# WEB SCRAPING USING PYTHON

**Mini Project Report**

**Submitted to**

**Sant Gadge Baba Amravati University**

In partial fulfillment of the Requirement for the Mini Project (III Year) of Degree of

**Bachelor of Engineering**

By

Mr. Yash Rajesh Balkhande Mr. Shreyash Narendra Waghmare

Mr. Mohit Vasant Gedam Miss. Aboli Chandrashekhar Deshmukh

Miss. Aachal Vilas Ganvir

**Under the guidance of**

**Dr. A. D. Raut**



**Department of Computer Science and Engineering**

**P. R. Pote Patil Education & Welfare Trust's Group of Institutions,**

**College of Engineering & Management.**

**Amravati – 444602 (M. S.) 20212022.**

**Department of Computer Science and Engineering**

**P. R. Pote Patil Education & Welfare Trust's Group of Institutions,**

**College of Engineering & Management.**

**Amravati – 444602 (M. S.) 2022-2023.**



# CERTIFICATE

This is to certify that the Mini Project report entitled

## Web Scraping using python

Submitted by

Mr. Yash Rajesh Balkhande Mr. Shreyash Narendra Waghmare

Mr. Mohit Vasant Gedam Miss. Aboli Chandrashekhar Deshmukh

Miss. Aachal Vilas Ganvir

In partial fulfillment of the requirements for the Mini Project (III Year) of Degree of Bachelor of Engineering in Computer Science and Engineering by Sant Gadge Baba Amravati University & is a bonafied work carried out during the session 2022-2023.

|  |  |
| --- | --- |
|  |  |
| Dr. V. B. Kute | Dr. A. D. Raut |
| Head of the Department | Guide |

**DECLARATION**

We hereby declare that we have completed the mini project work (III Year) towards the

Bachelor of Engineering Degree of Sant Gadge Baba Amravati University, Amravati, in Computer Science and Engineering discipline on the topic entitled Distance Measuring Wheel under supervision of **Dr. A. D. Raut** of Department of Computer Science and Engineering, P. R. Pote Patil Education & Welfare Trust's Group of Institutions College of Engg. & Management, Amravati.

This report embodies the original work done by us in fulfilment of the requirement of the Bachelor of Engineering Degree of Sant Gadge Baba Amravati University, Amravati, in Computer Science and Engineering discipline.

Amravati

Date:

Name of Students:

1.Mr. Yash Rajesh Balkhande

2.Mr. Shreyash Narendra Waghmare

3.Mr. Mohit Vasant Gedam

4.Miss. Aboli Chandrashekhar Deshmukh

5. Miss. Aachal Vilas Ganvir

## ACKNOWLEDGEMENT

It is our supreme duty and desire to express acknowledgement to the various torchbearers, who have rendered valuable guidance during the preparation of our Mini

Project.

First of all, we extend our deepest gratitude to our revered Principal **Dr. D.T. Ingole** without whose support, our Mini Project could not have been transformed into present form.

We are grateful to **Dr. V. B. Kute** HOD, Computer Science and Engineering Department; and guide **Dr. A. D. Raut** for providing immense support and guidance. We are beholden for guiding us at every step in the Mini Project. He has most honestly guided me throughout; never leaving me unanswered for any of our doubts. It was his constant persuasion, encouragement, inspiration and able guidance that helped us in completing our Mini Project successfully.

Name of Students:

1.Mr. Yash Rajesh Balkhande

2.Mr. Shreyash Narendra Waghmare

3.Mr. Mohit Vasant Gedam

4.Miss. Aboli Chandrashekhar Deshmukh

5.Miss. Aachal Vilas Ganvir

Third Year, **CSE**

### TABLE OF CONTENT

|  |  |  |  |
| --- | --- | --- | --- |
| **Chapter No.** |  | **Contents** | |
|  |  | Certificate | |
|  |  | Declaration | |
|  |  | Acknowledgement | |
|  |  |  | |
|  | **1** | **Introduction** | |
|  |  | 1.1 Introduction | |
|  |  | 1.2 Problem Definition | |
|  |  | 1.3 Objective | |
|  |  | 1.4 Organization of Report | |
|  | **2** | **Literature Survey (If any)** | |
|  | **3** | **Project Planning Implementation** | |
|  |  | 3.1 Proposed Methodology  3.2 Working Proposed System | |
|  | **4** | **System Implementation and Testing**  4.1 Setting Environment  4.2 System Execution Details | |
|  | **5 Advantages & Disadvantages**  5.1 Advantages  5.2 Disadvantages | |
|  | **6 Conclusion & Future Scope**  6.1 Conclusion  6.2 Future Scope | |

**7** **Reference**

**Appendix**

### Abstract

Main objective of Web Scraping is to extract information from one or many websites and process it into simple structures such as spreadsheets, database or CSV file. However, in addition to be a very complicated task, Web Scraping is resource and time consuming, mainly when it is carried out manually. Previous studies have developed several automated solutions. The purpose of this article is to revisit the different existing Web Scraping approaches, categories, and tools, but also its areas of application.

