mr. Bijoy Rahman Arif

Lecturer

Department of Computer Science & Engineering

Independent University, Bangladesh

September 12, 2021

Dissertation submitted in partial fulfillment for the degree of Bachelor of Science in Computer Science

Department of Computer Science & Engineering

Independent University, Bangladesh

Attestation

This is to certify that the report titled “Website Application” is completed by me, Mahmudul Islam Akhund (1620645), submitted in partial fulfillment of the requirement for the Degree of Computer Science and Engineering from Independent University, Bangladesh (IUB). It has been completed under the guidance of Mr. Bijoy Rahman Arif (Internal Supervisor). I also certify that all my work is original which I have learned during my internship. All the sources of information used in this project and report have been duly acknowledged in it.

Signature Date

Write Your Name Here

Name: Mahmudul Islam Akhund

Acknowledgement

Firstly, I would like to thank the Almighty for His blessings for keeping me healthy, safe and for making me able to complete my internship especially in this situation. Secondly, I am really grateful to my Advisor, Mr. Bijoy R. Arif, Lecturer, Department of Computer Science and Engineering, for his guidance, support and understanding throughout this internship period.

Then I would like to express my gratitude to Shikhao, for giving me the opportunity to complete my internship under their guidance and care and giving me this opportunity to take part in this internship program. The learning and experiences I have gathered here have groomed me a lot and this will surely help me in the next phase of life.

Finally, I would like to thank my parents for their immense love, support and letting me complete my internship during this pandemic. Without their blessings I could not have come this far.

Letter of Transmittal

August 15, 2021

Mr. Bijoy Rahman Arif

Lecturer

Department of Computer Science and Engineering

Independent University Bangladesh

Subject: Internship Report submission Autumn, 2021.

Dear Sir,

It is a great pleasure and honor to submit my Internship report on Data Scraping under your guidance. I have tried to present my project work, my experiences and my achievements in this report.

I have continued my Internship with Shikhao as an intern from March 15, 2021 to date. During this whole time period, I have gathered real-life working experience and knowledge in various aspects. This report includes all the project works, experiences and learning that I have achieved during this internship.

I would like to thank you for your patience, instructions and kindness. I have tried to complete this with the utmost honesty and sincerity. I hope and pray that this report fulfills all the requirements and is up to your expectations.

Sincerely,

Mahmudul Islam Akhund

Evaluation Committee

. . . . . .
. . . . . Signature

. . . . . .
. . . . . Name

. . . . . .
. . . . . Supervisor

. . . . . .

. . . . . Signature

. . . . . .
. . . . . Name

. . . . . .
. . . . . Internal Examiner

. . . . . .
Signature

. . . . . .
. . . . . Name

. . . . . .
. . . . . External Examiner
. . . . . .
Signature

. . . . . .
. . . . . Name

. . . . . .
Convener

Abstract

In this report, I have described the knowledge and experiences I have gathered and the work I have done throughout my internship at Shikhao as an intern. I have worked on Data Scraping where mostly my task was to scrape data from target public websites.

This work was divided into three parts; Scraping the Data, Storing the Scraped Data and Sorting the Data according to need. This report contains the update on as much of my work as I can share as I had to sign a Non-disclosure Agreement when joining the edTech company. All the shareable information is mentioned in this report. Data Scraping mainly uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured websites under the interdisciplinary field we know as Data Science. The requirement to apply data scraping is knowledge regarding standard frameworks. I was introduced to the python frameworks scrapy, beautiful soup and selenium through my internship experience.

Mentionable completed tasks include scraping the data from the websites of UGC and UGC approved Universities, Online tuition and student forums, sorting the data according to the directives of the supervisor and lastly mentioning a final demo work assigned by my inner supervisor Mr. Bijoy R. Arif where key data was scraped and stored in a way that it imitates a search engine but not for web pages, rather for the key data points or keywords.

