day04

- 复习
- urllib3 方法介绍

复习

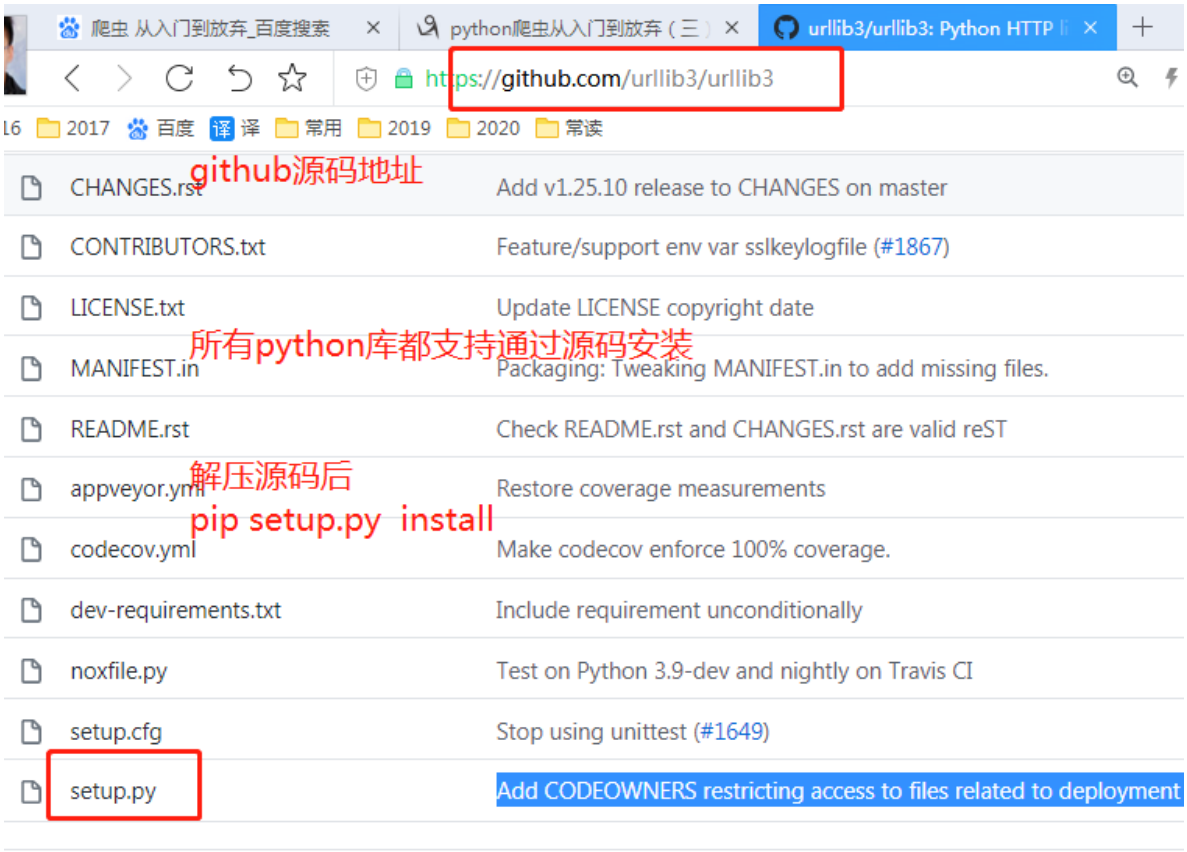
爬虫的基础库: urllib

```
# 1. 版本: python2.x 自带 urllibe 和urllib2
          python3.x 自带urllibe =(urllib+urllib2)   urllib3需要额外扩展
# 2. urllib 包, 包含4个模块
     urllib.request   发送请求   重要
     urllib.error     错误处理,直接使用Exception统一捕获
     urllib.parse     解析url(编码和参数生成)
     urllib.robotparses   解析爬虫协议模块
# 3. 相关方法

     参考表格
```

urllib3

> urllib3 是python3.x 中urllib的升级, 但是和urllib不能完全替代!
>
> 需要手动下载:
>
> pip install urllib3
>
> 所有pyhton库如果pip install 无法下载,可以去pipy中下载轮子文件 或直接下载git中的源码,通过源码安装
>
> git clone https://github.com/urllib3/urllib3
>
> 解压安装
>
> python setup.py install



