[WEB SCRAPING]

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology/Master of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

SRAVYA.CH	**(AP21110010977)**
NAVYA.K	**(AP21110010992)**
MANOHAR.K	**(AP21110010994)**
RENUKA.K	**(AP21110011250)**

Under the Guidance of

(DR. POONAM YADAV)

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[DECEMBER,2022]

Certificate

Date: 15-Dec-22

This is to certify that the work present in this Project entitled "**WEB SCRAPING**" has been carried out by [**Ch.Sravya, K.Navya, K.Manohar, K.Renuka]**under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

Supervisor

(Signature)

Prof. / Dr. Poonam Yadav

Designation,

Affiliation.

Acknowledgements

We Sravya, Navya, Manohar, Renuka would want to convey my heartfelt gratitude to Prof. Poonam Yadav, my mentor, for her invaluable advice and assistance in completing our project. She was there to help us in every step of the project, and her motivation helped us to accomplish our task effectively. We would also like to thank all of the other supporting personnel who assisted us by supplying the information and knowledge that was essential, without which we would not have been able to complete this project.

We would also want to thank SRM University for accepting our project in our desired field of expertise. We'd also like to thank our friends and parents for their constant support and encouragement as we worked on this project.

Table of Contents

Abstract

Main objective of our Web Scraping is to extract information of a type of job from one or many websites used for searching jobs and process it into simple structures such as spreadsheets, database or CSV file. web scraping extracts HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.

Abbreviations

CSV	Comma Separated Values
https	Hypertext transfer protocol secure
www	World Wide Web
UTF-8	8-bit Unicode transformation format
Writerow	used to enter single row data into CSV
Writerows	used to enter multiple rows of data into CSV

Introduction

Our main motive in this project is to scrap data from some job searching websites like Flexjobs and creating a website for job searching. In today's time of data science & engineering, it is entirely expected to gather information from sites for examination purposes, job purposes. The main reason to choose this topic is to bring different job availability details into one frame as the websites mentioned above has different prioritized job availability information but we can make a place where a person who is searching for a job can get in touch with all the available job offers being offered, it is mostly useful for the people who are looking for a job but not a particular one and that can be done by their capabilities. In this project we are going to use Beautiful soup library for scraping and parsing. First, we are going to collect the data from different job sites using Beautiful soup library and then we will store that data and using that we will create our own job website.

1.1 AIM AND IMPORTANCE

- This project is used to extract required data of work from home jobs

- Give the collected data in a csv file format, data includes name of the job,

place from where job have been provided ,job description, age of the job post and job url

- url of every job is given so that if the user is interested in job he can visit job

slide using that

1.1.1 NEED OF THE PROJECT

This project is mainly useful for people who is searching for a specific type of job ,here we are scraping work from home jobs details from flexijobs so that it is useful for people who are looking for work from home jobs if developed further we can scrape data of workfrom home from different job searching website so that searching for that specific job will be easy.

Methodology

We have used concepts like Web-Scrapping and CSV file handling for extraction and storing of data.

2.1 Extraction of Data

We have used Web-Scrapping for extraction of job . Web-Scrapping is a concept where we analyze the source html of a website and then try to find our desired information using the html tags and their attributes. We have used a module called beautiful soup to perform the web-scrapping task. This module simply first converts the data into html format then takes suitable arguments of the type of tag to find along with the attributes to match. Then it returns the tag or tags that match the conditions and then we work with those tags to extract the information.

2.2 Sorting Data

The data of the user as well as the data of the product is stored in CSV format files to make it simple, easy to access and analyse. CSV format is file format where the data is stored in columns and the columns are separated using a comma (,) delimiter.

Discussion

3.1 Modules

A number of modules have been used in this project. Their uses and short description are given below.