Keywords:

Web Scraping, World Wide Web, Data Extraction, API

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview/Background of the Work

In the 21st century, everything is dependent on information. People depend on data and information for everything. In the corporate sector, nowadays most employers use PCs or laptops for work purposes. They rely on the PC because it is easier to keep records of data there. Data extraction from documents kept in papers is much harder and time-consuming compared to Data extraction using a PC and data has a high chance of getting lost easily if recorded in Documents whereas there are fewer risks of data loss if stored on a PC. Data is easier to trace back from a PC but data tracing from documents is much more strenuous. The world revolves around data and information at present. The people who can control the flow of information and store it are ahead in life. As Shikhao is a new Ed Tech Company, the fastest way to get a good base would be to collect data and narrow down the steps necessary for them to follow so they can make profits. For such reasons, the opportunity to work for Shikhao and scrape data for their goals was available. Here employers have evaluated the work results from data scraping and assigned new work depending on their standards. It was my job to finish the assigned work and get new work.

1.2 Objectives

1. To scrape data from online and store it for future use
2. To keep records of all the data that is stored for future mining purposes
3. To understand what quality of data the company is looking for to get better yield of data

1.3 Scopes

Here are the features available in this web application for users, admin and engineers in ticket portal system:

1. Using past data from public websites, optimize the marketing strategy to reach target users.

2. Data available includes: currently successful edTech companies, online websites for tutors and students, university websites, ugc website(for standards) and many more.

3. If time permits and the necessity rises, an even more in depth solution to achieve the goals can be found after using the scraped data accordingly.

Chapter 2

Literature Review

2.1 Relationship with Undergraduate Studies

1. CSE100- Luckily, I was a student of IUB at a time when CSE100 was a mandatory course and was taught in the Python language instead of C++. The course handled the basics of coding but it got me comfortable with the language by the end of the semester.

2. CSE213- The course was conducted in Java language but it helped gain a broader understanding of code structure and standards under our very own Subrata Kumar Sir. The course especially helped because it taught us error handling which was most useful throughout my work as my knowledge on python and it's frameworks were fairly new to me and I made many errors that needed me to understand what kind of error I am facing and find how to handle the errors.

3. Other than the basic structures of coding and familiarity to the language, no other course in IUB was helpful in understanding my work and most of my learning regarding the frameworks I've used came from the Shikhao supervisor and online material. It was rough and rocky but the basics helped get the job done.

2.2 Related works

All the big tech companies starting from Google to Facebook have depended on data science to improve their marketing strategies and finding target customers. The billion dollar company owned by Jeff Bezos also uses data science to make many of their decisions.[2] The company was able to profit from this by relying on the conclusions that their analysts made for them, from the data that data scientists were able to scrape and mine.

Other than that in Bangladesh there is even a forum to understand Big Data and help to create data products with competitive intelligence and business analytics [3], which confirms the prospects of the work in the future in our country and not just relying on foreign countries to give opportunities for more work.

Chapter 3

Market Analysis and Research

Data collection from any type of online source became one of the most efficient market research methods in the 21st century. It not only offers a faster response, compared to classical surveying but also helps get a lot of data in a shorter time. While some may consider it best to utilize the traditional surveys, it is undeniable that Web-scraping is seen as a cost effective and efficient support for such instruments. To get a comprehensive picture and to gain knowledge of the tools in markets multiple sources should be used.(Raulamo-Jurvanen P., 2016)

Consumers are available in the online world and that's where they share their experiences, frustrations and motivations.Companies that wish to learn more from their customers or consumers in general, can always add online sources of information.Web scraping is just another one of the methods to collect such data.Targeted data collection from e-shop and advertising servers helps to update the knowledge regarding current market situation, based on frequently changing prices. More frequent update intervals are ensured through automated Web scraping and its implications.

With the increasing relevance and availability of on-line prices that we see today, it is natural to ask whether the prediction of the consumer price index (CPI), or related statistics, may usefully be computed more frequently than existing monthly schedules allow for.[9]

A small sample of 338 Craigslist listings to study the prevalence of secondary dwelling units in the San Francisco Bay Area was conducted by Wegmann and Chapple (2013). Finally, Feng (2014)web-scraped 6,000 craigslist listings to study Seattle’s housing market.(Daniel Glez-Pen,nedatováno) All these are the implications of data science and data scraping in the real world.

