

Day157 安全开发-Python

自动化挖掘项目&SRC目标

&FOFA资产&Web爬虫解析库



1.知识点

- 1、Python-Requests库技术
 - 2、Python-bs4-lxml库-数据筛选
 - 3、实战案例项目-edu_src自动化挖掘
-

2.演示案例

2.1 Python-WEB爬虫库&数据解析库

```
1  请求方式&POST提交&带入请求头&回显处理&加入代理等
2  #coding:utf-8
3  #author:xiaodi
4  import requests
5
6  #GET请求
7  # result=requests.get('http://www.xiaodi8.com/')
8  # print(result.text)
9
10 #post请求
11 #如果username=字典，password=字典，进行批量发送
12 #这不就是一个后台爆破脚本吗？
13 # url='http://www.xiaodi8.com/zb_system/cmd.php?
    act=verify'
14 #
    data='btnPost=%E7%99%BB%E5%BD%95&username=123123
    123&password=8fbff7db6624356b8f3874588d92ca02&sa
    vedate=1'
15 # result=requests.post(url,data=data)
16 # print(result.text)
```



```
1 #带入请求头
2 #比如: 某个地址只能手机访问, 电脑访问不了
3 #实现Python脚本尝试访问爬取数据, 那么一般怎么操作?
4 #直接利用手机数据包带入py脚本中去爬虫请求, 请求头更改!
5 #应用场景: 1、无法访问 2、绕过反爬虫 3、自定义设计(带入
  cookie)
6 # headers={
7 #     'User-Agent': 'Mozilla/5.0 (Windows NT
  10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like
  Gecko) Chrome/92.0.4515.107 Safari/537.36',
8 # }
9 #
  requests.get('http://www.xiaodi8.com', headers=headers)
```



```
1 #加入代理
2 #应用场景: 1、无法访问 2、绕过反爬虫 3、自定义设计
3 # proxy={
4 #     'http': '127.0.0.1:8080',
5 #     'https': '127.0.0.1:8080'
6 # }
7 #
  requests.get('http://www.xiaodi8.com/', proxies=proxy)
8
9 #解析库学习-Bs4_lxml库-数据筛选
10 #coding:utf-8
11 import re
12 from bs4 import BeautifulSoup
13
14 html_code='' 见测试代码
```

```
15 '''
16 #1、提取div便签数据
17 #2、提取div下class=item-top__1Z3Zo
18 #3、提取a标签targe=_blank下href
19 #4、提取所有职位信息(明文)
20
21 soup =BeautifulSoup(html_code,'lxml')
22
23 #1、提取div便签数据
24 # divs=soup.find_all('div')
25 # for div in divs:
26 #     print(div)
27
28 #2、提取div下class=item-top__1Z3Zo
29 # divs=soup.find_all('div',attrs={'class':'item-
    top__1Z3Zo'})
30 # for div in divs:
31 #     print(div)
32
33 #3、提取的a标签targe=_blank下href
34 # aaa=soup.find_all('a',attrs=
    {'targe':'_blank'})
35 # for aa in aaa:
36 #     print(aa['href'])
37 #     #print(aa.attrs['href'])
38
39
40 #4、提取所有职位信息(明文)
41 names=soup.find_all('a',attrs=
    {'target':'_blank'})
42 moneys=soup.find_all('div',class_="p-
    bom__JlNur")
```

```
43
44 for name in names:
45     print(name.string.strip())
46
47 #print(moneys)
48 for money in moneys:
49     print(money.get_text().strip())
```

2.2 Python-EDU_SRC-目标列表爬取

```
1 import requests,time
2 from bs4 import BeautifulSoup
3
4
5 #https://src.sjtu.edu.cn/rank/firm/0/?page=1
6
7
8 def edu_get_name():
9     url='https://src.sjtu.edu.cn/rank/firm/0/?
page='
10     for i in range(1,204):
11         url=url+str(i)
12         try:
13             print('获取第%s页数据'%str(i))
14             s=requests.get(url).text
15             soup = BeautifulSoup(s, 'lxml')
16             name=soup.find_all('tr',attrs=
{'class':'row'})
17             #print(name)
18             for n in name:
19                 names=n.a.string
```

```

20             #name=list(n.stripped_strings)
21             [1]
22             with open('edu.txt',
23             'a+',encoding='utf-8') as f:
24                 f.write(names + '\n')
25                 f.close()
26
27         except Exception as e:
28             time.sleep(0.5)
29             pass
30
31 if __name__ == '__main__':
32     edu_get_name()

```

2.3 Python-FOFA_API-资产信息爬取

```

1  import time,os
2  import requests,base64,json
3
4  #接口1
5  #https://fofa.info/api/v1/search/all?
6  email=your_email&key=your_key&qbase64=dG10bGU9Im
7  Jpbmci
8  #接口2
9  #"https://fofa.info/api/v1/search/stats?
10 fields=domain,port,server&qbase64=dG10bGU9IueZvu
11 w6piI%3D&email=your_email&key=your_key"
12 #查询语法
13 #title="xx大学" && country="CN""
14
15 def fofa_get_server(email,apikey):
16     # proxy = {

```

```
13      #      'http':
      'http://f777.kd1tps.com:15818',
14      #      #'https':
      'https://f777.kd1tps.com:15818'
15      # }
16      for name in open('edu.txt',encoding='utf-
      8'):
17          name=name.strip()
18          search_name='"%s" && country = "CN" &&
      title == "Error 404--Not Found"%name
19
      b=base64.b64encode(search_name.encode('utf-
      8')).decode('utf-8')
20
      url='https://fofa.info/api/v1/search/all?
      email=%s&key=%s&qbase64=%s'%(email,apikey,b)
21          s=requests.get(url).json()
22          #获取接口1
23          ss=s['size']
24          if ss!=0:
25              print('有数据来了')
26              print(search_name)
27              print(s)
28              ip=s['results']
29              for i in ip:
30                  ii=i[0]
31                  print(ii)
32                  with open('weblogic.txt', 'a+',
      encoding='utf-8') as f:
33                      f.write(ii + '\n')
34                      f.close()
35
```

```
36         # 获取接口2
37         #
        url='https://fofa.info/api/v1/search/stats?
        fields=domain,port,server&qbase64=%s&email=%s&key=%s'%(b,email,apikey)
38         # print(url)
39         # print(s)
40         # print('ip: '+str(s['distinct']['ip']))
41         # #print('server:' + str(s['distinct']
        ['server']))
42         # servers=s['aggs']['server']
43         # for server in servers:
44             #     count=server['count']
45             #     name=server['name']
46             #     print('%s个%s'%(count,name))
47             #
48         # time.sleep(10)
49
50 if __name__ == '__main__':
51     fofa_get_server('xxx@xx.com', 'xxxxxxxxxxxx')
```