# 

**Python Web Scraping Documentation**

WEB SCRAPING INTRODUCTION:

In this post, you will learn about the Python web scraping for the Zaubacorp website.

The code for this chapter is here. <https://github.com/Prasadvada/Web\_scraping/tree/main/Zaubacorp>`\_

URL = <https://www.zaubacorp.com/company-list/p-1-company.html>  
  
From above URL we are extracting the data of Companies incorporated details in India

This document assumes you have already installed Python 3, and you have used both \*pip\* and \*venv\*.

Python is the most popular programming language for web scraping. It provides many libraries that can handle web crawler related process smoothly. Scrapy and Beautiful Soup are among the widely used libraries.  
  
Python Web Scraping Library:  
  
In this article, we have used the Requests and Beautiful Soup modules to develop web scraping in Python.  
  
  
Install requests module

The **requests** module makes a HTTP request to the specified web page using Python and returns a response object. This module is used with the Python web scraping module. The installation process of this package is similar to the other Python packages.

##### **pip install requests**

Beautiful soup module

The beautiful soup library makes it easy to scrape the information from the HTML or XML files. The Beautiful Soup4 or bs4 works on Python 3. It is much faster and supports third party parsers like html5lib and lxml. The following command installs the Beautiful Soup module using pip tool.  
  
 **pip install bs4**  
  
Make a GET request

The GET method returns data of the requested URL. Suppose we want to fetch the HTML page content of a job search portal using requests get method   
  
 Document Load  
  
 Parsing  
  
 Extraction  
  
 Transformation  
  
**Web Scraping** is a process of data extracting from web sites. The extracted data can be content, URLs, contact information, CIN, RoC, Status etc, which we can store in a local file or database. This process can be done manually by code called scrapper or by an automated software implemented using a boat or web crawler. The web scraping is not always legal. Some sites have dis-allowed the scraping in the '*robots.txt*' file. Some popular sites provide APIs to access their data in a structured way. But not all websites. So, we need a web scraper for data extraction, data mining and storing in a structured way.

we will Find the URL that you want to scrape.  
Suppose, we want to get all the Companies incorporated details in India

from the URL.  
The URL is automated and after getting the response from the request

The response is sent to the server to client to get the https response like GET and POST   
After we will find the total number of companies found and no of pages present and I am taking the total pages and getting the total pages URL and appending into the list

By using the Thread Pool Executor and max workers these will open multiple browsers at a time by using the map function and I created a data extraction function and all the URLs are iterating in a function and getting response of the page and extracting the required fields what we need to extract. These is used to save the  
  
  
  
We are taking the total number of rows and by using the loop we are iterating all the rows and by giving the required x-paths we are fetching the required data

At finally we are saving the data in to the Mongo dB