Chapter 4

Methods of Web Scraping

The different methods of Web Scraping have evolved together with the internet. Not all listed methods were available at the beginning. There are two mention-worthy techniques; that are most used presently. The Document Object Model (DOM) has become more popular in DHTML than it was in 2000 after which a broader acceptance, later on, allowed the DOM Parsing technique to evolve from HTML Parsing technique. The second technique is the application of Application Programming Interfaces (APIs). This technique is the youngest on
the list, the growth of available content APIs is dated from 2005. According to ProgrammableWeb.com the number of APIs has grown within 8 years from 0 to 10302.(Berlind, 2015)

4.1 Manual Scraping

Manual scraping is a viable option in specific situations.

- For a minimal amount of data,
- For no repetitive data scraping tasks
- For automated scraping with time-consuming data collection
- Possibly security measures or specific characteristics of the website do not allow automated methods.

4.2 HTML Parsing

Websites don’t usually provide their data in traceable formats such as .csv or.json files. As a response to a user’s request, HTML Pages are created by the server. Analysis of the HTML structure in the provided web page will often show repeated elements. Each page has a similar pattern that can be used as a source of data with any programming language script or Web Scraping tool.

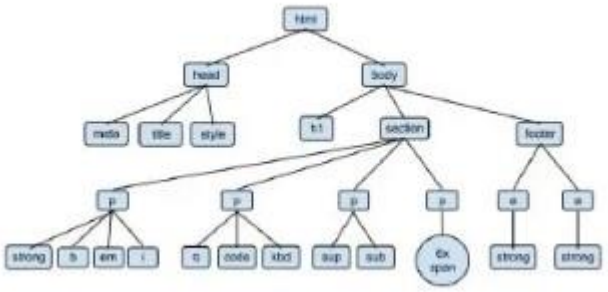

Figure 1 HTML Structure

4.3 DOM Parsing

A noteworthy evolution of HTML Parsing based on developments of the language and browsers is the Document Object Model (DOM) Parsing, which led to the start of the Document Object Model. DOM is heavily used for css and JavaScript. Integration of DOM revealed new possibilities for addressing some specific parts of the webpage, which made the data scraping process much easier than it used to be. These are used in Web Scraping for easier navigation through web page content.

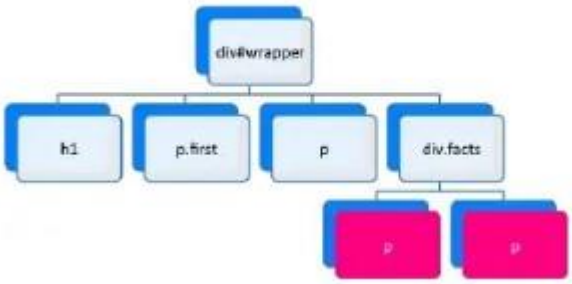

Figure 2 DOM Selector

4.4 XPath

Another similar addressing possibility with similar use as DOM provides XPath (XML Path Language). It is related to XML documents and is also applicable to HTML format. XPath requires a more precisely structured webpage than DOM and has the same possibility to address segments within the webpage and can be easily accessed using the available browser extensions. Figure 3 shows the document structure as interpreted in XPath.

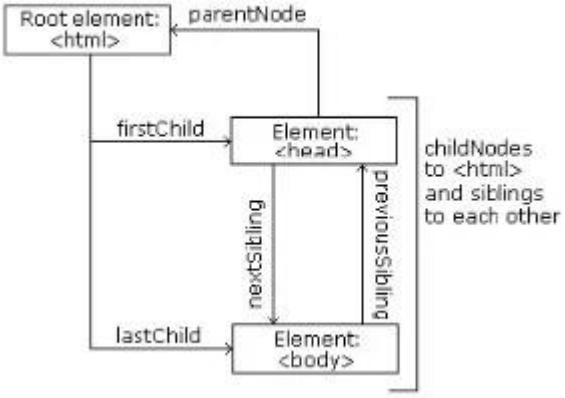

Figure 3 XPath Navigation

4.5 APIs

While the past strategies work to rub human-readable yields, Application Programming Interface(API) anticipates an application as a communication accomplice. Hence APIs are regularly named as machine-readable interfacing. Even APIs have presented much afterward than the WWW, and their development was unexpectedly exceptional and quick. The World of APIs can be found in fragments.

