**Taiyo assignment Details**

**-Sukkala Pravalika**

“*Thank you for giving me this opportunity to prove my skills and also from this assignment, I learned a lot of things. Hope I get to learn more and execute my skills in real life projects by working with you. As I managed to complete this assignment within in the given time, I hope I can manage the time for the future projects too. And I would like to be the part of the team taiyo.ai*”

-The assignment was about “web scraping” for the role ‘Python Data Engineering’.

-From the assignment pdf provided from the Intern Shala I completed the web scrapping of the China projects.

-From the website Of China projects there were a lot of links inside the website re-directing to the page of information of the projects.

Problem statement:

* From my understanding it is to Retrieve all the information of projects from the websites like project name, location, starting date, ending date, project id, handler name, details about the handler, project bids and also all the details present in the website. The clean data must be provided to examine this scraped data.

Perception:

* To retrieve the data from the website, first we need to inspect the html of the website.
* we have to observe all the parent and child tags, especially the links where the information of project is present.
* From there the actual procedure starts, going through the network and fetching all the xhr info, we need the GET link through the website.
* Now using that GET query link, form a code which gives all the information about the link.
* In the given time I scraped all the possible data from the link in the website and cleaned, geocoded, standardized the output data.

Process of Assignment completed:

Scraping

* First of all, it needed to scrape all the links from the website using requests, beautifulsoup, python.
* Using insomnia platform, I separated the information from the html network system collecting all the tags from the html using insomnia and visuoalcode studio
* And then I exported all the data output in a csv file which was clumsy and needed to be cleaned

Cleaning

* Using jupyter notebook, install all the required libraries like pandas, NumPy… Etc.
* And clean the data by dropping null values, duplicates values, and all the remaining unwanted data from the fiestresults.csv file.
* And export the cleaned data into cleaneddataodd.csv file

Geocoding

* Using geopandas, geopy separating the latitudes and longitudes.
* Then locating the actual coordinates on the map by using folium.

Standardization

* Using preprocessing. scale () function and StandardScaler() function by Python sklearn library standardization on the dataset and converted to a standardized format and stored the data in the csv file

Submission:

* All the required files that are to be evaluated are in the google drive link below.
* <https://drive.google.com/drive/folders/1WLd4zoY3SzGT2jlsVCkqxSZpT0v9fqUn?usp=share_link>

Conclusion:

From the task I completed, I hope I proved my skills through this assignment and I’m looking forward to the next recruitment process of taiyo.ai.