Name of Students:

1.Mr. Yash Rajesh Balkhande

2.Mr. Shreyash Narendra Waghmare

3.Mr. Mohit Vasant Gedam

4.Miss. Aboli Chandrashekhar Deshmukh

5. Miss. Aachal Vilas Ganvir

**Chapter 1**

**INTRODUCTION**

**1.1 INTRODUCTION**

Web scraping is a technique to fetch data from websites. While surfing on the web, many websites don’t allow the user to save data for personal use. One way is to manually copy-paste the data, which both tedious and time-consuming. Web Scraping is the automation of the data extraction process from websites. This event is done with the help of web scraping software known as web scrapers. They automatically load and extract data from the websites based on user requirements. These can be custom built to work for one site or can be configured to work with any website. Uses of Web Scraping: Web scraping finds many uses both at a professional and personal level.

**1.2 PROBLEM STATEMENT**

Web scraping is great when you have it all working, and can save you a ton of mind-numbing copy and pasting. Unfortunately, getting to that sweet spot can be filled with frustration. At first, things look easy - but as you start scraping more data, the problems begin to mount. Solving these is part science, part art. At axiom, we’re scraping experts that have seen 1000s of cases. These are the most common problems you’ll encounter and our recommended fixes.

Computers are brilliant in some ways, but quite dumb in others, and they don’t understand pages as well as a human does. Rather than parsing the data as meaningful chunks, they see the page as a generic data structure called the “Document Object Model” or DOM. Just because content looks the same on a page, doesn’t mean it has the same underlying DOM structure, and this can cause the web scraper to behave inconsistently.

Overall, web scrapingcan be a powerful tool for gathering data, but it requires careful planning and execution to overcome the challenges that come with it. By addressing these issues head-on, you can ensure that your web scraping efforts are successful and produce high-quality data.

#### 1.3 OBJECTIVES

Web Scraping is used for getting data. Access to relevant data, having methods to analyze it and performing intelligent actions based on analysis can make a huge difference in the success and growth of most businesses in the modern world. Data collection and analysis is important even for government, non-profit and educational institutions.

The following are few of the many uses of Web Scraping:

* Introduce the concept of structured data
* Discuss how data can be extracted from web pages
* Introduce the examples that will be used in this lesson
* In eCommerce, Web Scraping is used for competition price monitoring.
* In Marketing, Web Scraping is used for lead generation, to build phone and email lists for cold outreach.
* In Real Estate, Web Scraping is used to get property and agent/owner details.
* Web Scraping is used to collect training and testing data for Machine Learning projects.

**1.4 ORGANIZATION OF REPORT**

Distance Measuring Wheel system provides effective results of measurement of distance of different land. In which chapter 2 summerizes different types of traditional measuring tools. The system description and working are well illustrated in chapter 3. Actual implementation of distance measuring wheel system using hardware description is given in chapter 4. Chapter 5 provides experimental results

and application. Conclusions and future work of distance measuring wheel system are specified in chapter 6.

**Chapter 2**

### LITERATURE REVIEW

#### Big data analytics gives organizations a way to analyze huge data sets and gather new information. It helps answer basic questions about business operations and business performance. It also helps discover unknown patterns in vast datasets or combinations thereof. In the current data-driven world, it becomes increasingly essential that big data techniques are applied and analyzed for organizational growth. More specifically, with the large availability of data on the Web, whether from social media, websites, online portals, or platforms, to name but a few, it is important for organizations to know how to mine that data in order to extract useful knowledge. Web scraping represents a fundamental approach in this regard. Therefore, this paper aims to provide an updated literature review about the most advanced Web Scraping techniques to better equip scholars and managers with helpful knowledge on how to mine most effectively online data. The paper starts with presenting the basic design of a web scraper and the applications of web scraping in diverse sectors and areas. Next, the different Web scraping methods and Web scraping technologies are presented. Finally, a procedure to develop Web scraping with various tools is proposed before a conclusion wraps up the paper

**Chapter 3**

### Project Planning Implementation

#### 3.1 PROPOSED METHODOLOGY

#### Web Scraping is the process of automatically mining data or collecting information from the World Wide Web. There are methods that some websites use to prevent web scraping, such as detecting and disallowing bots from crawling (viewing) their pages. In response, there are web scraping systems that rely on using techniques such as DOM (Document Object Model), computer vision and natural language processing to simulate human browsing to enable gathering web page content for offline parsing. Current web scraping solutions range from the ad-hoc, requiring human effort, to fully automated systems that can convert entire websites into structured information, with limitations.