Most of the accessible APIs are enlisted and portrayed within the catalog with pertinent joins to the sources. Reasons for this style of cataloging are - Programmable Web (https://www.programmableweb.com) and APIs(https://apis.guru/). API Registries too give their own claimed API, which permits clients to look through their database for Sources of API. A standard HTTP request sent to an API endpoint returns the reply from the server. Each API has its determination and alternatives, exceptional to the API. The arrangement of the reply can be set as a choice within the request. The most widely used format for API communication is JSON.

Chapter 5

Web Scraping

5.1 Task Detail

The rapid growth of the World Wide Web has significantly changed the way we share, collect, and publish data. The vast amount of information is being stored online, both in structured and unstructured forms. Regarding certain questions or research topics, this has resulted in a new problem no longer is the concern of data scarcity and inaccessibility but, rather, one of overcoming the tangled masses of online data.(B.C., 2016)

Web Scraping involves the process of querying a source, retrieving the results page and parsing the page to obtain the results.(John J. Salerno, 2003)

The task has three parts; Scraping the Data, Storing the Scraped Data and Sorting the Data according to need. Data Scraping mainly uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured websites under the interdisciplinary field we know as Data Science. The requirement to apply data scraping is knowledge regarding standard frameworks. I was introduced to the python frameworks scrapy, beautiful soup and selenium through my internship experience which I used for most of my work. Even though I heavily focused on scrapy, use of beautiful soup became unavoidable for some of the issues I had to face. Mentionable completed tasks include scraping the data from the websites of UGC and UGC-approved Universities, Online tuition and student forums, sorting the data according to the directives of the supervisor.

5.2 Coding Prerequisites

- Access to selected websites
- Installed browser extensions namely: Selector Gadget and XPath Helper
- Sufficient bandwidth of an internet connection
- Sufficient Central Processing Unit(CPU) and RAM size –this use case is not intensive on computing power.

5.3 Software Selected

The software I used to complete my tasks includes Sublime Text, MiniConda, Web Browser, Python 3.8, default command prompt, and the extensions used in my web browser. This was done for reasons:

- Fast learning curve– the user interaction is simple to adapt to

- Simple setup– use of the software is easy as the commands are basic

- Available documentation– data on the use of the applications are available online.

5.4 Preparation for Coding

The Selector Gadget and Xpath Finder help to collect data from HTML of websites based on its templates. As a first step, I had to use the cmd or miniconda to install important packages and then either use Jupyter Notebook or Sublime Text to run the codes. While coding, first, I had to define the URL and start to define the parsing functions. At this moment it is necessary to provide a structure. Common usage requires defining the main object and binding the related ones.

When using the selector gadget or Xpath finder on a website it gives me the specifics of the page I want to scrape. With the data, I can specify how the function should scrape from the website and where it should scrape from. Other than that, there is also the option to scrape the entire webpage and work with that.

5.5 Execution of Code

Once I start coding the data from the Selector Gadget and Xpath can guide the code to scrape from the website in the URL using the specific data areas, which I want to collect. Additionally, related content like images, metadata was set to be collected in the same process. Each separate page of the URL can be scraped through accessing the link or if they use different selectors and paths then different codes have to be written to parse the data. Due to this, the execution of the task is dependent on the type of webpage being scraped.

Chapter 6

Results & Analysis

Most of the results were presented in the CSV format, a few in HTML format but for the early days of learning. Each of the tasks is only provided if and only if it meets the requirements successfully. Initially, some of these tasks showed bugs/errors but they were fixed after several tests and scraping. Some of these tasks have some shortcomings which have not been implemented yet have been mentioned above. The tasks that do not show any bugs or errors, so the error rate has been stated as zero. Apart from me, my supervisor also tested the code and checked the files that dropped from that code.

