**Python爬虫小结**

# 1.找到待爬网址，以及其html信息

## 如何访问互联网？🡪urlopen(‘URL’)

import urllib.request

**respond = urllib.request.urlopen(r'https://fishc.com.cn/')** // urlopen输入网址即可，为安全可以加r表示原始字符

**html = respond.read()** //读取网址源代码

print(html) //打印源代码，但未解码，较乱

**html = html.decode('gbk')** //按网址数据格式手动解码

print(html)



## 如何实现自动网页解码？（字符等信息需要解码）

采用chardet模块

import urllib.request

response = urllib.request.urlopen("https://fishc.com.cn/").read()

**import chardet**

**chardet.detect(response)**

>>> {'encoding': 'GB2312', 'confidence': 0.99, 'language': 'Chinese'}

#返回字典：编码格式，置信度，语言

>>> if chardet.detect(response)['encoding'] == 'GB2312':

**response.decode('GBK') //解码**

## 下载一只猫的图片

**import urllib.request**

**方法1：**

**response = urllib.request.urlopen('https://placekitten.com/500/600') //图片的url,而不是网站的url**

**cat\_img = response.read()**

**with open('cat\_500\_600.jpg','wb') as f:**

**f.write(cat\_img)**

方法2：

req = urllib.request.Request('https://placekitten.com/300/200')

response = urllib.request.urlopen(req)

cat\_img = response.read():

with open('cat\_300\_200.jpg','wb') as f:

f.write(cat\_img)

## 获取今日头条的图片

**import urllib.request**

**response=urllib.request.urlopen(r'https://p1.pstatp.com/list/190x124/pgc-image/RXwfudS7C0UzPl') #今日头条该图片的来源网址！**

**img = response.read()**

**#img = img.decode('uft-8') // 图片不需要解码，字符信息才需要**

**with open('toutiao.jpg','wb') as f:**

**f.write(img)**

**总结：图片不需要解码，抓取网站的图片，关键是找个网站图片的来源网址！**

# 2.设置代理或者延时，避免IP被屏蔽；

## 设置延时->time模块

**import time**

print('Hello!')

**time.sleep(5)**

print('World!')

## User-agent隐藏python操作

req = urllib.request.Request(url) #打开网址

req.addheaders=[('User-Agent','Mozilla/5.0 (Windows NT 6.1;WOW64) AppleWebKit/537.36 (KHTML,like Gecko) Chrome/55.0.2883.87 Safari/537.36')] #**设置User-Agent隐蔽**

response = urllib.request.urlopen(req) #爬取源数据

html = response.read().decode('UTF-8') #数据解码

req = urllib.request.Request(url)

req.add\_header('User-Agent','Mozilla/5.0 (Windows NT 6.1;WOW64) AppleWebKit/537.36 (KHTML,like Gecko) Chrome/55.0.2883.87 Safari/537.36')

response = urllib.request.urlopen(req)

如何获取本机的User-Agent?

方法1：在线获取网址<http://www.useragentstring.com/> (**推荐**)

方法2：在地址栏中输入：about:version（**推荐**）

方法3：地址栏中输入：javascript:alert(navigator.userAgent) （不能复制粘贴）

## 设置代理

1. 参数是一个字典{‘类型’:’代理ip：端口号’}

proxy\_support = urllib.request.ProxyHandler({})

1. 定制、创建一个opener

opener = urllib.request.build\_opener(proxy\_support)

1. 安装、调用opener

urllib.request.install\_opener(opener)

opener.open(url)

1. 实例

import urllib.request

url='https://www.whatismyip.com' #爬取的网址

proxy = {'http': '125.127.24.239:808'} #设置代理的ip,网上搜索代理ip，一大堆

proxy\_support = urllib.request.ProxyHandler(proxy) #发送代理请求

opener=urllib.request.build\_opener(proxy\_support) #定制、创建一个opener

#用User-Agent参数隐藏opener人机

opener.addheaders=[('User-Agent','Mozilla/5.0 (Windows NT 6.1;WOW64) AppleWebKit/537.36 (KHTML,like Gecko) Chrome/55.0.2883.87 Safari/537.36')]

urllib.request.install\_opener(opener) #安装opener

html=urllib.request.urlopen(url).read().decode('utf-8') #爬取数据并解码

print(html)

# 3.html源数据信息提取过滤：正则表达式

ip地址的正则表达式：re.search(r'>((25[0-5]|2[0-4]\d|((1\d{2})|([1-9]?\d)))\.){3}(25[0-5]|2[0-4]\d|((1\d{2})|([1-9]?\d)))<',str1)

端口的正则表达式: re.search(r'>([0-9]|[1-9]\d{1,3}|[1-5]\d{4}|6[0-5]{2}[0-3][0-5])<',str1)

# 4.数据保存

爬取妹子图实例

import urllib.request

import os

def url\_open(url):

req = urllib.request.Request(url)

req.add\_header('User-Agent','Mozilla/5.0 (Windows NT 6.1;WOW64) AppleWebKit/537.36 (KHTML,like Gecko) Chrome/55.0.2883.87 Safari/537.36')

proxy = {'http': '125.127.24.239:808'} #设置代理的ip,网上搜索代理ip

proxy\_support = urllib.request.ProxyHandler(proxy) #发送代理请求

opener=urllib.request.build\_opener(proxy\_support) #定制、创建一个opener

#用User-Agent参数隐藏opener人机

opener.addheaders=[('User-Agent','Mozilla/5.0 (Windows NT 6.1;WOW64) AppleWebKit/537.36 (KHTML,like Gecko) Chrome/55.0.2883.87 Safari/537.36')]

urllib.request.install\_opener(opener) #安装opener

response = urllib.request.urlopen(req)

html = response.read()

return html

def get\_page(url):

html = url\_open(url).decode('utf-8')

a = html.find('current-comment-page') + 23

b = html.find(']',a)

return html[a:b]

def find\_imgs(url):

html = url\_open(url).decode('utf-8')

img\_addrs = []

a = html.find('img src=')

while a != -1:

b = html('.jpg',a,a+255)

if b!= -1:

img\_addrs.append(html[a+9:b+4])

else:

b = a+9

a = html.find('img src=',b)

return img\_addrs

def save\_imgs(folder,img\_addrs):

for each in img\_addrs:

filename = each.split('/')[-1]

with open(filename,'wb') as f:

img = url\_open(each)

f.write(img)

def download\_mm(folder='OOXX',pages = 10):

os.mkdir(folder)

os.chdir(folder)

url = "http://jandan.net/ooxx/"

page\_num = int(get\_page(url))

for i in range(page):

page\_num -= i

page\_url = url + 'page-' + str(page\_num) + '#comments'

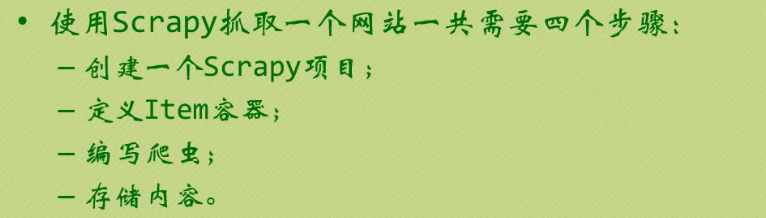
img\_addrs = find\_imgs(page\_url)

save\_img(folder,img\_addres)

if \_\_name\_\_ == '\_\_main\_\_':

download\_mm()

# scrapy网络爬虫框架



Pip install scrapy

Conda install scrapy