TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
**https://www.tejasjournals.com/**

# Apply different Efficient Data Extraction Techniques in real world application

Aman Yadav[a], Princy Srivastava [b] and Dr. Anil Kumar Pandey [c]

[a,b,c] Computer Science, Shri Ramswaroop Memorial University, Barabanki, India, 225003
aman310k@gmail.com, princysrivastava27@gmail.com, anipandey@gmail.com

KEYWORD

*Data Extraction; Machine Learning; Natural Language Processing (NLP); Digital Insights; Document Processing*

ABSTRACT

*In this research paper, we are performing data extraction by using Efficient Data Extraction Techniques tool for the headache-free manner. We inspect various strategies for extracting valuable information form huge amounts of data.*
*These techniques use from simple rules to more modern methods involving smart machines that can learn on their own. These tool work to save us time and effort by extracting the essential information of bits from the very large data landscape.*

## 1. Introduction

In this research paper, we are inspecting several techniques and methodologies of data extraction in a very effective manner to be more correct. we are analyzing various Data Extraction techniques like Traditional copy & paste, Web scraping software, and Web scraping tools such as import.io, Scrapy, etc.

We are going to discuss about assorted approaches to find the best content in this big digital library. They have many books and papers, but it might be like trying to find a needle in a haystack to find what you need. The goal of our research is all about to make this search easy.

In the digital world data is like a box of toys, and getting the information can be a big roll. Efficient Data Extraction Techniques using a super-smart tool to pick out only the puzzle pieces you need, without worrying about the action figures or filled animals. It works like having a digital detective that saves us time and effort.

## 2. Overview

Data extraction has important criteria to analyze the web pages. It works as a guide to help people to pull out information from a big digital library. At first, it was all about the basics - simple rules to find the exact data.

Web scraping is like a tool to take useful information from websites and organize neatly. By organizing messy data from websites, we can store it in databases or spreadsheets. People call it web data extraction or screen scraping. The main work to get the information from websites and make it easy to understand in things like spreadsheets or CSV files.

Think of it as a robot doing a copy-and-paste job. This robot are virtual computer agent, acts like a person, clicking links, and filling out forms. It more useful for getting exact details, like prices, stocks, or market trends, helping businesses. But, be careful—too many requests too quickly can stress out a server, almost like a website

**Corresponding Author:** Aman Yadav, Shri Ramswaroop Memorial University, Barabanki, India 225003
**Email:** aman310k@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

traffic jam. So, while web scraping is a useful tool for getting the goods from the internet, it's got to be done in a friendly way to not bug the servers too much.

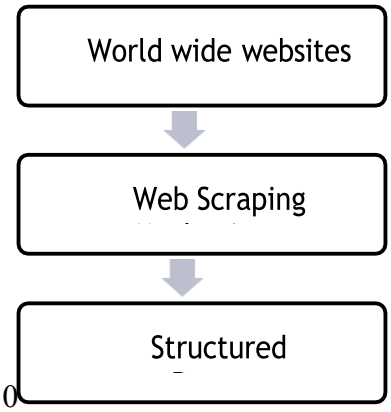


Figure 1. Architecture of Web

## 3. Web scraping uses:

Web scraping is a tool that have better automation and efficiency at higher level, it also provides helpful data for decision making at times. However, agreements & ethical considerations to legal standards are required for its execution.

- Data Analysis and Research
- Competitor Monitoring
- Price Tracking
- Content Aggregation
- Job Market Insights
- Real Estate Market Research
- Social Media Analysis
- Weather Data Collection
- Government and Public Data Retrieval
- Lead Generation
- Content Monitoring


Figure 2. Web Scraping uses

## 4. Methodologies:

Templates in Word that implement these instructions can be retrieved

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

### Traditional Copy-and-Paste

The simple way to doing this things where you just manually copy information from a webpage and paste it somewhere else. In this method a given text is easily copied and then pasted wherever you want. [1]

### Text Grapping and Regular Expression Matching

This one includes locating using regular expressions to find and pick out specific text of bits from a webpage.

### HTTP Programming

Simple way to asking a website and get information, then computation what it says back. [2].

### HTML Parsing

Find information by looking at the secret code, HTML of a website and getting the main parts. [1]

### DOM Parsing (Document Object Model)

Going through its different parts to find what you want and inspect the data of a website. [2]

### Web Scraping Software

Using computer programs that make web scraping easy and less immediate.

