

# Data Aggregation by Web Scraping using Python

Mr. H. Sathish<sup>1</sup>, Y. Shiva Sai<sup>2</sup>, J. Keerthana<sup>3</sup>, A. Anjali<sup>4</sup>, T. Vaishnavi<sup>5</sup>, D. Deekshitha<sup>6</sup>

Associate Professor, Department of Computer Science & Engineering<sup>1</sup>

UG Students, Department of Computer Science and Engineering<sup>2,3,4,5,6</sup>

Christu Jyothi Institute of Technology & Science, Jangaon, Telangana, India

**Abstract:** *The standard data examination are based on the root and effect relationship, formed a model tiny assessment, abstract and quantitative assessment, the level headedness approach of making extrapolation assessment. The Web Scraper's scheming morals and methodology are compared, it clarifies about the working of how the scrubber is planned. The strategy of it is distributed into three pieces: the web scrubber draws the ideal connections from web, and afterward the information is removed to get the information from the source joins lastly stowing that information into a CSV record. The Python language is carried out for the completing. Thusly, connecting every one of these with the ethical information of libraries and working skill, One can have a satisfactory Scraper in our grasp to produce the ideal outcome. Because of a gigantic local area and library assets for Python and the impeccableness of coding stylish of python language, it is generally proper one for Scraping wanted information from the ideal site.*

**Keywords:** Data analysis, Web Scraping, Implementing Web Scrape

## I. INTRODUCTION

Data analysis is the method of extracting solutions to the problems via interrogation and interpretation of data. The analysis process comprises of discovering problems, resolve the accessibility of suitable data, determining which method can help in finding the solution to the interesting problem and convey the result. For the purpose of analysis, the data has to segregate into various steps further on such as starting with its specification assembling, organizing, cleaning, re-analyzing, applying models and algorithms and the final result. Web information scraping and publicly supporting are outstanding strategies for naturally creating substance on web. A considerable amount of individuals utilized these strategies in research and business for creating substance or offering criticisms to expand the exactness of business advertising that enables individuals to deliver resources in advancing and developing the business. By and large, web scraping is notable for a "Screen Scraping", "Web Data Extraction". The scraper tool for the web is utilized for derived information from the web host, and as a portion of uses used for web orders, web mining and data mining, online esteem change observing and value correlation, element survey scratching (to watch the challenge), gathering land postings, atmosphere data checking, webpage change area, inspect, following on the web closeness and reputation, web mash up and, web data joining. Pages are manufactured utilizing content-based increase dialects (HTML and XHTML), and much of the time contain a profusion of cooperative info in the content structure. Be that it may be as most website pages are anticipated for human end users and not for minimalism of robotized use. Thus the toolbox that scrapes web info was made. As for the paper will be focused on the data analysis using python's effectiveness as a programming language, it's out to an apt choice as a single language for the data-centric application, For this, the version of Python used will be Python 3.6 for the analysis.

## II. AIM AND OBJECTIVE

### 2.1 Aim

The principle point of the paper is to contemplate a cycle of cleaning, changing, and demonstrating information to find helpful data for business dynamic. The motivation behind Data Analysis by web scratching is to extricate valuable data from information and taking the choice dependent on the information investigation.

### 2.2 Objective

The general objective of the study is to propose a reliable, convenient and accurate detection system. The study has the following specific objectives:

Copyright to IJAR SCT

[www.ijarsct.co.in](http://www.ijarsct.co.in)

DOI: 10.48175/IJAR SCT-18177



519

- The point of the paper is to remove the information from different sources with the assistance of programming known as the web crawler Scrapy.
- The software is used to extract data using an application programming interface or as a general-purpose web crawler required by the desired customer.
- To analyze the variation, comments, ratings or anything else with innumerable options.

**III. LITERATURE SURVEY**

**Paper 1: Data Analysis By Web Scraping Using Python:**

This paper depicts a standard data examination are based on the root and effect relationship, molded a model little assessment, abstract and quantitative assessment, the judiciousness approach of making extrapolation assessment. The method of it is dispensed into three parts: the web scrubber draws the ideal connections from web, and afterward the information is extricated to get the information from the source joins lastly stowing that information into a CSV document. Because of a gigantic local area and library assets for Python and the impeccableness of coding stylish of python language, it is most suitable one for Scraping wanted information from the ideal website.

