Important Libraries Do a check to gain more information about requests module: https://www.w3schools.com/python/module_requests.asp

Know More about Python List: https://www.w3schools.com/python/python_lists.asp, https://www.w3schools.com/python/python_lists_access.asp

If you want to know more about BeautifulSoup : https://beautiful-soup-4.readthedocs.io/en/latest/

Not possible to cover every concept today,but I can give you the basic understanding about it: 🙂

You can always remove the '#' and run the following commands , also '#' is used to make a comment in python

```
import requests
from bs4 import BeautifulSoup as bsp
import pandas as pd
```

```
requests.get('https://internshala.com/internships/','html.parser')
```

```
<Response [200]>
```

```
requests.get('https://internshala.com/internships/','html.parser').content
```

```
name="email" autofocus="true" placeholder="Enter your e-mail address">\n
<input type="hidden" name="src_url" value="/internships/%3Fhtml.parser">\n
<input type="hidden" id="subscription_location" name="subscription_location"
value="subscription_popup_new">\n                                <span class="input-
group-btn"><button type="submit" class="btn btn-primary btn-input"
id="but_subscription_popup_form">Subscribe</button></span>\n
</div>\n                    </div>\n                </form>\n                <div
class="subscription_alert_footer">\n                    <div
class="link_container">\n                        <a href="#"
id="already_subscribed">I\'m already subscribed</a>\n                    </div>\n
<div class="link_container">\n                        <a href="#"
id="no_thanks">No, thanks</a>\n                    </div>\n
</div>\n            </div>\n        </div>\n    </div>\n\n\n    <script
nonce="YyqjQ7b/o4iVffOzkJ9I0A==">\n        if
(document.getElementById("subs_timer") != null) {\n            var distance = new
Date(\'2024/06/01 00:00:00\').getTime() - new Date(\'2024/05/31
01:23:46\').getTime();\n\n            var x = setInterval(function () {\n
if (distance < 0) {\n                    clearInterval(x);\n
document.getElementById("subs_timer").innerHTML = "";\n                }\n\n
var hours = Math.floor((distance % (1000 * 60 * 60 * 24)) / (1000 * 60 * 60));\n
var minutes = Math.floor((distance % (1000 * 60 * 60)) / (1000 * 60));\n
var seconds = Math.floor((distance % (1000 * 60)) / 1000);\n\n                if
(hours == 0 && minutes == 0 && seconds == 0) {\n
clearInterval(x);\n
document.getElementById("subs_timer").innerHTML = "";\n                }\n\n
if (hours < 10) {\n                    hours = "0" + hours\n                }\n
if (minutes < 10) {\n                    minutes = "0" + minutes\n
}\n                if (seconds < 10) {\n                    seconds = "0" +
seconds\n                }\n\n                distance = distance - 1000;\n
```

```
url='https://internshala.com/internships/'
```

Getting Responses from the website

```
resp=requests.get(url,'html.parser')
print(resp.status_code)
```

```
200
```

```
fieldname=input("Enter your profile : ")
```

```
Enter your profile : web development
```

```
field=fieldname.replace(' ','-')
```

URL Handling

```
modified_url=url+str(field)+'-internship/'
print(f"Status :{requests.get(modified_url)}")
print(modified_url)
```

```
Status :<Response [200]>
https://internshala.com/internships/web-development-internship/
```

```
resp_new=requests.get(modified_url)


soup=bsp(resp_new.content,'html.parser')
#print(soup)


# type(soup)


pages=int(soup.find('span',id='total_pages').text)
# print(pages)


urlList = []
page = 1
while page <= pages:
  newUrl = modified_url+str(f"page-{page}/")
  urlList.append(newUrl)
  page +=1
# print(urlList)


soup2 = []
for url in urlList:
  resp_new=requests.get(url)
  soup3=bsp(resp_new.content,'html.parser')
  soup2.append(soup3)
# print(len(soup2))


# print(soup.prettify())
```