3.1.1 Beautiful Soup

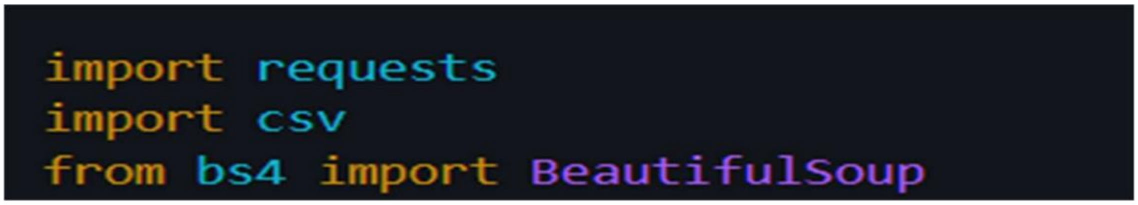
It is module used for web-scraping using tags and their attribute values. In this project we have used this module to extract the name of the product to track, the price of the product and the rating of the product from time to time.

3.1.2 Requests

This module is used to retrieve the source html of the webpage of the URL. This module provides the source harm in bytes format to the beautiful soup module for further processing.

3.1.3 csv

This is an inbuilt module in python for handling files. In this project we have used this module to write data into the csv files.

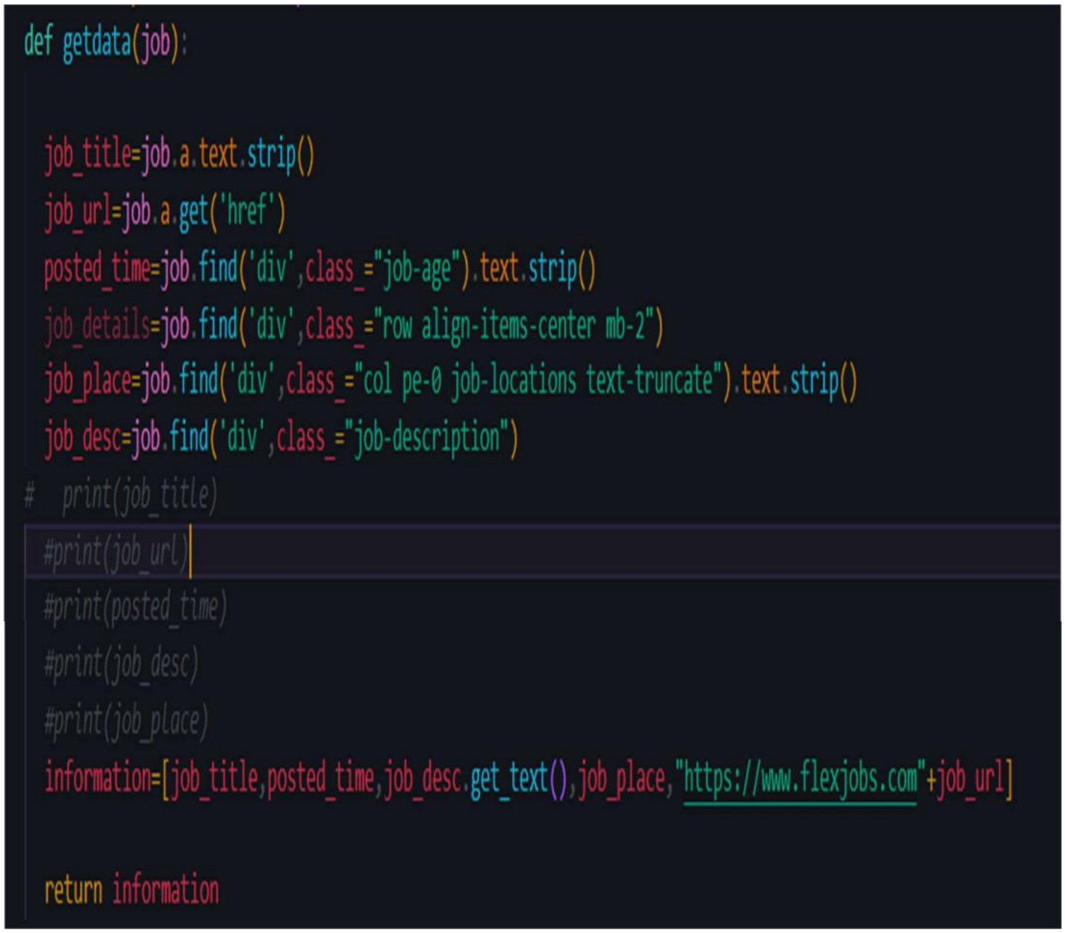


```
import requests
import csv
from bs4 import BeautifulSoup
```

