

SEMINAR ON

“Web scraping using python”



Presented By:

DEPARTMENT OF COMPUTER
SCIENCE ENGINEERING

- 1) Vishvanath Metkari
- 2) Shubham Nikam
- 3) Akash shinde
- 4) Saurabh Bane

Guided By:

Guide Name: Prof. Mr. Pathak.P.A

Designation: Assistant Professor

Department: Computer Science &
Engineering

College Name: Arvind Gavali College of
Engineering, Satara



Agenda

- Introduction
- Research/Project Area/Domain and its Justification
- Title of the Project/Problem Statement
- Objective of the Proposed Work
- Literature Review
- Methodology/Algorithm/Flowchart/Circuit Diagram
- Results and Discussion (100% Implementation)
- Conclusion
- Future Scope
- References



INTRODUCTION

- Python Web scraping is nothing but the process of collecting data from the web.
- Web scraping in Python involves automating the process of fetching data from the web. In order to fetch the web data, all we need is the URL or the web address that we want to scrape from.
- The fetched data will be found in an unstructured form. In order to make use of the data or collect useful insights from it, we transform it into a structured form. Once converted into a structured form, we need to store the data for further processing. The whole process is called web scraping.



RESEARCH/PROJECT AREA

Website

- A website is set of related web pages served from a single web domain.
- Website is hosted on one or more web servers.

Portal Area

- A portal is a private location on the internet, accessible with a unique URL (web address).

programming languages

- Web scraping with Python is easy due to the many useful libraries available.
- One of the Python advantages is a large selection of libraries for web scraping.
- There are several types of Python web scraping libraries from which you can choose.

Request, BeautifulSoup, lxml, Selenium

Problem Statement

- Now students are spending a lot of time searching for internship opening or job opening on the websites, many times they have to do crawling.
- So from this project you can save a lot of student time.
- For this, we can save the internship or job opportunities on internshala.com in a .CSV file and pass it on to the student. They will see all the data in.

Objective of the Proposed Work

- There is a lot of data on internshala.com but you have to visit every pages in it , so it seems like a big process.
- So from this project you can find out how many opening there are for internship or job on internshala.com
- Also, the name of the company , Domain name , Stipend , location , and a lot of things can be stored in one place and pass to the students.
- So they can see all the data on a single page.

Literature Review

SrNo	AUTHOR	METHOD FOLLOWED	OUTCOME
1)			



SrNo	AUTHOR	METHOD FOLLOWED	OUTCOME

Algorithm/Flowchart/Circuit Diagram

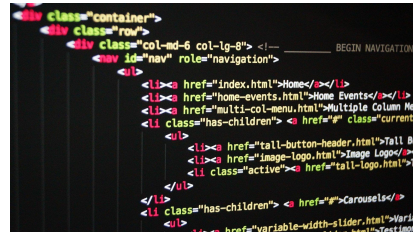
Request



Web datasource

Response

HTML Document



Extract the data from Response

Python script



Save data



URL(<https://internshala.com/>)

Methodology

Steps for Scraping Any Website

To scrape a website using Python, you need to perform these four basic steps:

- Sending an HTTP GET request to the URL of the webpage that you want to scrape, which will respond with HTML content. We can do this by using the Request library of Python.
- Fetching and parsing the data using BeautifulSoup and maintain the data in some data structure such as Dict or List.
- Analyzing the HTML tags and their attributes, such as class, id, and other HTML tag attributes. Also, identifying your HTML tags where your content lives.
- Outputting the data in any file format such as CSV, XLSX, JSON, etc.



Step 1: Importing the required libraries

```
import requests  
  
from bs4 import BeautifulSoup
```

Step 2: Getting the URL and storing it in a variable.

```
siteUrl = "https://internshala.com/internships/data%20science-internship "
```

Step 3: Making a request to the website using the requests library.

Here we use the requests library by passing “url” as a parameter, be careful don’t run this multiple times. If you get like Response 200 then its success, if you get something else then there is something wrong with maybe the code or your browser I don’t know.

```
response = requests.get(siteUrl)  
  
print(response)
```

```
<Response [200]>
```



Step 4: Using the BeautifulSoup library to get the HTML (raw) data from the website.

Here we use the BeautifulSoup by passing the response.text as a parameter and using the HTML parser.

```
main_container = BeautifulSoup(response.text , 'html.parser')
```

Step 5: Using main_container.findAll method to get the respected tag that we are looking for.

You can then copy the HTML tag and class/id if any, and then place it inside the main_container.findAll method. In this case, the HTML tag is 'div' and id is "list_container".

```
sub_container = main_container.find('div', {'id': 'list_container'})
```

Step 6: Removing all the HTML tags and converting it to a plain text format.