Scraping

```
name=[]
for soup in soup2:
  names=soup.find_all('div',class_='individual_internship_header')
  for i in names:
    name.append(i)
# print(len(name))
#print(name)


# print(type(name))
# profile=name.find_all('h3',class_='heading_4_5 profile')
```

```
# for i in name:
#   p=i.find('h3',class_='heading_4_5 profile')
  # print(p)
  # print(p.text)
  # print(p.text.strip())
  # break


profile=[]
for i in name:
  p=i.find('h3',class_='heading_4_5 profile')
  # print(p)
  # print(p.text)
  # print(p.text.strip())
  a=p.text.strip()
  profile.append(a)
  #break
# print(len(profile))
print(f"All profiles available are : {profile}")
```

```
All profiles available are : ['Web Development', 'PHP Development', 'Demo Post', 'Flu
```

```
company=[]
for i in name:
  com=i.find('p').text.strip()
  #print(com)
  company.append(com)
# print(len(company))
print(company)
```

```
['Stirring Minds', 'UI TECH LAB LLP', 'Seven Arc Info Systems LLP', 'AppyHigh Technol
```

```
detail=[]
for soup in soup2:
  detailList=soup.find_all('div',class_='individual_internship_internship')
  for i in detailList:
    detail.append(i)
# len(detail)


#print(detail[0])


location=[]
for i in detail:
  loc=i.find('a').text
  location.append(loc)
  #print(loc)
# print(len(location))
print(f"Locations are : {location}")
```

```
Locations are : ['Delhi', 'Patna', 'Gurgaon', 'Gurgaon', 'Jaipur', 'Work from home',
```

```
duration_detail1=[]
for soup in soup2:
  duraList=soup.find_all('div',class_='item_body')
  for i in duraList:
    duration_detail1.append(i)
duration=[]
i = 1
while i < len(duration_detail1):
  duration.append(duration_detail1[i].text.strip()[0])
  i +=3
# print(len(duration))
print(duration)
```

```
['6', '6', '1', '6', '6', '6', '6', '3', '6', '6', '3', '2', '3', '6', '6', '6', '3',
```

```
stipend=[]
for soup in soup2:
  stipList=soup.find_all('span',class_='stipend')
  for i in stipList:
    val=i.text
    stipend.append(val)
# print(len(stipend))
print(f"Stipend is : {stipend}")
```

```
Stipend is : ['₹ 7,000 /month', '₹ 5,000 /month', '₹ 10,000 /month', '₹ 15,000-18,000
```

```
cont=[]
for soup in soup2:
  coutList=soup.find_all('div',class_='cta_container')
  for i in coutList:
    cont.append(i)


application_link=[]
for i in cont:
  anc=i.find('a')
  link=anc.get('href')
  #print(link)
  updated_link='https://internshala.com/'+link
  #print(updated_link)
  application_link.append(updated_link)
# print(len(application_link))
print(f"Application Link is : {application_link}")
```

```
Application Link is : ['https://internshala.com//internship/details/web-development-i
```

```
dataTable = {
    'profile': profile,
    "company": company,
    "location": location,
    "stipend":stipend,
    "duration": duration,
    "application Link": application_link
    }

df = pd.DataFrame(dataTable)
#print(f"""{len(profile)} {len(company)} {len(location)} {len(duration)} {len(stipend)} {
```

```
filename='internship_data_'+str(fieldname.replace(' ','_'))+'.csv'
print(filename)
```

```
internship_data_web_development.csv
```

```
df.to_csv(filename, index=False)
```

New Section

Find Duration of each individual internship All the information u have got till now create a dataframe out of it using pandas Save the dataframe in CSV Create a general Code to fetch data from any number of page : 😉

Remember : https://internshala.com/internships/analytics-internship/page-1/ and https://internshala.com/internships/analytics-internship/ both are same