3.2 Program Execution Flow

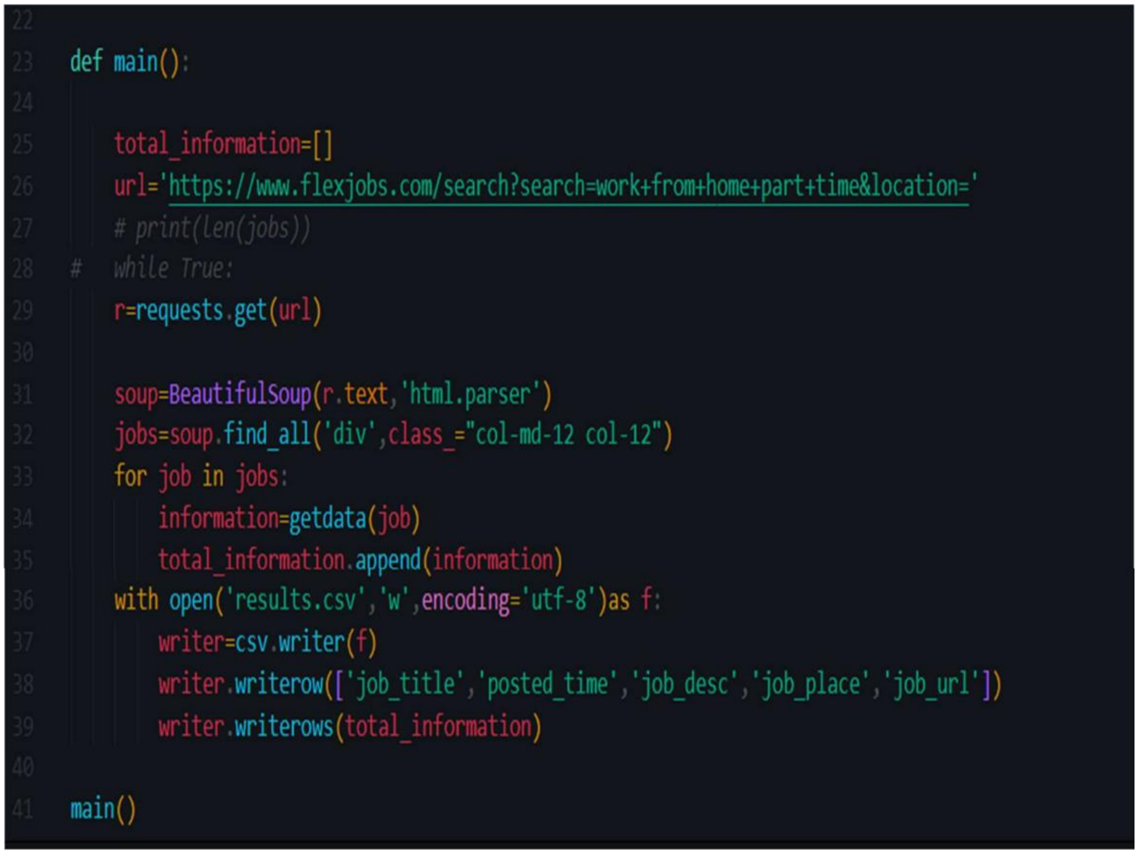
3.2.1 Sign-up Block

Code Block :



