在python/scripts下执行pip install Scrapy

URL就是同意资源定位符(Uniform Resource Locator)，它的一般格式如下(带方括号[]的为可选项)：

    protocol :// hostname[:port] / path / [;parameters][?query]#fragment

    URL的格式由三部分组成：

**(1)protocol：**第一部分就是协议，例如百度使用的就是https协议；

**(2)hostname[:port]：**第二部分就是主机名(还有端口号为可选参数)，一般网站默认的端口号为80，例如百度的主机名就是www.baidu.com，这个就是服务器的地址;

**(3)path：**第三部分就是主机资源的具体地址，如目录和文件名等。

    网络爬虫就是根据这个URL来获取网页信息的

使用urlib这个组件抓取网页：

**1**.urllib.request模块是用来打开和读取URLs的；

**2**.urllib.error模块包含一些有urllib.request产生的错误，可以使用try进行捕捉处理；

**3**.urllib.parse模块包含了一些解析URLs的方法；

**4**.urllib.robotparser模块用来解析robots.txt文本文件.它提供了一个单独的RobotFileParser类，通过该类提供的can\_fetch()方法测试爬虫是否可以下载一个页面。

from urllib import request

response=request.urlopen("http://fanyi.baidu.com")

html=response.read()

html=html.decode("utf-8")       #decode()命令将网页的信息进行解码，并显示出来

print(html)

* **geturl()**返回的是一个url的字符串；
* **info()**返回的是一些meta标记的元信息，包括一些服务器的信息；
* **getcode()**返回的是HTTP的状态码，如果返回200表示请求成功。
* 例如response.geturl()  response.info()  response.getcode()

urllib.error可以接收有urllib.request产生的异常。urllib.error有两个方法，URLError和HTTPError。

from urllib import request

from urllib import error

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

    url='http://www.douyu.com/Jack\_Cui.html'

    req=request.Request(url)

    try:

        response=request.urlopen(req)

    except error.HTTPError as e:

        print('HTTPError')

        print(e.code)

    except error.URLError as e:

        print('URLError')

        print(e.reason)

如果想用HTTPError和URLError一起捕获异常，那么需要将HTTPError放在URLError的前面，因为HTTPError是URLError的一个子类。

    url2='http://www.csdn.net/'

    head={}

    #设置user agent，模拟浏览器访问

    head['User-Agent']='Mozilla/5.0 (Linux; Android 4.0.4; Galaxy Nexus Build/IMM76B) AppleWebKit/535.19 (KHTML, like Gecko) Chrome/18.0.1025.133 Mobile Safari/535.19'

    req2=request.Request(url2,headers=head)

    #req2=request.Request(url2)

    response=request.urlopen(req2)

    html = response.read().decode('utf-8')

    print(html)

  req.add\_header('User-Agent', 'Mozilla/5.0 (Linux; Android 4.1.1; Nexus 7 Build/JRO03D) AppleWebKit/535.19 (KHTML, like Gecko) Chrome/18.0.1025.166 Safari/535.19')另外一种设置方式add\_header

    #class urllib.request.Request(url,data=None,headers={},origin\_req\_host=None,unverifiable=False,method=None)

IP代理

调用urlib.request.ProxyHandler()，proxies参数为一个字典；

创建Opener(类似于urlopen，这个代开方式是我们自己定制的)；

安装Opener

from urllib import request

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

    url='<http://www.whatismyip.com.tw/'>

   #代理ip

    proxy={'http':'106.46.136.112:808'}

   #创建ProxyHandler

    proxy\_support=request.ProxyHandler(proxy)

   #创建opener

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

   #添加UserAgent

    opener.addheaders=[('User-Agent','Mozilla/5.0 (Linux; Android 4.1.1; Nexus 7 Build/JRO03D) AppleWebKit/535.19 (KHTML, like Gecko) Chrome/18.0.1025.166 Safari/535.19')]

   #安装opener

    request.install\_opener(opener)

   #使用自己安装好的opener

    response=request.urlopen(url)

   #读取响应的信息并解码

    html=response.read().decode('utf-8')

   #打印信息

    print(html)

<http://www.lfd.uci.edu/~gohlke/pythonlibs/>，Ctrl+F搜索Lxml、Twisted、Scrapy，下载对应的版本（cp36代表python版本是3.6，amd64是win64位）

**pip3 install wheel**

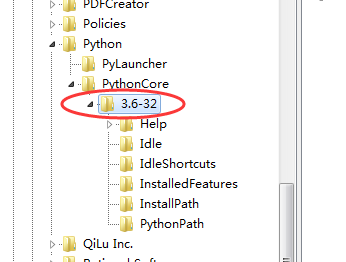
**pip3 install lxml-4.1.1-cp36-cp36m-win\_amd64.whl**

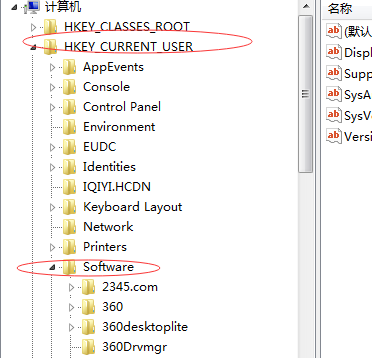
**pip3 install Twisted-17.9.0-cp36-cp36m-win\_amd64.whl**

**pip3 install Scrapy-1.5.0-py2.py3-none-any.whl**

Srapy已经安装成功，还要下载**pywin32**，找到对应版本下载，一路下一步安装即可。

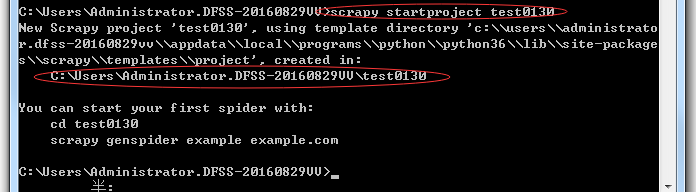
安装时提示安装找不到注册表信息，按windows+r---regedit--把python3.6修改为python3.6-32后再重新双击exe安装即可。安装成功后输入import win32com 不报错即安装成功

****

****

创建项目

scrapy startproject test0130

****