优点

```
urllib3 功能强大,提供urllib 标准库很多没有的重要特性
1. 线程安全
2. 连接池  [加快爬虫访问速度的手段通过连接池可以1次打开N个通道!]
3. 编码文件实现上传
4. 支持压缩
....
```

重点: urllib3 和urllib结构一样,都包含4个核心模块! 比如请求: urllib3.request

方法列表

- 发送请求
- 请求中添加参数
- post请求
- 发送json请求
- 上传文件

方法/属性	说明	参数
pools=urllib3.PoolManager(num_pools=10,timeout=3)	创建连接池对象	num_pools连接池数量,timeout表示超时时间秒
response=pools.request('GET',url,field,headers,retries=3)	通过连接池发送请求	方法,url,参数,请求头,retries表示重复请求次数默认3次!
respone.status	状态码	
respone.data.decode()	响应数据	
response =pools.request('POST',url=url,body=encode_data,headers={'Context-Type':'application/json'})	请求传输json数据	body是编码后的json数据,header中指定内容类型为json
pools.request("POST",url=url,headers=headers, fields={'mytxt':('1.txt',file_read,'text/plain')})	上传文本	fields要指定文件信息(名字,数据,类型)
response=pools.request('GET',url,field,headers,retries=False,redirect=Flase)	关闭重试和重定向	
response = pools.request("POST",url=url,body=file_read,headers={"Context-Type":"images/jpeg"})	上传图片	body指定数据,header中指定类型

```
import  urllib3

# 创建连接池(可以同时打开N个请求通道)
pools = urllib3.PoolManager(num_pools=10)
# 通过连接池对象发送请求
respone = pools.request('GET','https://cuiqingcai.com/')
print(respone.status) #获取状态
print(respone.data.decode())
```

向请求中传入json数据

```
import  urllib3
import  json #导入python自带的json模块

# 紧用于爬后台接口!!!
```

```
# JSON发起请求时,需要把数据转化为json字符串格式， 在请求头中需要指定内容类型为
application/json
data = {'uname':'zhangsan'}
url = "http://httpbin.org/post"
encode_data = json.dumps(data).encode() # 数据编码

pools = urllib3.PoolManager()

response =pools.request('POST',url=url,body=encode_data,headers={'Context-
Type':'application/json'})
print(response.data.decode())
```

上传文件

- 表单上传数据需要指定数据传输文件编码方式:enctype

```
<form action="http://www.jd.com" enctype='multipart/form-data' >
////
</form>

# 在文件上传时，所使用的编码类型应当是“multipart/form-data”，它既可以发送文本数据，也支持
二进制数据上载。
# 默认表单传输数据格式:“application/x-www-form-urlencoded”。 key:value 名字值的形式，
只能传送字符串！
```

上传文本

```
import  urllib3

with open('1.txt',mode='r+',encoding='utf-8') as f:
    file_read = f.read()

headers={
    'User-Agent':'Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/78.0.3904.108 Safari/537.36'
}
url = "http://httpbin.org/post"
pools = urllib3.PoolManager()
response = pools.request("POST",url=url,headers=headers,
                         fields={'mytxt':('1.txt',file_read,'text/plain')})
print(response.data.decode())
```

上传图片

```
import  urllib3

# 模式:read bytes 读取二进制
with open('img2.jpg',mode='rb') as f:
    file_read = f.read()

url = "http://httpbin.org/post"
pools = urllib3.PoolManager()
response = pools.request("POST",url=url,body=file_read,headers={"Context-
Type":"images/jpeg"})
print(response.data.decode())
```

Requests

工作中常用的爬虫库Requests,封装了urllib3, 更加简洁!

- 安装

```
pip install requests
```

- 中文API手册

```
# github源码
https://github.com/kennethreitz/requests
# 中文官方手册
https://requests.readthedocs.io/zh_CN/latest/
```

方法列表

发送请求方法

方法	说明	参数介绍
request.get(url,headers,params)	get请求	
request.post(url,headers,data)	发送post请求	

Response响应对象方法

方法	说明	参数
response.status_code	状态码	
response.headers		

总结

- urllib 和urllib3 作为基础库了解!!!
- 工作中使用是升级或简化的 requests

任务

- 整理urllib 和urllib3方法手册
- 整理requests手册
- 练习案例库中的爬虫案例!