A PROJECT REPORT ON

**WEB SCRAPING USING PYTHON**

Submitted in partial fulfillment of the requirement for the IV semester.

Bachelor of Technology

By

Vashu Sharma

2017785



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

GRAPHIC ERA DEEMED TO BE UNIVERSITY

DEHRADUN

2022-2023

DECLARATION

I, Vashu Sharma student of B-tech, Semester 4, Department of Computer Science and Engineering, Graphic Era Deemed To Be University, Dehradun, declare that the technical project work entitled “**WEB SCRAPING USING PYTHON**” has been carried out by me and submitted in partial fulfillment of the course requirements for the award of degree in B- tech of Graphic Era Deemed University during the academic year 2022-2023. The matter embodied in this synopsis has not been submitted to any other university or institution for the award of any other degree or diploma.



CERTIFICATE

This is to certify that the project report entitled “**WEB SCRAPING USING PYTHON**” is a bonafide project work carried out by Dr Garima Sharma. In partial fulfillment of award of degree of B tech of Graphic Era Deemed To Be University, Dehradun during the academic year 2022-2023. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated. The project has been approved as it satisfies the academic requirements associated with the degree mentioned.

Dr Garima Sharma

ACKNOWLEDGEMENT

Here by I am submitting the project report on “**WEB SCRAPING USING PYTHON**” as per the scheme of Graphic Era Deemed University, Dehradun.

I would like to express our sincere gratitude to Dr Devesh Pratap Singh, Head of Dept. of Computer Science, for providing a congenial environment to work in and carry out our project.

I consider it mine cardinal duty to express the deepest sense of gratitude to Dr Upma Jain, Department of Computer Science and Application for the invaluable guidance extended at every stage and in every possible way.

I would like to also thanks skyfi labs for helping me in better understanding each component of topic in an interesting way.

Finally I am very much thankful to all the faculty members of the Department of Computer Science and Technology, friends and our parents for their constant encouragement, support and help throughout the period of project conduction.

Vashu Sharma

2017785

**INTRODUCTION**

Suppose you want some information from a website? Let’s say a paragraph on Donald Trump! What do you do? Well, you can copy and paste the information from Wikipedia to your own file. But what if you want to get large amounts of information from a website as quickly as possible? Such as large amounts of data from a website to train a [Machine Learning](https://www.geeksforgeeks.org/machine-learning/) algorithm? In such a situation, copying and pasting will not work! And that’s when you’ll need to use **Web Scraping**.



What is Web Scraping?

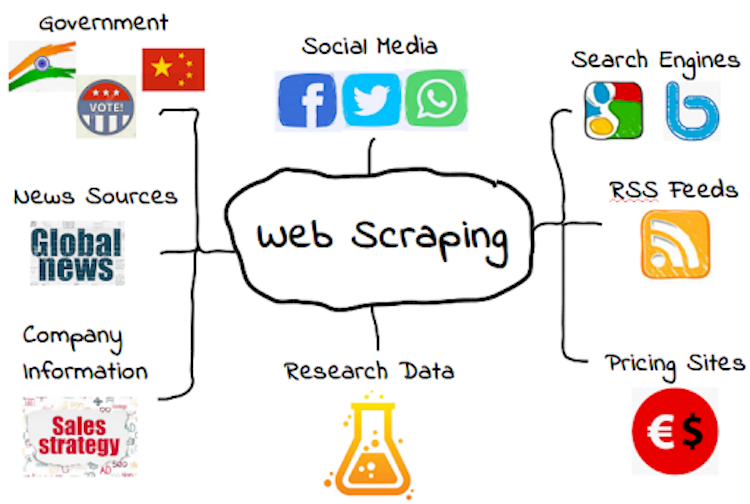
Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. Many large websites, like Google, Twitter, Facebook, StackOverflow, etc.

Web scraping requires two parts, namely the **crawler** and the **scraper**. The crawler is an artificial intelligence algorithm that browses the web to search for the particular data. The scraper, on the other hand, is a specific tool created to extract data from the website.

### How  Web Scrapers Work?

Web Scrapers can extract all the data on particular sites or the specific data that a user wants. Ideally, it’s best if you specify the data you want so that the web scraper only extracts that data quickly. For example, you might want to scrape an Amazon page for the types of juicers available, but you might only want the data about the models of different juicers and not the customer reviews.

### Why is Python a popular programming language for Web Scraping?



[Python](https://www.geeksforgeeks.org/python-programming-language/) seems to be in fashion these days! It is the most popular language for web scraping as it can handle most of the processes easily. It also has a variety of libraries that were created specifically for Web Scraping. [**Scrapy**](https://scrapy.org/) is a very popular open-source web crawling framework that is written in Python. It is ideal for web scraping as well as extracting data using APIs. [**Beautiful soup**](https://pypi.org/project/beautifulsoup4/) is another Python library that is highly suitable for Web Scraping. It creates a parse tree that can be used to extract data from HTML on a website. Beautiful soup also has multiple features for navigation, searching, and modifying these parse trees.

### What is Web Scraping used for?

Web Scraping has multiple applications across various industries. Let’s check out some of these now!

#### 1. Price Monitoring

Web Scraping can be used by companies to scrap the product data for their products and competing products.

#### 2. Market Research

Web scraping can be used for market research by companies.

#### 3. News Monitoring

Web scraping news sites can provide detailed reports on the current news to a company.

#### 4. Sentiment Analysis

If companies want to understand the general sentiment for their products among their consumers, then Sentiment Analysis is a must.

#### 5. Email Marketing

Companies can also use Web scraping for email marketing. They can collect Email ID’s from various sites using web scraping.

Methodology

* **Copy-pasting.** The manual human examination and copy-pasting method may sometimes prove irreplaceable. At times, this technique may be the only practical method to use especially when websites are setup with barriers and machine automation cannot be enabled.
* **DOM Parsing**. In order to dynamically modify or inspect a web page, client-side scripts parse the contents of the web page into a DOM tree. By embedding a program into the web browser, you can then retrieve the information from the tree.
* **HTTP Programming**. Using socket programming, posting HTTP requests can help one retrieve dynamic as well as static web page information.
* **Recognizing Semantic Annotation.** Most web pages have semantic annotations/markup or metadata that can be easily retrieved. This could be a simple case of DOM parsing if the metadata is just embedded in the web page.
* **Text Grepping**. Using Python programming languages or Perl, one can use the UNIX grep command to extract valuable data and information from web pages.
* **Web scraping Software**. If you do not want to manually use web-scraping codes, you can make use of a software that can do the web scraping for you.

Conclusion

* The completion of the project went quiet well, I learned much new things while I was building up it, and I get up to know various platforms which help us to learn all this stuff. I was able to learn the practical use of web. The practical helped me to learn the debugging of code and development tools of home automation Project.
* Overall working on this project was great fun as I came up with great piece of knowledge and understanding of the topic.
* Reference:
* Youtube
* www. Googlr.com