Table 1: Result Table

Scraping Task	Description	Conditions	Success Rate (%)	Error Rate (%)	Shortcomings	Store Scraped Data
BD Tutors Website	Data on Students, Data on Teacher that have been made public	Public website so seeking permission was not mandatory	100	0	<Nil>	Yes
Teacher Website	Data on Students, Data on teachers that have been made public	Public website so seeking permission was not mandatory	50	50	Most of the target data was unavailable or could not be scraped due to redirect error	No
Dikkha Website	Data on Students, Data on Teacher that have been made public	Public website so seeking permission was not mandatory	50	50	Most of the target data was unavailable or could not be scraped due to redirect error	No
Desh Tutor Website	Data on Students, Data on Teacher that have been made public	Public website so seeking permission was not mandatory	100	0	<Nil>	Yes

Scraping Task	Description	Conditions	Success Rate (%)	Error Rate (%)	Shortcomings	Store Scraped Data
UGC	Data on	Public	100	0	<Nil>	Yes
Website	Universities, Allowed Courses, Rules and Regulations	website so seeking permission was not mandatory				
Target Private Universities	Data on Academics,Departments, Faculty Members and Student data	Public website so seeking permission was not mandatory	90	.10	Some of the links could not be scraped because the data privacy	Yes
<NDA>	<Nil>	Due to signing an NDA, I can only share parts of my work that I have the permission to.	<Nil>	<Nil>	<Nil>	<Nil>

Chapter 7

Work as an Engineering Problem Analysis

7.1 Sustainability of the Work

Web scraping definitely will not go out of trend within the next 3 decades unless an easier form of data collection and organization is established. The prospects of the scraping and the existing statistics of profits that came from market analysis based on data that was web scraped speak for itself, louder than any other fact does.

Web scraping, data scraping; recently caught the eyes of prominent businesses in our country. The future of the field is just expanding and with new data consistently being added to the web, it has become almost imperative to find good ways to organize the necessary data from the noise, which is something web scraping is used for.

My work in the company similarly will only get iterations, no drastic changes will happen. Because the need of the modern world will remain information and data, for a significant time, thus making web scraping a sustainable and prominent aspect of data science.

7.2 Social Effects and Analysis

Web scraping has social effects, especially if the web scraping of social media platforms is considered. Given that modern people are concerned about their data privacy and many web scraping spiders can have malicious intentions, it is safe to assume that web scraping can affect society negatively. On the other hand, there are less intrusive and informative ways web scraping has helped people of the society as well.

As the social media platform owners have become more aware of the misuse of access of data to 3rd parties, this is a prime time to use web scraping, as many of the negative effects that could have dawned upon the society have been handled either by the owners restricting mass access to third party owners for all data online or legislatures have been passed to stop the exploitation of web scraping to harm others or profit from others without due compensation.

7.3 Addressing Ethics and Ethical Issues

To cut to the chase, a specific piece of legislation or law that forbids Web Scraping to gather information is yet to be made. But the owners of websites who know about scraping usually have legal rights against the company under intellectual property law and contract law. Each case will at the end of the day be decided based on the facts and this is very much dependent upon what kind of information was scraped from the websites. Companies are usually advised to be aware of the contractual provisions which they have agreed to in respect to the terms of use of a website – these sometimes relinquish the right of the user from taking and using the data of the websites. The only way to be truly certain that the rights of a website owner have not been infringed is to obtain their express consent to the screen scraping and subsequent use of the information. (Rezai, 2017)

The overall outcome of forbidden Web Scraping is usually considered negative for the owner of the website because of the possibility of losing visitors, content aggregator websites with fewer links and advertising yielding less income. For this reason, data hosts choose to only use legal actions against scrapers when they come off as a threat to the core business and the data host has a strong enough claim to prevail legally against them. From a law perspective, it is necessary to adjust the terms of use on many websites. Restrictions on Scraping techniques can be directly included in the terms. Such a step does not require the use of many resources and it also allows a direct argument to take to the court. Which summarizes the ethical issues regarding scraping.

Chapter 8

Lesson Learned

8.1 Problems Faced During this Period

2020-2021 the time period when the world has been brought down to its knees against the might of the viral Corona Virus was when everyone was expected to find an internship job and expected to meet work deadlines without any consideration for circumstances. The hardest part of the internship was finding a job. But the inner supervisor was not very lenient through my experience. That is something that will not be forgotten till the death bed because, in a deadly situation, I learned that apparently compassion is not obligatory, rather a guideline.