**Paper 2: Web Scraping Using python:**

Learn web scratching and creeping procedures to get to limitless information from any web source in any organization. Ideal for developers, security experts, and web managers acquainted with Python, this book trains essential web scratching mechanics, yet in addition digs into further developed subjects, for example, investigating crude information or utilizing scrubbers for frontend site testing.

**Paper 3: Web Scraping with Python: Successfully scrape data from any website with the power of Python:**

The Internet contains the most helpful arrangement of information at any point collected, generally openly open free of charge. Notwithstanding, this information isn't effectively reusable. It is implanted inside the design and style of sites and should be painstakingly separated to be valuable. Web scratching is getting progressively valuable as a way to effortlessly assemble and sort out the plenty of data accessible on the web. Utilizing a straightforward language like Python, you can creep the data out of complex sites utilizing basic programming.

**IV. EXISTING SYSTEM**

In Existing system is the manual web data extraction process has two major problems. Firstly, it can't measure costs efficiently and can escalate it very quickly. The data collection costs increase as more data is collected from each website. In order to conduct a manual extraction, businesses need to hire large number of staff, this increases the cost of labour significantly. Secondly, each manual extraction is known to be error prone. Further, if any business process is very complex then cleaning up the data can get expensive and time consuming. The below figure explains the errors and data cleanup processes problems with the Manual method<sup>[1]</sup>.

**V. COMPARTIVE STUDY**

SR NO.	PAPER TITLE	METHOD	ADVANTAGE	DISADVANTAGE
1.	Data Aggregation by web Scraping Using Python	Python, Web Scraping , (Beautiful Soup), Implementing Web Scrape	Easy to implement	Time Consuming
2.	Web Scraping Using python	Python, Web Scraping , (Beautiful Soup)	Accuracy of results	Difficult to understand
3.	Web Scraping with Python: Successfully scrape data from any website with the power of Python	Python, Web Scraping , (Beautiful Soup)	Low maintenance and speed	Protection policies

## **VI. PROBLEM STATEMENT**

The world of retail is changing rapidly. Many brick and mortar locations are closing and being replaced by online stores, direct to consumer brands, and subscription services. However, while the breadth of assortment is something that drives customers to website, a lot of E-Commerce platforms fail to sell through a high percentage of merchandise

## **VII. PROPOSED SYSTEM**

Web Scraping (web harvesting or web data extraction) is a computer software technique to extract information from websites. Usually, such programming programs recreate human investigation of the World Wide Web by either executing low-level Hyper content Transfer Protocol (HTTP), or installing a completely fledged internet browser, like Internet Explorer or Mozilla Firefox. Web Scraping is firmly identified with web ordering, that lists data on the web utilizing about web crawler and is a widespread method received by most web indexes. Conversely, Web Scraping centers more around the change of unstructured information on the web, ordinarily in HTML design, into organized information that can be put away and investigated in a focal neighborhood data set or accounting page. The pressure identification module examines the parallel picture from the limit left top to record the co-ordinates of the eyebrow. The stress detection module scans the binary image from the extreme left top to record the co-ordinates of the eyebrow. The offline displacement calculation sub-module calculates the shifting of eyebrow using the obtained eyebrow co-ordinates which is subsequently followed by variance calculation of the displacement. The classifier sub-module is trained offline are employed to determine the presence of emotion. The integrated decision of individual frames eventually determines the level of stress involved. Web Scraping is a technique to extract structured data from websites. WSAPI is the platform that enables an organization to extend their existing web based system, as well designed set of services for creating new channels, developer integration or partner integration.

## **VIII. MATHEMATICAL MODEL**

### **Logistic Regression**

It is a type of classification algorithm, it is used when there would be only Binary output, i.e., the result belongs to one class or another e.g., 0 or 1. Logistic Regression should only be used when the target variables are discrete. Logistic Regression is a kind of powerful machine learning algorithm it uses a sigmoid function, it is best suitable for binary classification problems, but it can be used in multiclass classification problems can be used with "one vs all" method. sigmoid function

