python剩余+python爬虫

datetime

```
from datetime import datetime
now=datetime.now()
print(now)
date=datetime(2020,1,1,12,10)
print(date)
print(date.timestamp())
times=date.timestamp()
print(datetime.fromtimestamp(times))
strday=datetime.strptime('2015-6-1 18:19:59','%Y-%m-%d %H:%M:%S')
print(strday)#转datetime
print(now.strftime('%a,%b %d %H:%M'))#转字符串
```

```
"E:\OneDrive - cumt.edu.cn\program\python\py
2022-01-15 15:45:20.536128
2020-01-01 12:10:00
1577851800.0
2020-01-01 12:10:00
2015-06-01 18:19:59
Sat,Jan 15 15:45
```

base64

Base64是一种用64个字符来表示任意二进制数据的方法。

用记事本打开`exe`、`jpg`、`pdf`这些文件时，我们都会看到一大堆乱码，因为二进制文件包含很多无法显示和打印的字符，所以，如果要让记事本这样的文本处理软件能处理二进制数据，就需要一个二进制到字符串的转换方法。Base64是一种最常见的二进制编码方法。

pillow

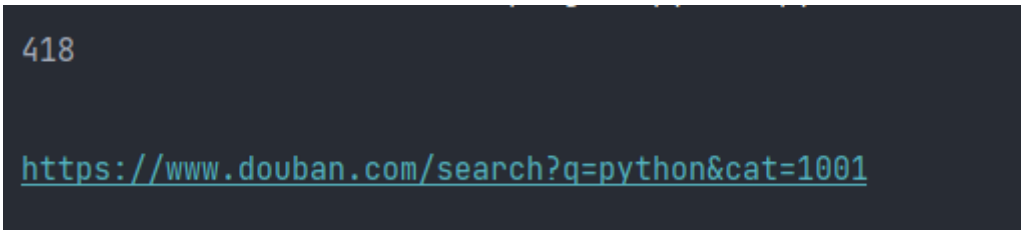
```
from PIL import Image
im=Image.open('2021-11-01.png')
w,h=im.size
print(w,h)
im.thumbnail((w//2,h//2))
im.save('2021-11-01.png','png')
```

```
"E:\OneDrive - cumt.edu.cn
960 540
```

requests

通过GET访问一个页面

```
import requests
r=requests.get('https://www.douban.com/')
print(r.status_code)
print(r.code)
r=requests.get('https://www.douban.com/search',params=
{'q':'python','cat':'1001'})
print(r.url)
```



python爬虫应用

爬虫：通过编写程序，模拟浏览器上网，然后让其去互联网上抓取数据的过程

通用爬虫：抓取系统重要组成部分，抓取的是一整张页面数据

聚焦爬虫：建立在通用爬虫的基础上，抓取的是页面中特定的局部内容

增量式爬虫：检测网站中数据更新的情况，抓取网站中最新更新出来的数据

反爬机制：门户网站，可以通过制定相应的策略或者技术手段，防止爬虫程序进行网站数据的爬取

反反爬策略：通过技术手段破解反爬机制

http协议：服务器和客户端进行数据交互的一种形式

User-Agent:请求载体的身份标识

Connection:请求完毕后，是断开连接还是保持连接

Content-Type：服务器响应回客户端的数据类型

https协议：安全的超文本传输协议（数据传输有加密）

加密方式：对称密钥加密，非对称密钥加密，证书密钥加密

requests模块

使用：指定URL，发起请求，获取响应数据，持久化存储

```
#需求：爬取搜狗首页的页面数据
import requests
#指定URL
url='https://www.sogou.com/'
response=requests.get(url=url)
page_text=response.text#字符串形式的响应数据
print(page_text)
with open('./sohou.html','w',encoding='utf-8')as fp:
    fp.write(page_text)
print('爬取结束')
```

bs4模块实战

爬取三国演义章节标题和内容

《三国演义》全集在线阅读 史书典籍诗词名句网 (shicimingju.com)

```
#爬取三国演义小说章节标题和内容
import requests
from bs4 import BeautifulSoup
if __name__=="__main__":
    #对首页页面数据爬取

    headers={
        "User-Agent": "Mozilla / 5.0( Windows NT10.0;Win64;x64) AppleWebKit /
537.36(KHTML, likeGecko)Chrome / 97.0.4692.71Safari / 537.36Edg / 97.0.1072.55"
    }
    url='https://www.shicimingju.com/book/sanguoyanyi.html'
    page_text=requests.get(url=url,headers=headers).text
    #解析
    #实例化Beautifulsoup
    soup=BeautifulSoup(page_text,'html.parser')
    li_list=soup.select('.book-mulu > ul > li')
    fp=open('./sanguo.txt','w',encoding='utf-8')
    for li in li_list:
        title=li.a.string
        detail_url='https://www.shicimingju.com'+li.a['href']
        #详情页请求
        detail_page_text=requests.get(url=detail_url,headers=headers).text
        #解析内容
        detail_soup=BeautifulSoup(detail_page_text,'html.parser')
        div_tag=detail_soup.find('div',class_='chapter_content')
        content=div_tag.text
        fp.write(title+':'+content+'\n')
        print(title,'爬取成功')
```

xpath模块实战

全国城市名称爬取

```
#解析所有城市名称
import requests
from lxml import etree
if __name__=="__main__":
    headers={
```

```
        "User-Agent": "Mozilla / 5.0(Windows NT 10.0;Win64;x64) AppleWebKit /
537.36(KHTML, likeGecko) Chrome / 97.0.4692.71Safari / 537.36Edg / 97.0.1072.55"
    }
    url='https://www.aqistudy.cn/historydata/'
    page_text=requests.get(url=url,headers=headers).text
    tree=etree.HTML(page_text)
    #热门城市
    host_li_list=tree.xpath('//div[@class="bottom"]/ul/li')
    all_city_names=[]
    for li in host_li_list:
        hot_city_name=li.xpath('./a/text()')[0]
        all_city_names.append(hot_city_name)
    #全部城市
    city_names_list=tree.xpath('//div[@class="bottom]/ul/div[2]/li')
    for li in city_names_list:
        city_name=li.xpath('./a/text()')[0]
        all_city_names.append(city_name)
    print(all_city_names,len(all_city_names))
```