Phase - I:

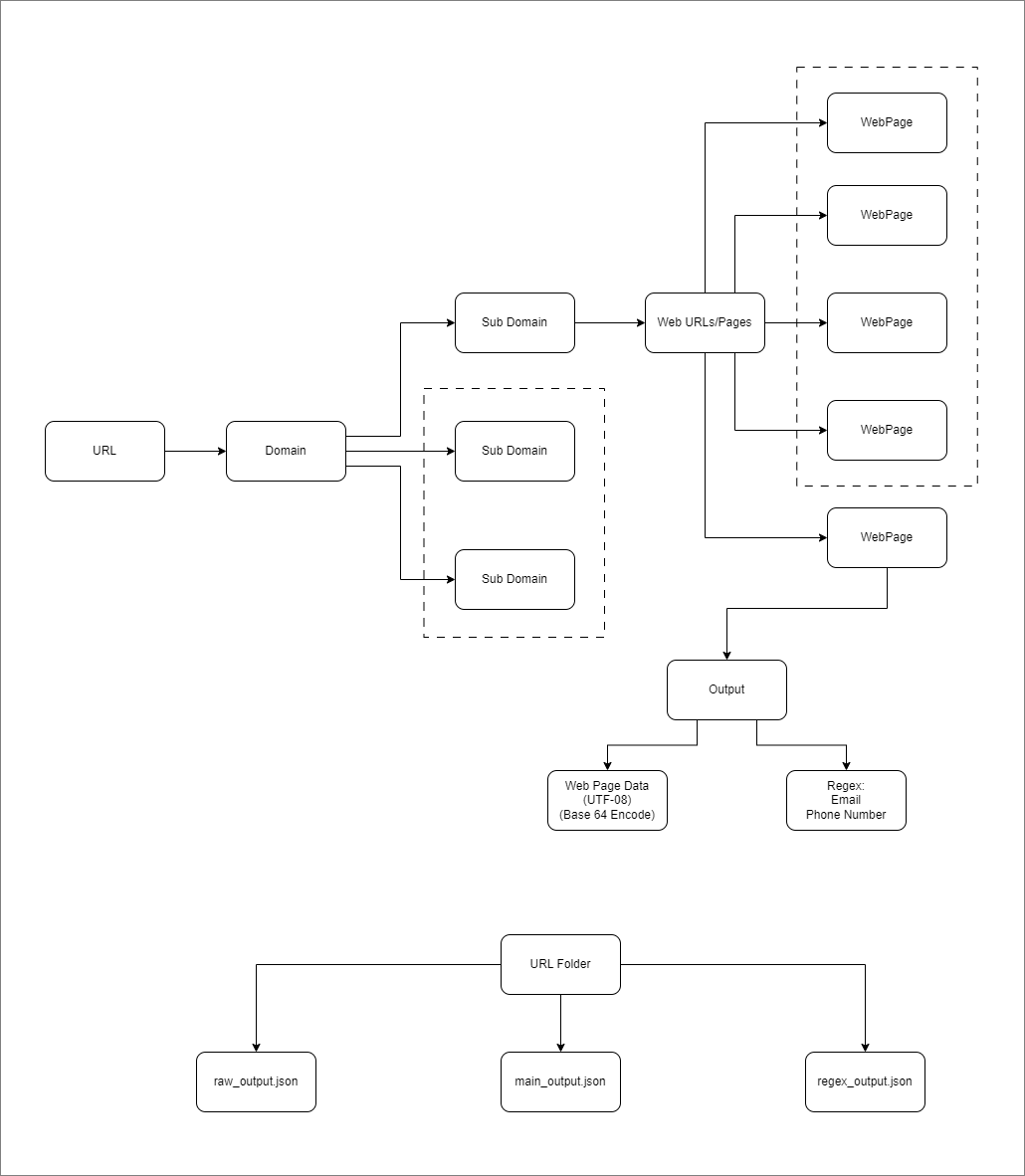
* List out target engineering college websites (Andhra/Telengana Engineering Colleges approved by JNTUK/H)
* Collected required data in csv format.

Phase - II:

* Design Data extraction process.

Tools selection:

* URL > Domain
* Domain > Subdomains
* Subdomain > Web Pages > Page Data > Regex > Email / Phone numbers



Phase - III: Output formatting (Strcture Data)

/URL\_Folder

raw\_output.json:   
{

"urlScanResult": {

"url": "string", // The URL that was scanned

"domain": "string", // The main domain extracted from the URL

"subdomains": [

{

"subdomain": "string", // The subdomain

"webPages": [

{

"page\_url": "string", // The URL of the web page

"webpageContent": "string", // Base64 encoded format.

"regexFilterOutput": "string" // The output after applying a regex filter on the web content

},

{

"page\_url": "string", // The URL of the web page

"webpageContent": "string", // Base64 encoded format.

"regexFilterOutput": "string" // The output after applying a regex filter on the web content

}

// ... more web pages

]

},

{

"subdomain": "string", // The subdomain

"webPages": [

{

"page\_url": "string", // The URL of the web page

"webpageContent": "string", // Base64 encoded format.

"regexFilterOutput": "string" // The output after applying a regex filter on the web content

},

{

"page\_url": "string", // The URL of the web page

"webpageContent": "string", // Base64 encoded format.

"regexFilterOutput": "string" // The output after applying a regex filter on the web content

}

// ... more web pages

]

}

// ... more subdomains

]

}

}

main\_output.json

{

"urlScanResult": {

"url": "string", // The URL that was scanned

"domain": "string", // The main domain extracted from the URL

"subdomains": [

{

"subdomain": "string", // The subdomain

"webPages": [

{

"url": "string", // The URL of the web page

"regexFilterOutput": {

"emailAddresses": ["string"], // Array of email addresses found in the web content

"phoneNumbers": ["string"] // Array of phone numbers found in the web content

}

},

// ... more web pages

]

},

// ... more subdomains

]

}

}

Regex\_output.json:

{

"urlScanResults": [

{

"url": "string", // The URL that was scanned

"domain": "string", // The main domain extracted from the URL

"combinedRegexOutput": {

"emailAddresses": ["string"], // Combined array of email addresses from all web pages under the URL

"phoneNumbers": ["string"] // Combined array of phone numbers from all web pages under the URL

}

}

// ... more URLs

]

}

Phase - IV:

Architecture Design:

Tool Selection:

1. URL > Domain
   1. Python function
2. Domain > Subdomain (Include main website also as an one subdomain)
   1. Identify subdomains (tool name)
3. Subdomain > Web URLs
   1. List out all Web URLs
4. WebURL > Data Extraction
   1. Extract Web Data (Python)
   2. Convert it to readable format (UTF format)
5. WebPage Content > Regex
   1. Regex Apply (Email / Phone number data)
6. WebPage\_content > Base64 Encode
   1. Python function

Radha:

* Keep update all relevant tools and to get final output format.