Team Name : **Runtimeterror**

Problem Statement : **Solution for Crowd Sourcing of Records**

Brief about the Idea:

• Data crowdsourcing from the internet using phone numbers or email addresses involves two separate processes that must be automated and combined.
• The first step is to search the internet for information related to the email address or phone number in public domains, and then scrape those public domains with web scrapers enabled with proxy chains to obtain detailed information.
• To collect information related to email addresses and phone numbers, we use tools such as PhoneInfoga and UserRecon. We will use the web scrapper tool to scrape the web pages containing information relating to the given phone number or email address after collecting the list of domains and URLs.
• In most cases, social media websites will block IP addresses after a few attempts at web scraping. Proxy chains are a possible solution to this problem.
• Data scraped from these sources will be analyzed and processed to make the information more understandable.

Opportunity :

How different is it from any other existing ideas out there? How will it be able to solve the problem?

• Existing crowd-sourcing tools fail to provide complete information about the phone number/email entered. In this case, they are just providing information regarding the usage of a specific email address in a domain.

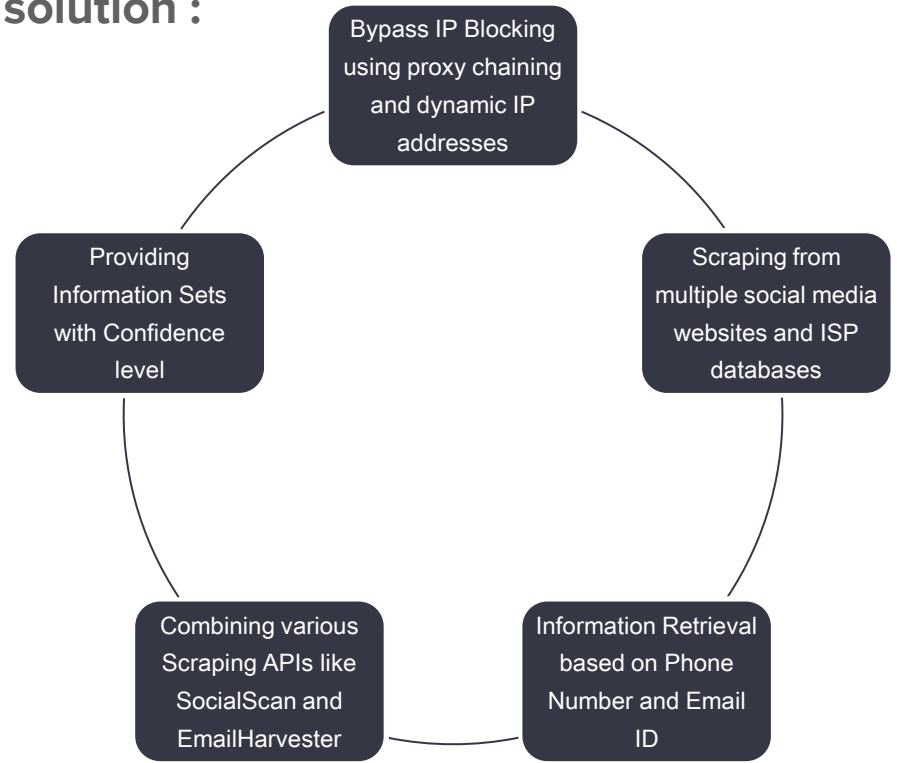
• Our development project will offer one-stop access to all internet information available in public domains. In this approach, both the presence of an account is detected as well as the public information associated with it is extracted. There also exists the problem of websites blocking IP addresses of the scrappers who scrape heavy loads of publicly available data from the sites. This can be solved by making use proxy chain technique to extract complete information.

• There is also a lack of a universal web scraping service for scraping different types of websites. The application that we develop will solve this problem and provide a convenient and time-efficient solution for extracting information and analysis of extracted information.

List of features offered by the solution :

All the features listed in the figure will be ***integrated into a single application***, which will be an one stop solution for the crowd sourcing of the data from publicly available domains of the internet.