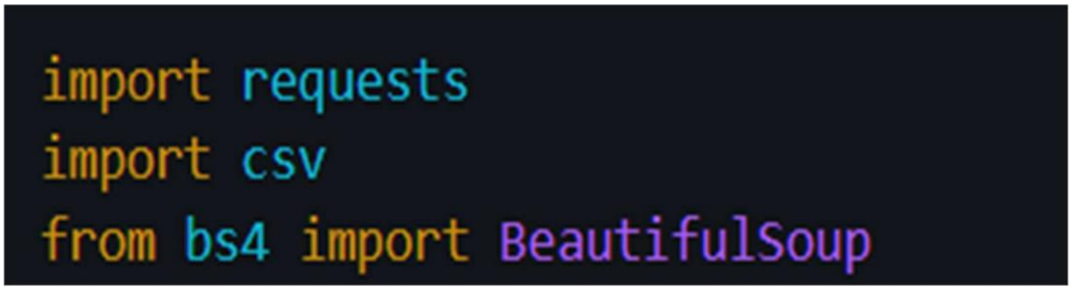
```
def getdata(job):

  job_title=job.a.text.strip()
  job_url=job.a.get('href')
  posted_time=job.find('div',class_="job-age").text.strip()
  job_details=job.find('div',class_="row align-items-center mb-2")
  job_place=job.find('div',class_="col pe-0 job-locations text-truncate").text.strip()
  job_desc=job.find('div',class_="job-description")
#   print(job_title)
  #print(job_url)
  #print(posted_time)
  #print(job_desc)
  #print(job_place)
  information=[job_title,posted_time,job_desc.get_text(),job_place,"https://www.flexjobs.com"+job_url]

  return information
```

3.2.2 Login Block
Code Block:

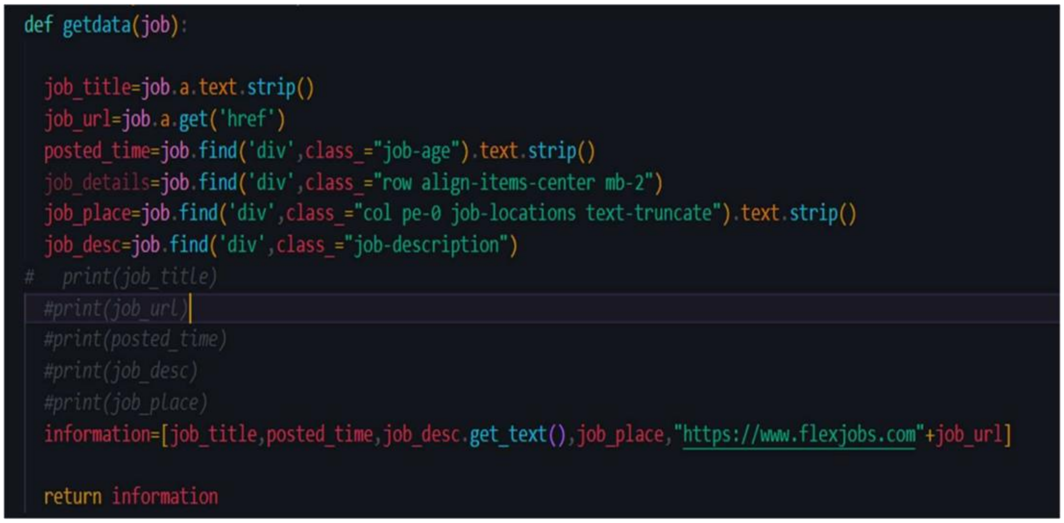


```python

def main():

    total_information=[]
    url='https://www.flexjobs.com/search?search=work+from+home+part+time&location='
    # print(len(jobs))
#   while True:
    r=requests.get(url)

    soup=BeautifulSoup(r.text,'html.parser')
    jobs=soup.find_all('div',class_="col-md-12 col-12")
    for job in jobs:
        information=getdata(job)
        total_information.append(information)
    with open('results.csv','w',encoding='utf-8')as f:
        writer=csv.writer(f)
        writer.writerow(['job_title','posted_time','job_desc','job_place','job_url'])
        writer.writerows(total_information)

main()
```