The library of codes we are going to use for this project are :

* **Requests Library**
* **Beautiful Soup Library**
* **Selenium Webdriver**
* **Beautiful Soup Library**
* **MongoDB**
* **EJS**

#### *Request Library :-* The [requests](https://requests.readthedocs.io/en/latest/) library is the de facto standard for making HTTP requests in Python. It abstracts the complexities of making requests behind a beautiful, simple API so that you can focus on interacting with services and consuming data in your application.

#### *Selenium :-* Selenium WebDriver is a web framework that permits you to execute cross-browser tests. This tool is used for automating web-based application testing to verify that it performs expectedly. Selenium WebDriver allows you to choose a programming language to create test scripts. As discussed earlier, it is an advancement over Selenium RC to overcome a few limitations. Selenium WebDriver is not capable of handling window components, but this drawback can be overcome by using tools like Sikuli, Auto IT, etc.

#### *Beautiful Soup Library :-* Beautiful Soup is a [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) package for parsing [HTML](https://en.wikipedia.org/wiki/HTML) and [XML](https://en.wikipedia.org/wiki/XML) documents (including having malformed markup, i.e. non-closed tags, so named after [tag soup](https://en.wikipedia.org/wiki/Tag_soup)). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for [web scraping](https://en.wikipedia.org/wiki/Web_scraping). Beautiful Soup was started by Leonard Richardson, who continues to contribute to the project, and is additionally supported by Tidelift, a paid subscription to open-source maintenance.

#### *MongoDB :-* MongoDB is an open-source document-oriented database that is designed to store a large scale of data and also allows you to work with that data very efficiently. It is categorized under the NoSQL (Not only SQL) database because the storage and retrieval of data in the MongoDB are not in the form of tables.

#### *EJS :-*[EJS](https://ejs.co/) (Embedded JavaScript Templating) is one of the most popular template engines for JavaScript. As the name suggests, it lets us embed JavaScript code in a template language that is then used to generate HTML.

**3.2 WORKING PROPOSED SYSTEM**

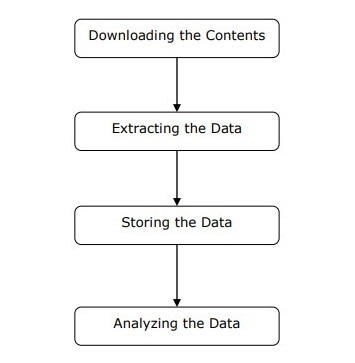
The data extracted above is not suitable for ready use. It must pass through some cleaning module so that we can use it. The methods like String manipulation or regular expression can be used for this purpose. Note that extraction and transformation can be performed in a single step also.

**Storage Module**

After extracting the data, we need to store it as per our requirement. The storage module will output the data in a standard format that can be stored in a database or JSON or CSV format.

**Working of a Web Scraper**

Web scraper may be defined as a software or script used to download the contents of multiple web pages and extracting data from it.



**Chapter 4**

### SYSTEM IMPLEMENTATION AND TESTING

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr No.** | **Name of the Resources** | **Specification** | **Qty.** | **Remarks** |
| 1. | Hardware : Computer System | Computer (I3-I5 Preferable) RAM minimum 4 GB and Onwards | 1 |  |
| 2. | Operating System | Windows 10-11 | 1 |  |
| 3. | Development Software | Python IDE, EJS, Visual Studio | 1 |  |

