Sabudh Foundation 8 Weeks Internship Programme

Submitted By : Kartika , BTECH - CSE , GNDEC, Ludhiana

Web Scraping :

Web scraping is an automated method used to extract large amounts of data from websites. The data on the websites are unstructured. Web scraping helps collect these unstructured data and store it in a structured form. There are different ways to scrape websites such as online Services, APIs or writing your own code. In this notebook, we'll see how to implement web scraping with python.

Web Scraping On Job Finder Website Like https://www.timesjobs.com/

We can scrape a website after checking if there are no explicit rules about scrapping . So I had scrapped https://www.timesjobs.com/ website
We can scrape this website to extract information regarding particular job .All Jobs related to python were scraped from this website by me from the link https://www.timesjobs.com/candidate/job-search.html?searchType=personalizedSearch&from=submit&txtKeywords=python&txtLocation=

The website was scrapped using BeautifulSoup in Python.To extract data using web scraping with python, you need to follow these basic steps:

1) Find the URL that you want to scrape
2) Inspecting the Page
3) Find the data you want to extract
4) Write the code
5) Run the code and extract the data
6) Store the data in the required format

We can download the page using **requests** in Python . The requests library will make a GET request to a web server, which will download the HTML contents of a given web page for us.

For example :

```
html_text=requests.get('https://www.timesjobs.com/candidate/job-search.html?searchType=personalizedSearch&from=submit&txtKeywords=python&txtLocation=')
html_text
```

The above commands will return status code 200 ,which means text is successfully downloaded

```
In [5]: html_text=requests.get('https://www.timesjobs.com/candidate/job-search.html?searchType=personalizedSearch&from=subm
        html_text

Out[5]: <Response [200]>
```

To extract the text from this site we will use **text** functionality

html_text=requests.get('https://www.timesjobs.com/candidate/job-search.html?searchType=personalizedSearch&from=submit&txtKeywords=python&txtLocation=').text

We will create **soup instance of BeautifulSoup ,**and we will use **lxml parser** for parsing because default HTML Parser doesn't work well with broken HTML Code

soup=BeautifulSoup(html_text,'lxml')

The following is the code to Scrap all jobs related to Python from the website

```
jobs=soup.find_all('li',class_="clearfix job-bx wht-shd-bx")
for job in jobs:
    published_date=job.find('span',class_="sim-posted").text
    if 'few' in published_date:
        company_name=job.find('h3',class_="joblist-comp-name").text.replace('  ','')
        skills=job.find('span',class_="srp-skills").text.replace('  ','')
        more_info=job.header.h2.a['href']
        print(f"Company Name :{company_name.strip()}")
        print(f"Required Skills:{skills.strip()}")
        print(f"More Info: {more_info}")
        print()
```

In this Code using various functionalities of Beautiful Soup ,Information related to all jobs related to Python is extracted .
Using find_all on soup instances ,all information with class_name=”clearfix job-bx wht-shd-bx” is stored in **jobs.** Then the company name,Required Skills and more info(url related to job) is scrapped using various HTML and CSS selectors.

Only those jobs are displayed in which the Published Date is “Posted a few Days ago “ . We stored the value in the published date and only if the Published Date contains “few” ,then only we will display the job post , if condition is used for that.

