# Abstract

In this project, I worked on the Government Services Portal Of India which is developed with an objective to enable a single window access to services being provided by the various Indian Government entities. This Portal is an initiative under the India Portal Project. It is designed and developed by National Informatics Centre (NIC), Ministry of Electronics & Information Technology, Government Of India.

The project aims to list all the information available on the portal and give a short summary about every service that the Government of India is providing the community.

The proposed system for this project is a web scraper program that is able to access and extract data

from the Government Services Portal Of India. The extracted data is stored in an excel file.

Web Scraping is the process through which we extract data from a website, and save it in a form which is easy to read, to understand and to work on. Web Scraping also known as web extraction or harvesting is a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis. Commonly, web data is scrapped utilizing Hypertext Transfer Protocol (HTTP) or through a web browser.

### Aims and Objectives

This project sets out to conduct research in the area of web scraping and how it can be used as a tool for listing all the available services that government is providing the citizens of India. The main focus of this project is to give a short summary of every service listed on the portal.

The core aims of this project are :

1. To build an understanding of the various services that are being provided by the government.
2. To create a Web Scraping Program to extract the important information available on the portal about a particular service.
3. To log the findings in an Excel File for the further manipulation.

The Deliverables of this project are:

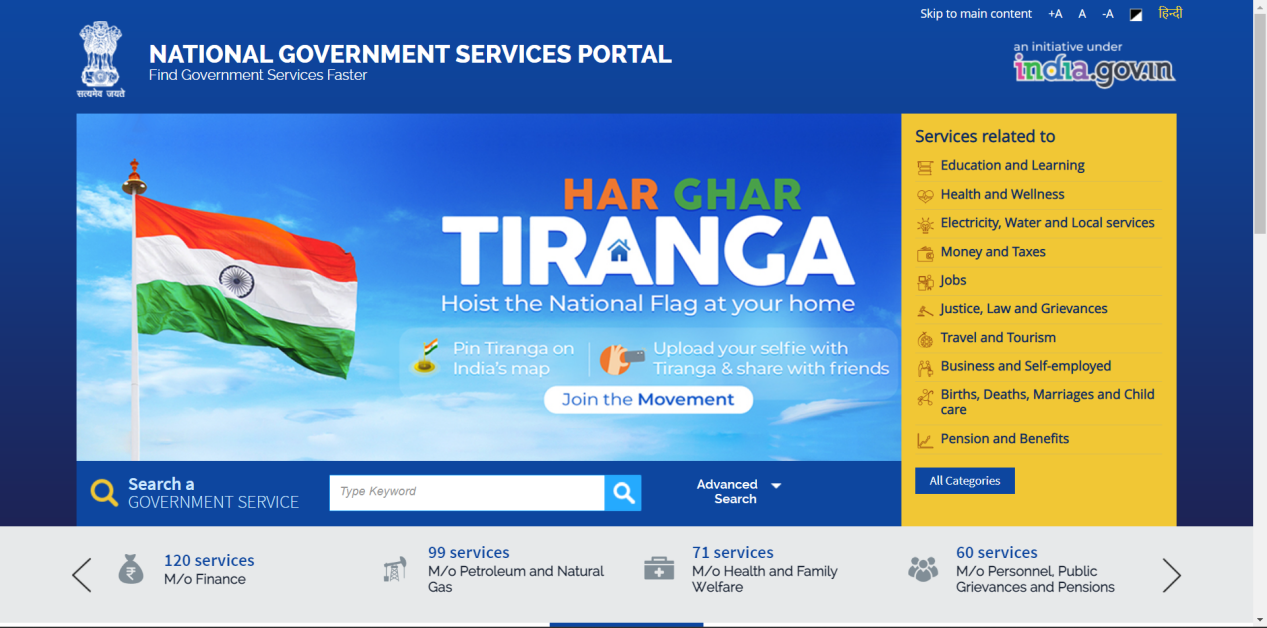
1. Web Scraping Python Program
2. Excel Database
3. Final Report

#### About the Website

Many Government entities at the Centre, State, District and Local Levels are providing online services that have made the life of the citizens simpler and have also increased transparency and efficiency. These services are provided through multiple websites.