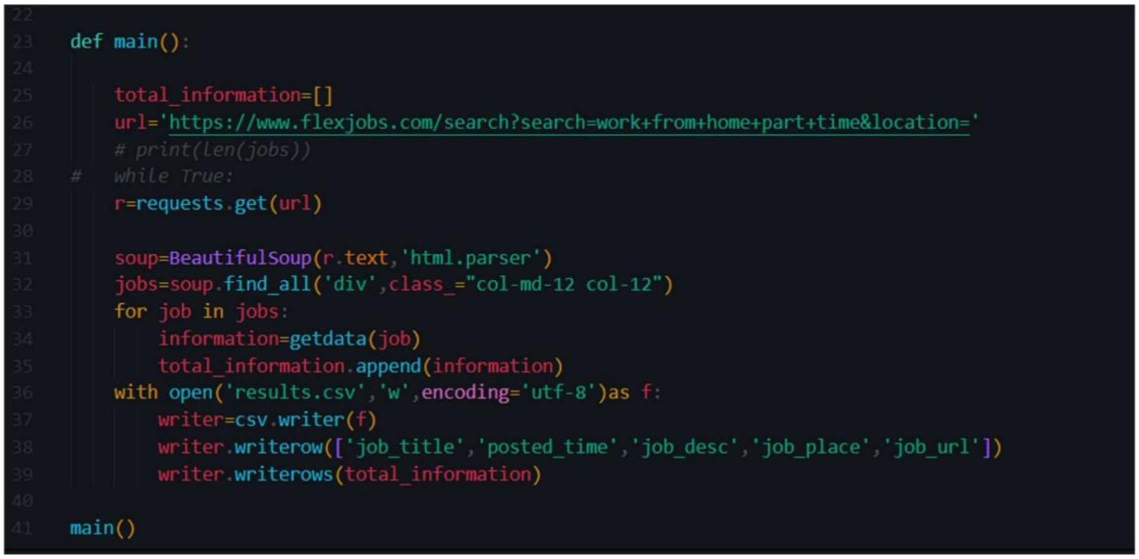
Source code:



```
import requests
import csv
from bs4 import BeautifulSoup
```



```
def getdata(job):

  job_title=job.a.text.strip()
  job_url=job.a.get('href')
  posted_time=job.find('div',class_="job-age").text.strip()
  job_details=job.find('div',class_="row align-items-center mb-2")
  job_place=job.find('div',class_="col pe-0 job-locations text-truncate").text.strip()
  job_desc=job.find('div',class_="job-description")
#   print(job_title)
  #print(job_url)
  #print(posted_time)
  #print(job_desc)
  #print(job_place)
  information=[job_title,posted_time,job_desc.get_text(),job_place,"https://www.flexjobs.com"+job_url]

  return information
```



```python

def main():

    total_information=[]
    url='https://www.flexjobs.com/search?search=work+from+home+part+time&location='
    # print(len(jobs))
#   while True:
    r=requests.get(url)

    soup=BeautifulSoup(r.text,'html.parser')
    jobs=soup.find_all('div',class_="col-md-12 col-12")
    for job in jobs:
        information=getdata(job)
        total_information.append(information)
    with open('results.csv','w',encoding='utf-8')as f:
        writer=csv.writer(f)
        writer.writerow(['job_title','posted_time','job_desc','job_place','job_url'])
        writer.writerows(total_information)

main()
```

4. Concluding Remarks

We did web scaping of the flexjobs website to collect work from home jobs which helps one to search that particular job without going through many scraped data will be converted into csv format where link of every work from home is provided.

5. Future Work

In the near future, Web scraping particular type of job from different jobs searching website will be of great use as the user can get particular information about the job they want and no need to go through all other type of jobs. It will have great demand in the future as people prefer website which has all jobs regarding type of job they are looking for than a website which contains multiple types of jobs.

References

1. Library Carpentry is another Software Carpentry spinoff focused on software skills for librarians.
2. W3School
3. Tutorials Point Python
4. Getting Started Image Source
5. wikipedia