$$S(z) = 1/(1+e^{-z})$$

### **Linear Regression**

Straight relapse is one of the simplest and most well known Machine Learning calculations. It is a measurable strategy that is utilized for prescient examination. Straight relapse makes expectations for persistent/genuine or numeric factors like deals, compensation, age, item cost, and so on Direct relapse calculation shows a straight connection between a reliant (y) and at least one autonomous (x) factors, consequently called as direct relapse. Since straight relapse shows the direct relationship, which implies it discovers how the worth of the reliant variable is changing as indicated by the worth of the free factor. The direct relapse model gives a slanted straight line addressing the connection between the factors. Mathematically, it can represent a linear regression as:

$$y = a_0 + a_1x + \epsilon$$

## **IX. SYSTEM ARCHITECTURE**

### **Description:**

Web Scraping is a strategy to separate organized information from sites. WSAPI is the stage that empowers an association to expand their current electronic framework, too planned arrangement of administrations for making new channels, designer mix or accomplice joining. It assists with offering spotless and organized information from existing sites, so the information can be easily devoured by unique frameworks. The innate plan assists engineers with fusing site changes without influencing the extraction rationale by moving them to designs. There are numerous particular reasons why organizations might need to scratch their site; one of the essential explanation being the inaccessibility of APIs.

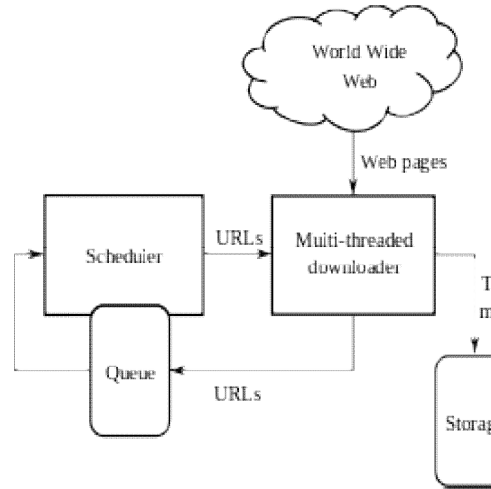


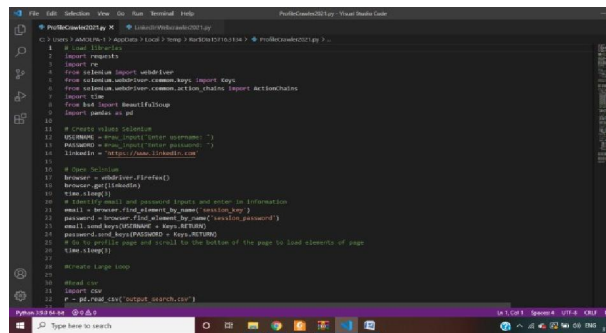
Fig.1: System Architecture

**X. ADVANTAGES**

- o Output in which result is modified image or report that is based on image analysis.
- o Stress Detection System enables employees with coping up with their issues leading to stress by preventative stress management solutions.
- o Create Applications for Tools that don't have a public developer API.
- o Web scraping services provide an essential service at a low cost. It is paramount that data is collected back from websites and analyzed so that the internet functions regularly.

**XI. METHODOLOGY**

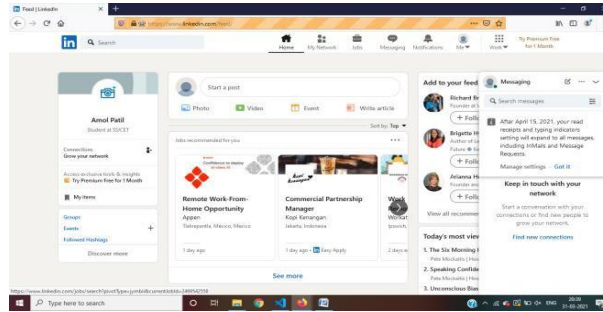
The methodology used for the paper is to gather all the data extracted from various sources by using the vivid features of the web crawler scrapy using the scripts written in python language and further analyze it as per the requirements of the customer where the data is stored in the company's database. Coding The basic web crawling script used for the project which shows the data crawled and stored in the database of the products from a social network site.



```

1 # scrapy spider
2 import scrapy
3
4 import re
5 from scrapy.spiders import Spider
6 from scrapy.selector import Selector
7 from scrapy.http import Request
8 from scrapy.http import Response
9 from scrapy import Item
10
11 class SpiderName(scrapy.Spider):
12     name = 'SpiderName'
13     allowed_domains = ['example.com']
14     start_urls = ['http://example.com']
15
16     def parse(self, response):
17         # Extract data from the page
18         selector = Selector(response)
19         items = []
20         # Loop through all the items on the page
21         for item in selector.xpath('//div[@class="item"]'):
22             # Extract the item's text
23             text = item.xpath('text()').get()
24             # Create a new item
25             item = Item()
26             item['text'] = text
27             items.append(item)
28
29         return items
  