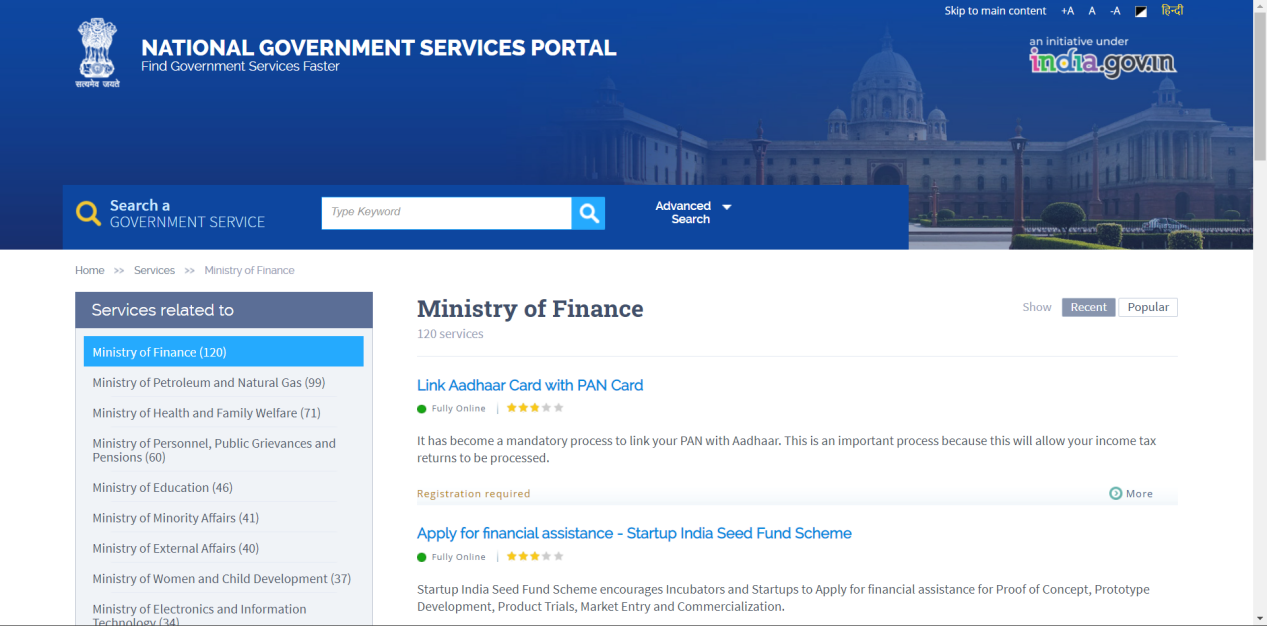
In order to list these services in a well categorised and searchable interface, National Government Services Portal (https://services.india.gov.in) has been developed under the ambit of India POrtal project which is being executed by NIC (National Informatics Centre).

The purpose of this portal is to facilitate the listing of online services provided by various government entities under one platform and ensuring standardization with respect to content architecture and classification of services.



The above picture shows the homepage of the Government Services Portal Of India.

All the services are grouped on the basis of the ministry which is providing that particular service.



This picture clearly shows the basic information about a particular service. For eg, the service “Link Aadhaar Card with PAN Card” comes under the Ministry of Finance.