Regardless, the scraping was not much of a problem once I adapted to the necessary knowledge. The problems faced include: the error 302 which redirected the scraping from the target website to a different one, which made it hard to scrape any data from any public website, also mention-worthy are the issues with some University webpages with really bad HTML coding, with the codes varying from page to page and it's not even different departments, only NSU uses perfect organized coding is the conclusion I came to after scraping data from multiple Private Universities, other than that there were issues with the EWU data, as the extraction would often return empty arrays, which made me realize some form of restriction was used stopping from the extraction of some key data from their websites.

8.2 Solution of those Problems

The problem of different coding used in different Webpages was not that hard to get over, it just took more time to make more individual spiders for each weird website and the data was then added to the files where scraped data were being stored.

The data that could not be accessed due to the redirect error, were handled more manually and only the key parts of it were taken to meet the goals but it's less related to the scraping. A similar format of code was used but it wasn't directly scraped from the webpage.

The data that was subject to the use of API, approvals, and permissions of websites were directly handled by the company, regarding those works, I had more of a grunt work position, helping by being part of the division of labor, instead of handling it myself. The company needs to assign me a more permanent position to allow handling those things

Chapter 9

Future Work & Conclusion

9.1 Future Works

In the future, the plan is to either help the edTech startup Shikhao grow into a much more prominent company so they can influence the people of the society. It starts with data collection, gathering information and mining according to goals.

The option to complete an online masters degree in Data science and web scraping is also an option if I consider getting an in-depth understanding of the subject matter. This can help get jobs in established data science firms or companies working to strengthen their web scraping arsenal.

Other than that, if a more immediate concern is addressed, the goal is to keep providing the company with my services until the goals have been met, reiterated or a better opportunity for learning and practice comes up.

9.2 Conclusion

Depending on the main goals and purpose, different Web Scraping techniques can be used, the amount of data scraped its periodicity and considering the required outcome into consideration. There is a huge collection of Web scrapers, selection of tools most efficient to the need is helpful.

All Web Scraping projects should be able to address the legal aspect of the specific job examined and have their necessary steps identified. Data hosts can choose to assess the benefit that scrapers can provide and take a pragmatic approach to grow or establish data scientists who scrape their data. Web Scrapers can choose to keep the connection to the
Data hosts and as a source of presented information, allow identification of the Data Host.

Bibliography

1. [Wien,04.02.2018]https://www.academia.edu/35901535/BACHELOR_PAPER_Web_S craping_Data_Extraction_from_websites?fbclid=IwAR0BHqZZkDJlVG2C4QMO4njq2o3-8KbWDtIXRCJt9zbLpuEQq1r7qmHfM6w
2. https://medium.com/dataflair/how-data-science-has-taken-amazon-on-top-360043e146ec
3. http://datasciencebd.org/
4. B.C., B., 2016. Scraping Data. In: Data Wrangling with R. Use R!.. Cham: Springer 5. John J. Salerno, D. M. B., 2003.Method and apparatus for improved web scraping. United States of America, Patentnr. US 7072890 B2
6. Berlind, D., 2015.APIs Are Like User Interfaces--Just With Different Users in Mind.
7. [Online] Available at:: https://www.programmableweb.com/news/api-economy-delivers-limitless-possibilities/ analysis/2015/12/03 [Zugriff am 17 November 2017]
8. Rezai, A., 2017.Beware of the Spiders: Web Crawling and Screen Scraping– theLegal Position.[Online]Available at: https://parissmith.co.uk/blog/beware-spiders-web-crawling-screen-scraping-legal-posit ion/ [Zugriff am 28 November 2017]
9. Jie Yang & Brian Yecies, 2016. Mining Chinese social media UGC: a big-dataframework for analyzing Douban movie reviews.Journal of Big Data
10. Raulamo-Jurvanen P., K. K. M. M., 2016. Using Surveys and Web-Scraping to SelectTools for Software Testing Consultancy. In:Lecture Notes in Computer Science, vol10027.Cham: Springer
11. Daniel Glez-Pen, M. R.-J. a. F. F.-R., kein Datum Webscrapingtechnologies in an APIworld.s.l., s.n.