## python 爬虫框架scrapy安装步骤

python相关包下载地址https://www.lfd.uci.edu/~gohlke/pythonlibs/#twisted

安装pip 下载pip-9.0.1 命令行进入下载的pip文件夹下 执行命令python setup.py install 安装结果检查:pip --version

安装lxml和twisted 下载.whl文件 执行 pip install lxml.whl和 pip install twisted.whl 安装结果检查:python下执行命令 import lxml和 import twisted

安装pywin32和pyOpenSSL:windows下直接使用安装程序安装，注意pywin32需要对应python的安装版本

安装scrapy 执行pip install scrapy 安装结果检查::python下执行命令 import scrapy

## 创建python scrapy项目

命令行下执行scrapy startproject projectName 会在当前文件夹下生成一个scrapy工程 pycharm open 可以打开这个项目

项目创建后会自动创建相关目录与文件：

spiders 爬虫目录，如：创建文件，编写爬虫规则

scrapy.cfg 项目的配置信息，主要为Scrapy命令行工具提供一个基础的配置信息。（真正爬虫相关的配置信息在settings.py文件中）

items.py 设置数据存储模板，用于结构化数据，如：Django的Model

pipelines 数据处理行为，如：一般结构化的数据持久化

settings.py 配置文件，如：递归的层数、并发数，延迟下载等

spiders文件夹下新建一个scrapydemo.py文件,import scrapy,定义一个类并集成scrapy.spoders.Spider

import scrapy

class scrapydemo(scrapy.spiders.Spider):

name = "scrapydemo"#必须定义name

start\_urls = ["https://list.jd.com/list.html?cat=9987,653,655"]#定义要抓取的网页url列表，scrapy源码是一个for循环

def parse(self,response):#必须实现parse方法 scrapy源码中默认callback函数的函数名

#筛选div标签中 属性class=p-name 的<a>/<em>字节点的元素值

phonename = response.xpath('//div[@class="p-name"]/a/em/text()').extract\_first()

print(phonename)

运行scrapy工程:命令行下进入工程目录 执行scrapy crawl projectname --nolog

## scrapy查询语法：

查询子子孙孙中的某个标签(以div标签为例)：//div

查询儿子中的某个标签(以div标签为例)：/div

查询标签中带有某个class属性的标签：//div[@class=’c1′]即子子孙孙中标签是div且class=‘c1’的标签

查询标签中带有某个class=‘c1’并且自定义属性name=‘alex’的标签：//div[@class=’c1′][@name=’alex’]

查询某个标签的文本内容：//div/span/text() 即查询子子孙孙中div下面的span标签中的文本内容

查询某个属性的值（例如查询a标签的href属性）：//a/@href

## scrapy parse使用yield

yield关键词,作用是返回一个对象(parse会调用pipelines处理)后继续执行以实现递归处理

使用了yield parse函数会被当做一个生成器使用.scrapy会逐一获取parse方法中生成的结果，并判断结果是什么样的类型

如果是request则加入爬取队列,如果是item类型则使用pipeline处理,其他类型返回错误信息

## scrapy request对象：

class scrapy.http.Request(url[, callback, method='GET', headers, body, cookies, meta, encoding='utf-8', priority=0, dont\_filter=False, errback])

一个request对象代表一个HTTP请求，通常有Spider产生，经Downloader执行从而产生一个Response。

参数：url(string): 用于请求的URL

callback(callable):指定一个回调函数，该回调函数以这个request是的response作为第一个参数。如果未指定callback，

则默认使用spider的parse()方法。

method(string):HTTP请求的方法，默认为GET（看到GET你应该明白了，过不不明白建议先学习urllib或者requets模块）

meta(dict):指定Request.meta属性的初始值。如果给了该参数，dict将会浅拷贝。

body(str):the request body.

headers(dict):request的头信息。

cookies(dict or list):cookie可以使用dict和list

当request的response被下载时会调用callback指定的函数并以response对象作为参数

http://blog.csdn.net/github\_38196368/article/details/72755669

http://blog.csdn.net/qq\_28057541/article/details/61921632

http://blog.csdn.net/willib/article/details/51873259