**Web Scraping Using Python**

**Problem Statement:**

Information commonly lies in silos in multiple places and websites. One such data source is Zubacorp, which provides information on multiple startups in India. The current task is to build a scraper and collect as much information as possible from these profiles.

For example,

● YNOS Venture Engine - [zaubacorp.com/company/YNOS-VENTURE-ENGINE-CC-PRIVATE-LIMITED/](https://www.zaubacorp.com/company/YNOS-VENTURE-ENGINE-CC-PRIVATE-LIMITED/U74999TN2017PTC115985)

● BusinessOnBot - [zaubacorp.com/company/BUSINESSONBOT-PRIVATE-LIMITED/](https://www.zaubacorp.com/company/BUSINESSONBOT-PRIVATE-LIMITED/U72900KA2020PTC136387)

These are profiles of a couple of startups in India. The scraper should also work for other startup profiles from the same website

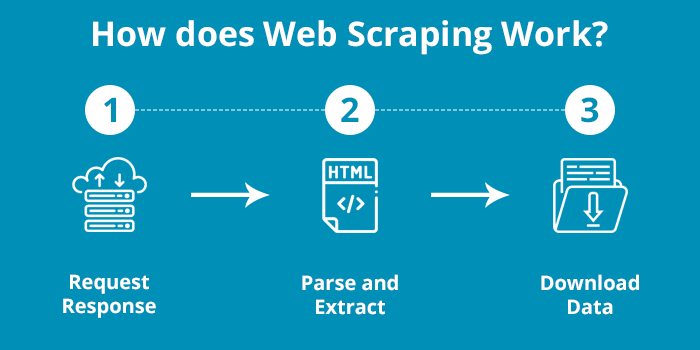
**Dependencies Required:**

1. Python
2. Beautiful Soup (bs4)
3. Html5lib
4. Requests
5. CSV

**Web Scraping:**

Web scraping is the process of using bots to extract content and data from a website.

Unlike screen scraping, which only copies pixels displayed onscreen, web scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate the entire website content elsewhere.

e

**Approach:**

Step 1: Setting up the Environment

Step 2: Get the HTML as a string

Step 3: Parse the HTML using Beautiful Soup (lxml, HTML parser)

Step 4: Parsing the data from the table and taking the values from rows and columns.

Step 5: Writing the obtained into a CSV file

Step 6: Done

**Conclusion:**

Using the Python Beautiful Soup as leverage, the HTML page is parsed, and the data is obtained. The required information is obtained from the tables present in the HTML. Parsing of HTML is done, and the data present in the tables are processed using the modules in python. The data is written into the CSV for various use cases.

Code can be refined to get the data in a format necessary for Machine Learning Algorithms.

**References:**

1. [zaubacorp.com/company/YNOS-VENTURE-ENGINE-CC-PRIVATE-LIMITED/](https://www.zaubacorp.com/company/YNOS-VENTURE-ENGINE-CC-PRIVATE-LIMITED/U74999TN2017PTC115985)
2. [zaubacorp.com/company/BUSINESSONBOT-PRIVATE-LIMITED/](https://www.zaubacorp.com/company/BUSINESSONBOT-PRIVATE-LIMITED/U72900KA2020PTC136387)
3. <https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/>
4. <https://www.dataquest.io/blog/web-scraping-python-using-beautiful-soup/>