Output of the above code

```
Company Name :Surya Informatics Solutions Pvt. Ltd.
Required Skills:python,web technologies,linux,mobile,mysql,angularjs,javascript
More Info: https://www.timesjobs.com/job-detail/python-surya-informatics-solutions-pvt-ltd-chennai-0-to-3-yrs-jobid-UVlLes58wutzpSvf__PLUS__uAgZw==&source=srp

Company Name :Pure Tech Codex Private Limited
Required Skills:rest,python,database,django,debugging,mongodb
More Info: https://www.timesjobs.com/job-detail/python-pure-tech-codex-private-limited-pune-2-to-3-yrs-jobid-OHwfF0d6EhNzpSvf__PLUS__uAgZw==&source=srp

Company Name :GEMINI SOFTWARE SOLUTIONS
Required Skills:python,mobile,svn,nosql,python scripting,git,sql database
More Info: https://www.timesjobs.com/job-detail/qa-python-python-sdet-gemini-software-solutions-gurgaon-4-to-7-yrs-jobid-jsOuZLK8chlzpSvf__PLUS__uAgZw==&source=srp

Company Name :Gemini Solutions
Required Skills:python,mobile,svn,nosql,python scripting,git,api,sql database
More Info: https://www.timesjobs.com/job-detail/qa-python-python-sdet-gemini-solutions-gurgaon-4-to-7-yrs-jobid-eGMLzwOk2QlzpSvf__PLUS__uAgZw==&source=srp

Company Name :WHITE FORCE
Required Skills:python,mobile,debugging
More Info: https://www.timesjobs.com/job-detail/python-developer-white-force-chennai-2-to-4-yrs-jobid-hLnxtF76yRtzpSvf__PLUS__uAgZw==&source=srp

Company Name :Angel and Genie
Required Skills:python,security,debugging,opencv
More Info: https://www.timesjobs.com/job-detail/python-developer-angel-and-genie-noida-greater-noida-3-to-6-yrs-jobid-Plg0JDJmGwFzpSvf__PLUS__uAgZw==&source=srp

Company Name :FresherMart
Required Skills:python,django, Django Framework
More Info: https://www.timesjobs.com/job-detail/python-developer-freshermart-navi-mumbai-mumbai-0-to-1-yrs-jobid-PUYMPEa4NaFzpSvf__PLUS__uAgZw==&source=srp
```

The jobs we are not familiar with can be filtered out. By taking unfamiliar skills as an input from the user and removing the jobs containing that skill .

Code -

```
print("Enter the skills which you are not familiar with ")
not_familiar_with_skill=input('>')
print(f"Filtering out Non Familiar Skills:{not_familiar_with_skill}")
jobs=soup.find_all('li',class_="clearfix job-bx wht-shd-bx")
for job in jobs:
    published_date=job.find('span',class_="sim-posted").text
    if 'few' in published_date:
        company_name=job.find('h3',class_="joblist-comp-name").text.replace('  ',")
        skills=job.find('span',class_="srp-skills").text.replace('  ',")
        more_info=job.header.h2.a['href']
        if not_familiar_with_skill not in skills:
            print(f"Company Name :{company_name.strip()}")
            print(f"Required Skills:{skills.strip()}")
            print(f"More Info: {more_info}")
            print()
```

Output : Since user had entered django as an unfamiliar skill so all job posts containing **django** as skill will be removed

```
Enter the skills which you are not familiar with
>django
Filtering out Non Familiar Skills:django
Company Name :Surya Informatics Solutions Pvt. Ltd.
Required Skills:python,web technologies,linux,mobile,mysql,angularjs,javascript
More Info: https://www.timesjobs.com/job-detail/python-surya-informatics-solutions-pvt-ltd-chennai-0-to-3-yrs-jobid-UVlLes58wutzpSvf__PLUS__uAgZw==&source=srp

Company Name :GEMINI SOFTWARE SOLUTIONS
Required Skills:python,mobile,svn,nosql,python scripting,git,sql database
More Info: https://www.timesjobs.com/job-detail/qa-python-python-sdet-gemini-software-solutions-gurgaon-4-to-7-yrs-jobid-jsOuZLK8chlzpSvf__PLUS__uAgZw==&source=srp

Company Name :Gemini Solutions
Required Skills:python,mobile,svn,nosql,python scripting,git,api,sql database
More Info: https://www.timesjobs.com/job-detail/qa-python-python-sdet-gemini-solutions-gurgaon-4-to-7-yrs-jobid-eGMLzwOk2QlzpSvf__PLUS__uAgZw==&source=srp

Company Name :WHITE FORCE
Required Skills:python,mobile,debugging
More Info: https://www.timesjobs.com/job-detail/python-developer-white-force-chennai-2-to-4-yrs-jobid-hLnxtF76yRtzpSvf__PLUS__uAgZw==&source=srp

Company Name :Angel and Genie
Required Skills:python,security,debugging,opencv
More Info: https://www.timesjobs.com/job-detail/python-developer-angel-and-genie-noida-greater-noida-3-to-6-yrs-jobid-Plg0JDJmGwFzpSvf__PLUS__uAgZw==&source=srp
```