Here we remove all the HTML tags and convert it to a text format, this can be done with the help of .text method. This converts the HTML into the text format.

```
profile_name_title = job_div.find('div', class_='heading_4_5 profile')  
profile_name_title = profile_name_title.text
```

Results and Discussion (100% Implementation)

```
Runt Internshala_web_scrapping (1) x
Apply_by : 31 Dec '21
Apply : https://internshala.com/internship/detail/data-analytics-internship-in-bangalore-at-abg-it-services1637870917

=====

183)profile_name : Business Analytics
company_name : Minnano Consultancy
location_name : Work From Home
Joining : Immediately
Duration : 2 Months
stipend : 2000 /month
Apply_by : 31 Dec '21
Apply : https://internshala.com/internship/detail/business-analytics-work-from-home-job-internship-at-minnano-consultancy1637880950
company_link : http://www.minnanoconsultancy.com

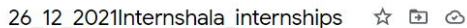
=====

184)profile_name : Mobile App Development
company_name : Uruvyacas Research And Technological Solutions
location_name : Guwahati, Silchar
Joining : Immediately
Duration : 2 Months
stipend : 30000-40000 /month
Apply_by : 31 Dec '21
Apply : https://internshala.com/internship/detail/mobile-app-development-internship-in-guwahati-silchar-at-uruvyacas-research-and-technological-solutions1637845099

=====

185)profile_name : Summer Research
company_name : Indraprastha Institute Of Information Technology, Delhi
location_name : Delhi
Joining : 6 May '21
Duration : 2 Months
stipend : 5000 /month
```

PyCharm 2020.2.5 available
Update...



File Edit View Insert Format Data Tools Extensions Help Last edit was seconds ago

 Share


A1	f_X	profile_name
----	-------	--------------

	A	B	C	D	E	F	G	H	
1	profile_name	company_name	location_name	Joining	Duration	stipend	Apply_by	Apply	company_link
2									
3	Machine Learning	Spritle Software	Chennai, Coimbatore, Madurai,	Immediately	4 Months	5000 /month	8 Jan' 22	https://internshala.com/internship	https://www.spritle.com
4									
5	Data Science	Fittlyf	Work From Home	Immediately	4 Months	3000 /month	8 Jan' 22	https://internshala.com/internship	https://fittlyf.com
6									
7	Data Science	Ellipsonic	Bangalore	Immediately	3 Months	5000 /month	8 Jan' 22	https://internshala.com/internship	http://www.ellipsonic.com
8									
9	Teaching Assistance (Data Science)	Pianalytix Edutech Private Limited	Work From Home	Immediately	2 Months	15000 lump sum	8 Jan' 22	https://internshala.com/internship/detail/teaching-assistance-data	
10									
11	Business Analytics	Guidona Softpedia Private Limited	Work From Home	Immediately	3 Months	5000 /month	8 Jan' 22	https://internshala.com/internship/detail/business-analytics-work	
12									
13	Customer Experience Design	Walkover Web Solutions Private Limited	Indore	Immediately	2 Months	4000-7000 /month	8 Jan' 22	https://internshala.com/internship	https://walkover.in/
14									
15	Equity Research Analysis	Value Ethics Private Limited	Pune	Immediately	3 Months	8000 /month	8 Jan' 22	https://internshala.com/internship/detail/equity-research-analysis	
16									
17	Financial Analysis	Artists Association Of India	Work From Home	Immediately	2 Months	Unpaid	8 Jan' 22	https://internshala.com/internship/detail/financial-analysis-work-f	
18									
19	Business Analytics	Sterling Technolabs	Work From Home	Immediately	3 Months	3000-6000 /month	8 Jan' 22	https://internshala.com/internship/detail/business-analytics-work	
20									
21	International Marketing	V Ganesh Agro	Indore	Immediately	3 Months	2000 /month	7 Jan' 22	https://internshala.com/internship	http://ganeshagro.com
22									
23	Business Analytics	Stirring Minds	Work From Home	Immediately	1 Month	5000-10000 /month	7 Jan' 22	https://internshala.com/internship	http://www.stirringminds.com
24									
25	Business Analytics	Growth Natives	Mohali	Immediately	6 Months	16000 /month	7 Jan' 22	https://internshala.com/internship	https://growthnatives.com
26									
27	Systems Applications & Products (SAP)	Geodrive Solutions Private Limited	Dehradun, Delhi, Kangra, Chandigarh	Immediately	6 Months	10000 /month	7 Jan' 22	https://internshala.com/internship	http://geodrive.in/
28									



26 12 2021Internshala internships ▼



- 
- So this way we see that the data are going to get is real, there is a lot of information in it , and we can store this data in our database and use it as we want.....

Is Web Scraping Legal in India?

- When it comes to whether web scraping legal or illegal, it is the biggest query people have about web scraping. However, most websites do not allow people to web scrap their websites. And why would they want to? They may not include this information on the home page, of course, but they do write about this in their Terms and Conditions section.
- There is no legal statement out there against web scraping, however, if they write about it on their website, they can file a case against you, this is why there are many web scraping legal cases. Although it varies from country to country.



Conclusion

- The main goal of this seminar was to explain how to use web scraping techniques to gather data from the web and display it in a meaningful way.
- By Scraping the web, we can store the data in a format and use it properly, Thus saving our searching time .
- Also you can use it as you wish, you can analyze it, you can preprocess that data and get a lot more information.

Future Scope

Web Scraping is becoming very popular currently and its popularity in future is surely going to increase due to the benefits it is providing to many companies.

Currently these are some areas where data scraping usage has grown drastically are:

- Marketing
- Finance
- SEO
- eCommerce
- Social media

Due to increasing competitive environment the usage of data scraping in future will increase gradually which will enhance the quality of services

References

- [1] https://en.wikipedia.org/wiki/Web_scraping
- [2] <https://webscraper.io/>
- [3] <https://realpython.com/beautiful-soup-web-scraper-python/>
- [4] <https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/>
- [5] <https://www.researchgate.net/publication/337545583> Web Scraping Wikipedia using Python and BeautifulSoup



Thank You

Please let me know if you have any questions