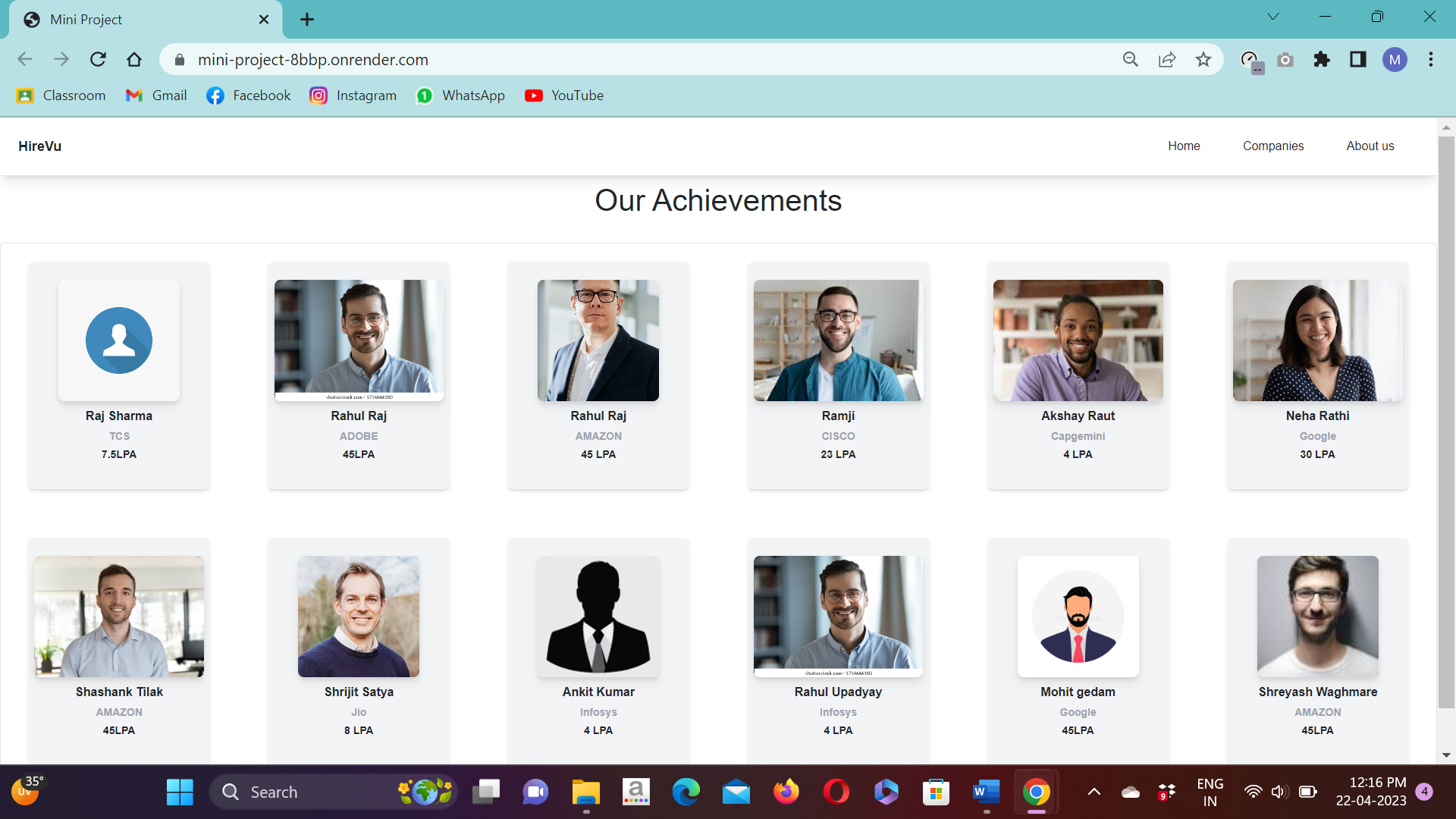
#### 4.1 SETTING ENVIRONMENT

#### 

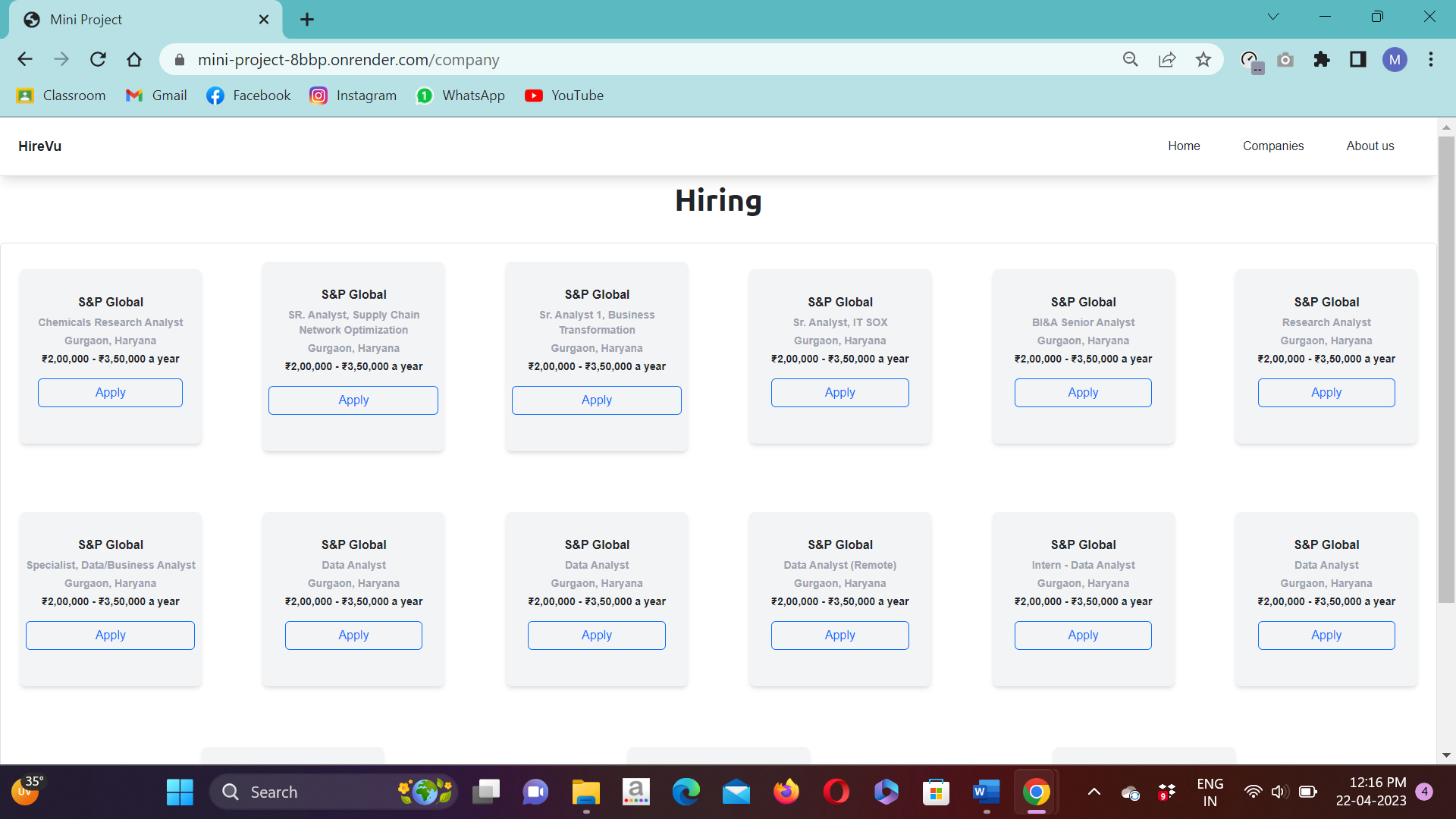
##### 

#### 4.2 SYSTEM EXECUTION DETAILS

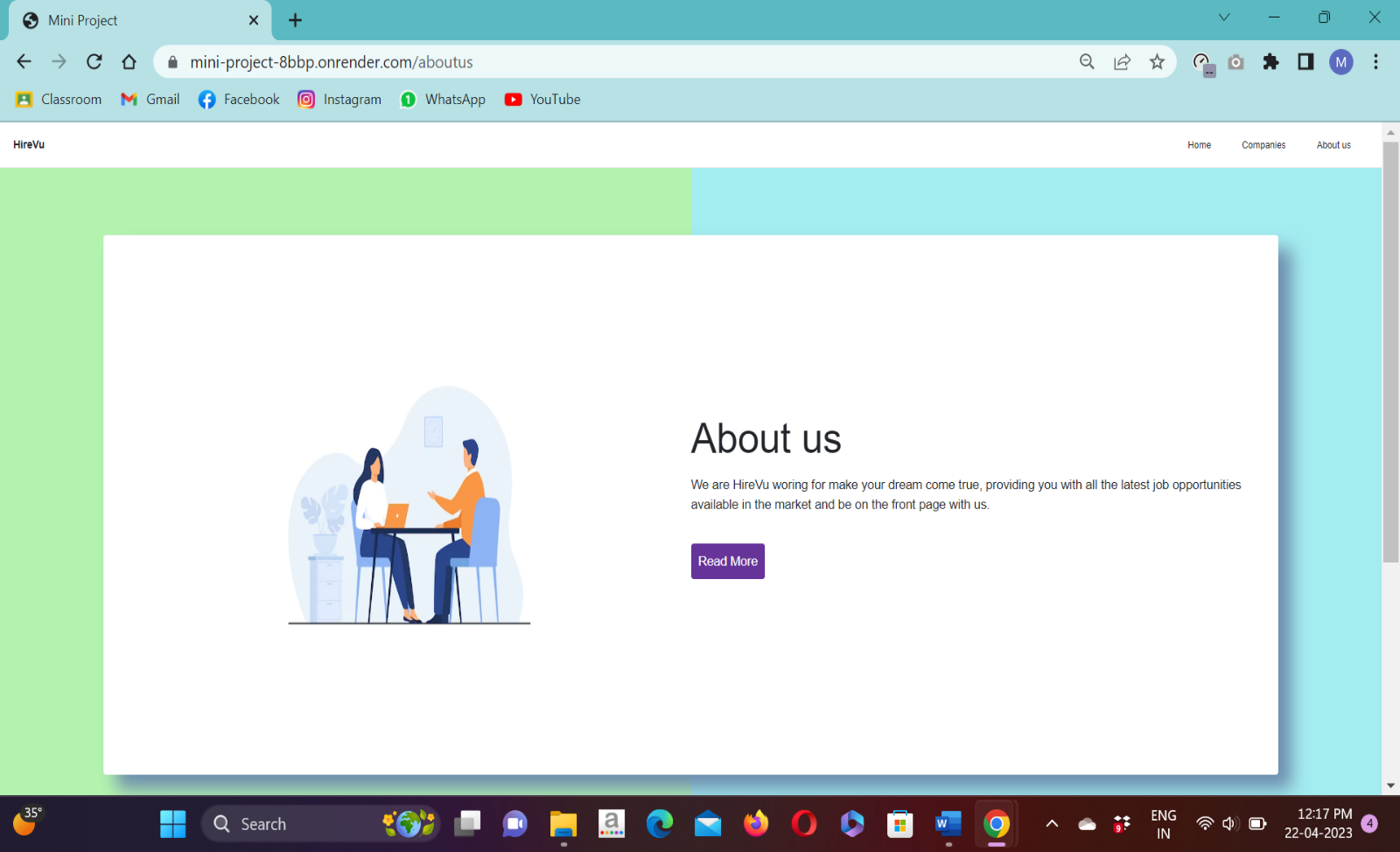
Page 1 : Home Page-

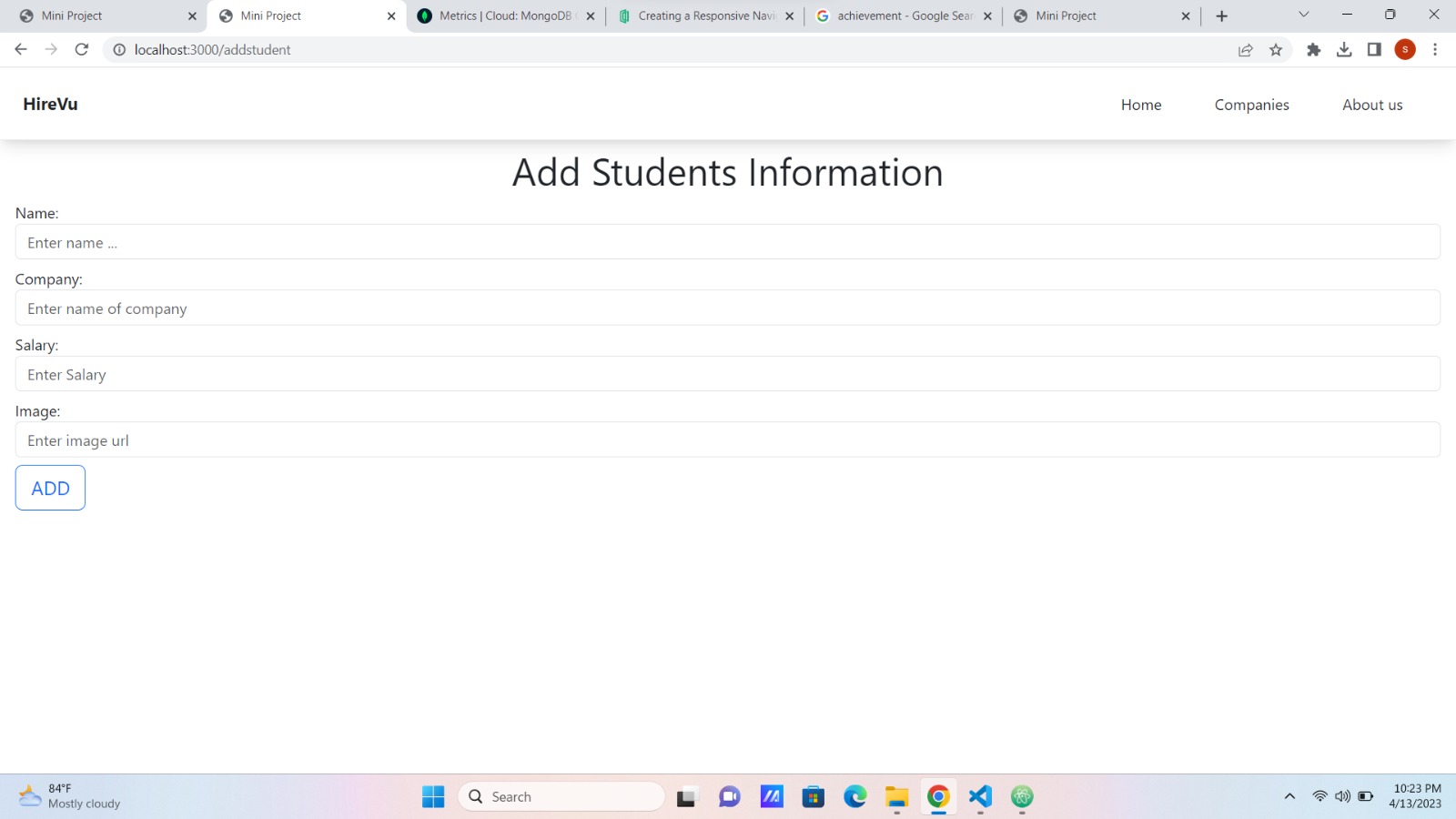


Page 2 : Companies Page for hiring -

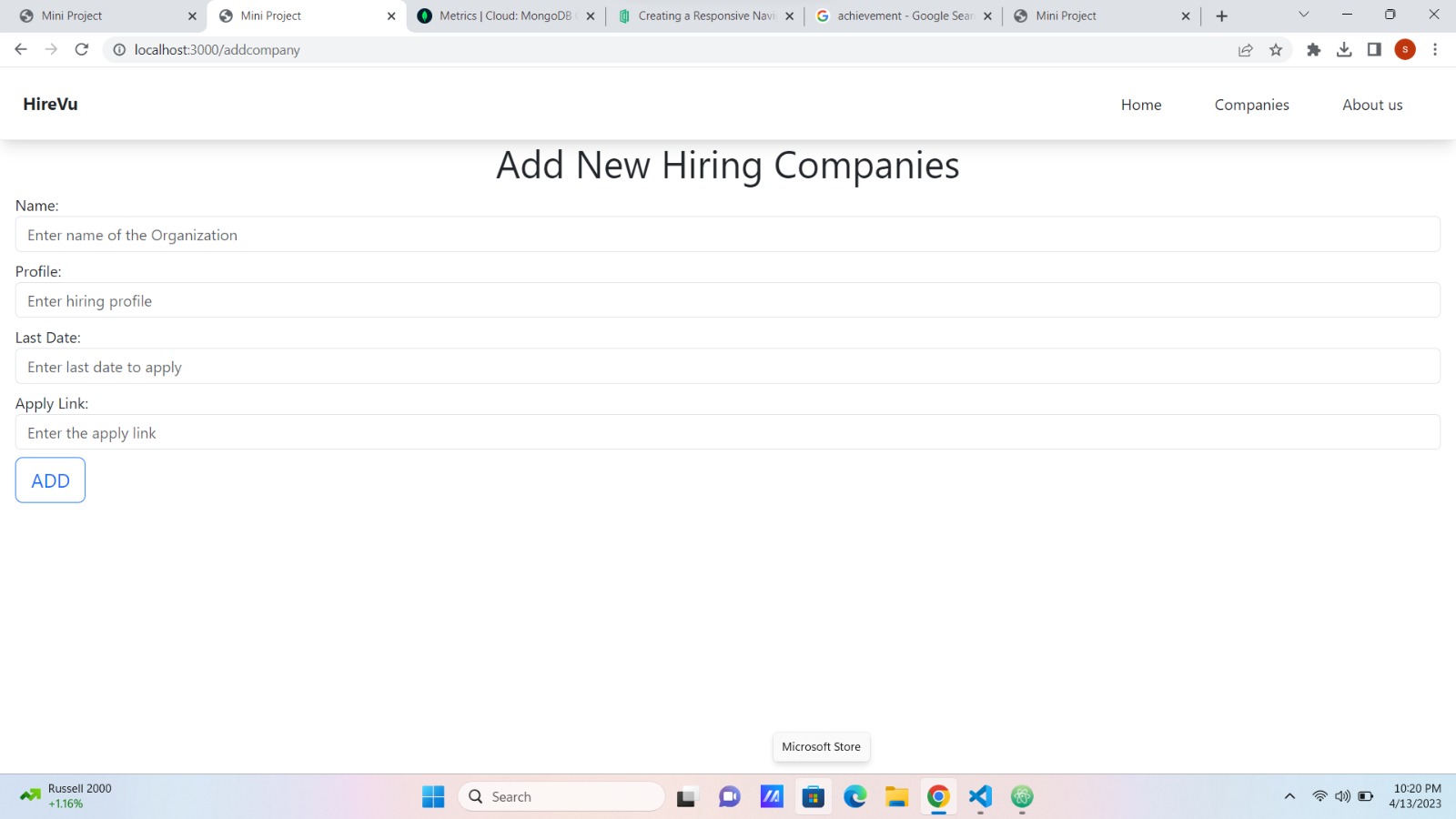


Page 3 : About us -



Page 4 : Add Student

Page 5 : Add new Company Hiring



**Chapter 5**

**ADVANTAGES & DISADVANTAGES**

**5.1 ADVANTAGES**

#### Cost-Effective

#### Web scraping services provide an essential service at a competitive cost. The data will have to be collectedback from websites and analyzed so that the internet functions regularly. Web scraping services manage to do this in a cost-effective and budget-friendly manner.

#### Low Maintenance and Speed

#### Web Scrapingdoes have a very low maintenance cost associated with it over a while. In that way, it helps plan the budget accurately. Also,scraping web saves a lot of time, as it can do a day’s manual work in a few hours.