```

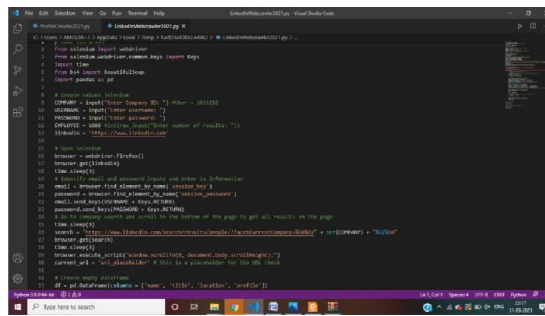
Fig 2: Code for Implementation of Scrapy



**Fig 3: Scrapping of Website**

**Testing**

The project was tested by using the various components as defined earlier and made to run on the browser. The extraction done turns out to be completely relevant and the analysis made is estimated.

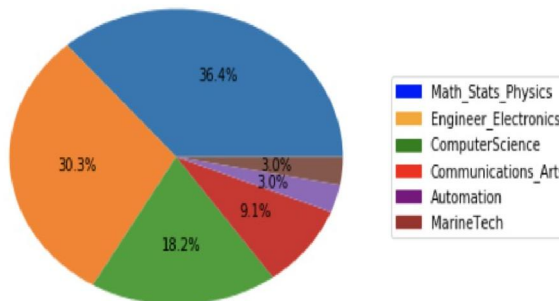


**Fig4: Code for analyzing the data after scrapping**

**X. RESULTS**

The overall results of the project turn out to be helpful to understand. The Web scrappy extracted the data and made into csv file format. The script which was written to extract the data turned out to be both of finding each of these sources provided with great ease. Moreover, the analysis done has shown the most searched content in the site taken for test in the percentage format.

**Bachelor Degree Received**



### **XI. CONCLUSION**

Thus, we have tried to implement the paper "David Mathew Thomas, Sandeep Mathur", "Data Analysis by Web Scraping Using Python" in IEEE 2019 and according to the extraction of data hidden web data is a major challenge nowadays because of autonomous and heterogeneous nature of hidden web content traditional stress engine has now become an ineffective way to search this kind of data. The main outcomes of this project were user friendly search interface, indexing, query processing, and effective data extraction technique based on web structure, form submission analysis and new submission plan. Hidden web data need synthetic and semantic matching to fully achieve automatic integration in this thesis fully automatic and domain dependent prototype system is proposed that extract and integrate the data lying behind the search form.

### **REFERENCES**

- [1]. "Renita Crystal Pereira, Vanitha T. "Web Scraping of Social Networks." International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, pp.237-239, Oct. 7, 2018"
- [2]. "Ghazvinian, Holbert, Viswanathan. "SimpleWebScraping." Internet: <https://seanolbert.wordpress.com/2011/07/15/scrapy-simple-webscraping/>, Jun. 2015"
- [3]. "Bellarosey. "Crowdsourcing-Definition." Internet: [http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing\\_a.html](http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html), Jun. 02, 2006"
- [4]. "BrightPlanet.com Deep web White Paper. <http://www.completeplanet.com/Tutorials/DeepWeb/index.asp>."
- [5]. "Kolari, Pand Joshi A. , "Web mining : research and practice , Computing in Science & Engineering", IEEE Transactions on Knowledge and Data Engineering, vol. 6, no. 2, Vol. 6 , No. 4, 2004"
- [6]. "Kengtel, W: Wagner, M. Proteins 1999, 37, 334-345."
- [7]. "Datahen." 3 Advantages of web scraping for your enterprise" Internet: <https://www.datahen.com/3-advantages-web-scraping-enterprise/>, May. 17, 2017"
- [8]. "http://resources.distilnetworks.com/h/i/53822104-is-webscraping-illegal-depends-on-what-the-meaning-of-the-word-is-is/181642"