The ***user interface*** of the application will make it easy and convinient for the usage of tool to gather all the required information ***with a single click***



Business Logic of the solution :

- The business logic for the solution is to filter out relevant information from all the information available on the internet by writing our algorithm to identify the relevant information for the entered mobile number or email.
- The filtering will be done based on the websites, section of the profile, relevancy, authenticity, etc.
- For example, two people might have email addresses with slight tweaks in the address, so there is a high chance that the information from other emails is also extracted by the web scraper. Therefore once the information is extracted, it has to be processed with help of an algorithm before presenting it.
- We can label the information obtained with a confidence interval, and only the information above certain confidence value will be taken into consideration for further processing.

Technology used :

1. We use **python** and **shell scripting** to integrate existing tool and our novel module for data acquisitoin and data processing.
2. We make use of tools such as **PhoneInfoga** which is one of the most advanced tools to scan international phone numbers. It allows you to first gather standard information such as country, area, carrier, and line type on any international phone number, then searches for footprints on search engines to try to find the VoIP provider or identify the owner
3. We can also make use of tools such as **Emailharvester** and **SocialScan** to identify the domains in which a given email address is being used .
4. The **UserRecon** tool is used to find usernames across over several social networks. It is very useful when you are running an investigation to determine the usage of the same username across different social media platforms such as Twitter, Instagram, MySpace, Youtube, Reddit, WordPress, GitHub, and many more.When we have the links by using the above tools, the pages can be scrapped using web scrapper tools such as octaparse , scrapy etc.
5. However scrappers may get blocked from any of the websites. Hence we can use **proxy chains** on the browsers that redirects connections to periodically assign dynamic IP addresses. ProxyChains can string multiple proxies together to make it harder to identify the original IP address Thus making enabling efficient scraping

The **PhoneInfoga** can be invoked automatically using **python shell scripting** and the result of the PhoneInfoga, which is the list of all the ***websites with the digital footprint*** of the entered number will be redirected to an intermediatery file and then this output will be further processed by the **webscraper (BeautifulSoup module)** to gather all the required data from the website with the presence of that phone number.
Similarly email address can also be used to crowdsource the publicly available data from the internet by integrating different open-source tools like **Emailharvester** and **SocialScan** with the inhouse module developed to ***create an application*** for data gathering from all the available sources.

```shell
$ phoneinfoga scan -n +4176418xxxx

Results for googlesearch
Social media:
    URL: https://www.google.com/search?q=site%3Afacebook.com+intext%3A%224176418xxxx%22+OR+intext%3A%22%2B4176418xxxx%22+OR+intext%3A

    URL: https://www.google.com/search?q=site%3Atwitter.com+intext%3A%224176418xxxx%22+OR+intext%3A%22%2B4176418xxxx%22+OR+intext%3A%

    URL: https://www.google.com/search?q=site%3Alinkedin.com+intext%3A%224176418xxxx%22+OR+intext%3A%22%2B4176418xxxx%22+OR+intext%3A

    URL: https://www.google.com/search?q=site%3Ainstagram.com+intext%3A%224176418xxxx%22+OR+intext%3A%22%2B4176418xxxx%22+OR+intext%3

    URL: https://www.google.com/search?q=site%3Avk.com+intext%3A%224176418xxxx%22+OR+intext%3A%22%2B4176418xxxx%22+OR+intext%3A%22076
Disposable providers:
    URL: https://www.google.com/search?q=site%3Ahs3x.com+intext%3A%224176418xxxx%22

    URL: https://www.google.com/search?q=site%3Areceive-sms-now.com+intext%3A%224176418xxxx%22+OR+intext%3A%22076418xxxx%22

    URL: https://www.google.com/search?q=site%3Asmslisten.com+intext%3A%224176418xxxx%22+OR+intext%3A%22076418xxxx%22
```

The figure shows the output of PhoneInfoga which is going to be further processed by the python code to extract data.

Estimated cost of/after implementing the solution :

• The application makes use of open-source libraries and tools, hence it requires absolutely **zero investment** for development (Developer salary is exempted).
• As discussed in the technology slide, the entire project will be python based and will also make use of open-source libraries which are **free** to use. This application doesn't require hosting any online platform, therefore there is no cost of purchasing a domain or a database as it does not require any huge amount of storage. It directly scrapes data from the internet each time and processes it in the local system.
• In case of high usage of the tool, there might be a requirement of a cloud platform for processing the information which might cost about Rs.1 Lakh in AWS platform (it is free for the period of first 12 months).

THANK YOU