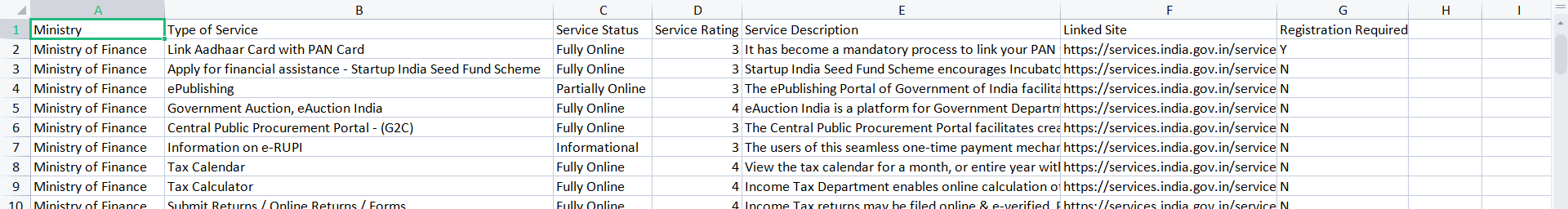
Now, a particular service has the following attributes :

1. The Name of the Service
2. The Status of the Service
3. The Short Description of the service
4. The Rating of the Service
5. The Linked URL associated with that service
6. Whether Registration is required for that service.

#### About the Program

Now, the project is a python program which aims to extract all the information mentioned above in an Excel Database which can be further manipulated.

The Data extracted will be shown in the following format.



The above picture shows a part of the Excel file which is generated by the python program.

For the Creation of the web scraper program, there are certain tools that are utilized. Python 3.10.5 will be used, with beautifulsoup4 4.11.1 for the web scraper, and requests 2.28.1.

The Hardware Requirements of this project are:

1. A PC that is connected to the internet.

The Software Requirements of this project are:

1. Python 3.10.5, beautifulsoup4 4.11.1 and requests 2.28.1

#### Implementation

To store the data I will be using a spreadsheet which will be filled using the python csv package.

urls = []

url = 'https://services.india.gov.in/service/ministry\_services?ln=en&cmd\_id=1126'

page = requests.get(url)

soup = BeautifulSoup(page.content, 'html5lib')

ministries = soup.find('nav', *class\_*='naver').find('ul').find\_all('li')

*for* ministry *in* ministries:

    anchor = ministry.find('a')

    Link = anchor['href']

    urls += [Link]

To Get all the links which lists the services on the portal the above code is used. It creates a list of all the urls of the pages on which the services are listed.

*with* open('servicesDesc.csv', 'w', *newline*='', *encoding*='utf8') *as* f:

    thewriter = writer(f)

    header = ['Ministry', 'Type of Service', 'Service Status', 'Service Rating', 'Service Description', 'Linked Site', 'Registration Required']

    thewriter.writerow(header)

This creates a csv file named ‘servicesDesc.csv’ and gives it a header for the format of data which is to be extracted.

*for* url *in* urls:

*for* i *in* range(12):

*if* (i != 0):

            url = url + '&page\_no=' + str(i+1)

        page = requests.get(url)

*if*(page.content):

            soup = BeautifulSoup(page.content, 'html5lib')

            services = soup.find\_all('div', *class\_*='edu-lern-con')

            category = soup.find('div', *class\_*='title-left').find('h2')

            categoryText = category.text.strip()

*with* open('servicesDesc.csv', 'a', *newline*='', *encoding*='utf8') *as* f:

                thewriter = writer(f)

*for* service *in* services:

                    anchor = service.find('a', *class\_*='ext-link')

                    status = service.find('li', *class\_*='status\_icon')

                    rating = service.find('span', *class\_*='star\_rating\_stars')

                    ratingNo = rating['style'].lstrip('width:').rstrip('%;')

                    ratingNo = int(ratingNo)

                    ratingFinal = ratingCalc(ratingNo)

                    desc = service.find('p')

                    desc = desc.text.strip()

                    Link = anchor['href']

                    RegReq = 'N'

*if*(service.find('li', *class\_*='tag\_light')):

                        RegReq = 'Y'

                    Row = [category.text.strip(), anchor.text.strip(), status.text.strip(), ratingFinal, desc, Link.strip(), RegReq]

                    thewriter.writerow(Row)

*else*:

*continue*

f.close();