### Vertical Aggregation Platforms

Vertical Aggregation Platforms that bring together data from lots of different places to give you an approach.

### Semantic Annotation Recognizing

Getting important data on a website based on special tags or labels in the secret code..

## 5. Web Scraping Tools

### 5.1 Beautiful Soup

It helps read and understand HTML and XML stuff. Beautiful Soup is like a confidant tool for Python lovers. You can use it to search and find things in the website information. People get it because it's simple and stable to get the job done in web scraping.

### 5.2 Selenium

It can control web browsers and chat with JavaScript data. Selenium is a superhero tool for websites. Perfect if you are working with websites that play a lot with JavaScript.

### 5.3 Scrapy

It's working like a set of cheat codes for web scraping. Scrapy is a Python thing that makes web scraping less of a headache. It knows what you want and helps you to get information from websites without too much commotion. [1]

### 5.4 Octoparse

It works like, "Hey, you don't need to be a tech wizard for this." Click on what you need, and Octoparse helps you get the data. Octoparse is the easy to use on web scraping tools .

### 5.5 ParseHub

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
**https://www.tejasjournals.com/**

It makes getting data as easy line. It works on both Windows and Mac, so everyone can use the party. ParseHub works like the artist of web scraping tools.

### 5.6 Apache Nutch

It's a big deal for handling lots of data on web scraping. Developed in Java, it is part of a cool tools by the Apache folks. Apache Nutch works like the giant web crawler on the block

### 5.7 Import.io

Import.io work like the neighborhood web data constructer. It changes website data for you. You can use simple clicks or dive into the tech side with APIs. [1]

### 5.8 OutWit Hub

OutWit Hub is your childhood friend for scraping data. It's there for both Windows and Mac users, making the whole process of getting information from the website easy. [1]

Indicate each footnote in a table with a superscript lowercase letter.

### 5.9 Content Grabber

Content Grabber works like the handyman for web scraping. It gives you a computer graphics setup to getting data. Perfect for getting information from websites that keep changing..

### 5.10 Mechanical Turk

Number tables consecutively in accordance with their appearance in the text. Place footnotes to tables below the table body and indicate them with superscript lowercase letters. Avoid vertical rules. Be sparing in the use of tables and ensure that the data presented in tables do not duplicate results described elsewhere in the article.

## 6. Discussion

In this table [1] the different Web Scraping tools are showing with the Operating System they support. These web scraping tools can only work on their comparable operating systems but there have some tools that are web based such as import.io & Mechanical Turk and it works on every operating system through web.

Table 1. Web Scraping Tools and OS they support

| **Tools** | **Supporting OS** |
|---|---|
| Beautiful Soup | Windows, macOS, Linux |
| Selenium | Windows, macOS, Linux |
| Scrapy | Windows, macOS, Linux |
| Octoparse | Windows, macOS |
| ParseHub | Windows, macOS, Linux |
| Apache Nutch | macOS, Linux |
| Import.io | Web-based |
| OutWit Hub | Windows, macOS |
| Content Grabber | Windows |
| Mechanical Turk | Web-based |

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
**https://www.tejasjournals.com/**

## 7. Acknowledgment

First and foremost, we would like to express our deepest gratitude to our mentor, Dr. Anil Kumar Pandey, for his unwavering support and guidance throughout our research on web scraping techniques. His expertise and mentorship have been invaluable in shaping this journal article. We would like to thank our university, Shri Ramswaroop Memorial University, for providing us with the necessary resources and facilities to carry out this research. Lastly, we would like to express our appreciation to our colleagues, friends, and family for their encouragement and support during this research journey. In conclusion, we are deeply grateful for the contributions of everyone who has played a part in the completion of this journal article on Apply different Efficient Data Extraction Techniques in real world application.

## 8. References

[1]. S.C.M. de S Sirisuriya, 2015, A Comparative Study on Web Scraping. Proceedings of 8th International Research Conference, KDU.

[2]. Anand V. Saukar, Kedar G. Pathare, Shweta A. Gode, An Overview On Web Scraping Techniques And Tools, International Journal on Future Revolution in Computer Science & Communication Engineering.

[3]. Kishor Kumar Reddy, A Text Mining using Web Scraping for Meaningful Insights, Journal of Physics: Conference Series.