#### Data Accuracy

#### Simple errors in data extraction can lead to major issues. Hence it is needed to ensure that the data is correct. Data scrapingis not only a fast process, but it’s accurate too. This reputationhelps while collecting important data such as sales price, financial data to name a few.

#### Easy to Implement

#### Once a website scraping service starts collecting data, you can rest assured that you are getting data from not just a single page but from the whole domain. With a one time investment,it can have a high volume of data.

#### 5.2 DISADVANTAGES

#### Data Analysis of Data Retrieved through Scraping the Web

#### To analyze the retrieved data, it needs to be treated first. This often becomes a time-consuming work.

#### Difficult to Analyze

#### For those who are not much tech-savvy and aren’t an expert, web scrapers can be confusing. Even though it’s not a major issue.

#### Speed and Protection Policies

#### Most of the web [scraping services](https://www.promptcloud.com/blog/web-scraping-tool-vs-web-scraping-services/) are slower than API calls. Many websites don’t allow screen scraping. It is a huge challenge to Web Scraping. Also, if any code of the target website gets changed, web scrapersstops capture the data.

**Chapter 6**

### CONCLUSIONS AND SCOPE FOR FUTURE WORK

**6.1 CONLUSIONS**

Building on previous topical work, this study reviews the recent literature relating to the applications of web scraping in various domains, web scraping techniques, and tools that employ web scraping techniques. We use this study to improve our web scraping process, and we discovered that most of the web scrapers are often quite similar and general in nature designed to carry out generic and easy jobs. By comparing the performance and features of different tools and frameworks, we found that Scrapy provides better results as it is fast, extensible, and powerful. Since Scrapy handles requests asynchronously, the results can be scraped rapidly. Furthermore, Scrapy’s architecture is based on a web crawler which enables easy data extraction. Scrapy’s selectors like CSS and XPath can be employed to extract the required data. Scrapy is the perfect tool for complex projects because of its flexible