The code above traverse through the list of urls and for every url it fetches the necessary information of every service available on that particular page and stores it in the database. If there are multiple pages for services providing by a single minstry then it adds the neccassary changes in the url.

def ratingCalc(*x*):

*if*(*x* == 0):

*return* 0

*elif* (*x* > 0 and *x* <= 25):

*return* 1

*elif* (*x* > 25 and *x* <= 45):

*return* 2

*elif* (*x* > 45 and *x* <= 65):

*return* 3

*elif* (*x* > 65 and *x* <= 85):

*return* 4

*elif* (*x* > 85 and *x* <= 100):

*return* 5

*else*:

*return*

The function above named ratingCalc returns the rating of a particular service in a number format.

*import* requests

*from* bs4 *import* BeautifulSoup

*from* csv *import* writer

The python packages mentioned above are used in this program.

#### About the Python Packages

###### Requests

The requests library is the standard for making HTTP requests in Python. It abstracts for complexities of making requests behind a beautiful, simple API so that you can focus on interacting with services and consuming data in your application.

###### Beautifulsoup4 or bs4

The bs4 stands for BeautifulSoup version 4. It is a Python library which is used for pulling out data of the HTML and XML files using Python Program. The BeautifulSoup library was created basically for the purpose web scraping. The BeautifulSoup library works with a parser which then provides a way to us for navigating, modifying and searching the parsed tree in the web scraping process.

#### Code

*import* requests

*from* bs4 *import* BeautifulSoup

*from* csv *import* writer

def ratingCalc(*x*):

*if*(*x* == 0):

*return* 0

*elif* (*x* > 0 and *x* <= 25):

*return* 1

*elif* (*x* > 25 and *x* <= 45):

*return* 2

*elif* (*x* > 45 and *x* <= 65):

*return* 3

*elif* (*x* > 65 and *x* <= 85):

*return* 4

*elif* (*x* > 85 and *x* <= 100):

*return* 5

*else*:

*return*

urls = []

url = 'https://services.india.gov.in/service/ministry\_services?ln=en&cmd\_id=1126'

page = requests.get(url)

soup = BeautifulSoup(page.content, 'html5lib')

ministries = soup.find('nav', *class\_*='naver').find('ul').find\_all('li')

*for* ministry *in* ministries:

    anchor = ministry.find('a')

    Link = anchor['href']

    urls += [Link]

*with* open('servicesDesc.csv', 'w', *newline*='', *encoding*='utf8') *as* f:

    thewriter = writer(f)

    header = ['Ministry', 'Type of Service', 'Service Status', 'Service Rating', 'Service Description', 'Linked Site', 'Registration Required']

    thewriter.writerow(header)

*for* url *in* urls:

*for* i *in* range(12):

*if* (i != 0):

            url = url + '&page\_no=' + str(i+1)

        page = requests.get(url)

*if*(page.content):

            soup = BeautifulSoup(page.content, 'html5lib')

            services = soup.find\_all('div', *class\_*='edu-lern-con')

            category = soup.find('div', *class\_*='title-left').find('h2')

            categoryText = category.text.strip()

*with* open('servicesDesc.csv', 'a', *newline*='', *encoding*='utf8') *as* f:

                thewriter = writer(f)

*for* service *in* services:

                    anchor = service.find('a', *class\_*='ext-link')

                    status = service.find('li', *class\_*='status\_icon')

                    rating = service.find('span', *class\_*='star\_rating\_stars')

                    ratingNo = rating['style'].lstrip('width:').rstrip('%;')

                    ratingNo = int(ratingNo)

                    ratingFinal = ratingCalc(ratingNo)

                    desc = service.find('p')

                    desc = desc.text.strip()

                    Link = anchor['href']

                    RegReq = 'N'

*if*(service.find('li', *class\_*='tag\_light')):

                        RegReq = 'Y'

                    Row = [category.text.strip(), anchor.text.strip(), status.text.strip(), ratingFinal, desc, Link.strip(), RegReq]

                    thewriter.writerow(Row)

*else*:

*continue*

f.close();