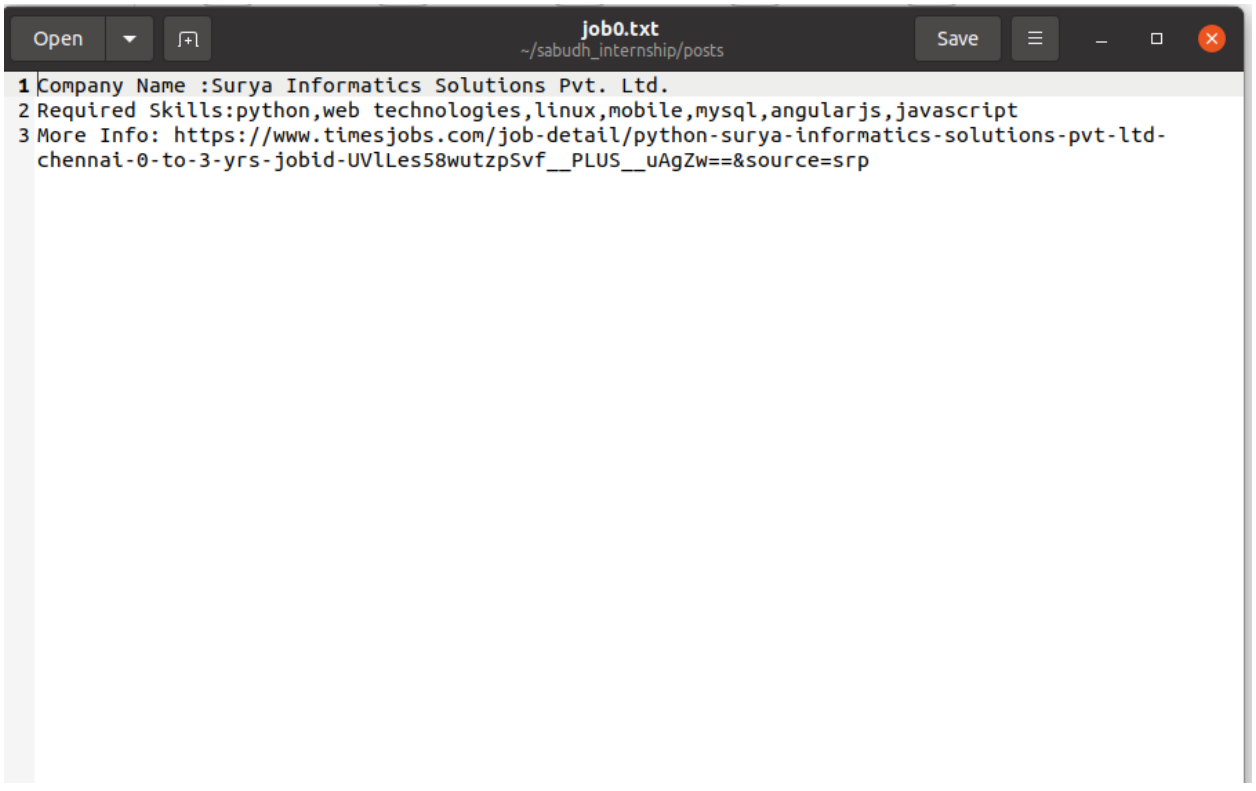
We can store this data about job Posts in text files using the below code :

```
print("Enter the skills which you are not familiar with ")
not_familiar_with_skill=input('>')
print(f"Filtering out Non Familiar Skills:{not_familiar_with_skill}")
jobs=soup.find_all('li',class_="clearfix job-bx wht-shd-bx")
for index,job in enumerate(jobs):
    published_date=job.find('span',class_="sim-posted").text
    if 'few' in published_date:
        company_name=job.find('h3',class_="joblist-comp-name").text.replace('  ',")
        skills=job.find('span',class_="srp-skills").text.replace('  ',")
        more_info=job.header.h2.a['href']
        if not_familiar_with_skill not in skills:
            with open(f'posts/job{index}.txt','w') as f:
                f.write(f"Company Name :{company_name.strip()} \n")
                f.write(f"Required Skills:{skills.strip()} \n")
                f.write(f"More Info: {more_info} \n")
                print()
            print(f"file saved as :job{index}.txt")
```

Output:-

```
Enter the skills which you are not familiar with
>sql database
Filtering out Non Familiar Skills:sql database

file saved as :job0.txt

file saved as :job1.txt

file saved as :job10.txt

file saved as :job11.txt

file saved as :job13.txt
```

All the jobs related to Python and not containing sql database as skill are stored in different files .The data inside "**job0.txt"** is :



The Similar data is stored in other text files.