and extensible capabilities, making integration with VPNs and proxies easier. In addition, Scraper API supported browsers, proxies, and CAPTCHAs, allowing you to get raw HTML from any website with a single API call.

**6.2 FUTURE WORK**

A lot of future work can be done in this area. Developers are just now discovering the potential behind web scrapers and the one developed for this project is very simple. Future work can be done on building an automated web scraper where you user can press what they want to scrape from the website. A user interface can also be developed for the web scraper so that the user can easily understand how the web scraper work. The scraper can also be developed further internally, by adding more things to scrape and making sure that it handles the different types of data that exists on TimeEdit. Future research can be done on how data mining methods can be applied to the dataset. If more data was scraped, or data over an entire year, different data mining methods can be applied to analyze the data set. Research in this area could show where students succeed or fail and how different arrangement of courses affect the students. With the help of web scraping and data mining good statistics can be produces and these statistics can be used to further develop the program. Perhaps, developing a website where the user can add a link which then gets web scraped by a Python script in the background is a good start for future work. This project shows that the potential upsides of web scraping is massive but that more research can be done on how to implement it.

### 

### REFERENCES

1. <https://en.wikipedia.org/wiki/Web_scraping>
2. <https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/>
3. <https://docs.python.org/3.9/tutorial/index.html>
4. <https://www.w3schools.com/js/js_htmldom_navigation.asp>
5. Sirisuriya, D. S. (2015). A comparative study on web scraping. In the Proc. 8th Int. Res. Conf. KDU, 135– 140.
6. Phan, H. (2019). Building Application Powered by Web Scraping. Doctoral Thesis
7. Rahmatulloh, A. and Gunawan, R. (2020). Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar. Indonesian Journal of Information Systems, 2(2):95-104.