1. 用正则匹配整个网页内容时，需注意要在传入的参数中加上re.DOTALL,表示和整个网页匹配，如：re.match(‘(.\*)’,response.text,re.DOTALL),否则将只能匹配到第一行的内容
2. 用requests库发出请求可以用request.get(url,headers=header)要附加提交数据

用post(url,data=data,headers=header);scrapy中发出

请求用scrapy.http.Request(url,headers=header,callback=fuc),

要附加提交数据用scrapy.FormRequest(url=url,formdata=data,headers=header,callback=fuc),

其中callback是回调函数，能对提交的请求返回的response作出处理,定义如fuc(response);

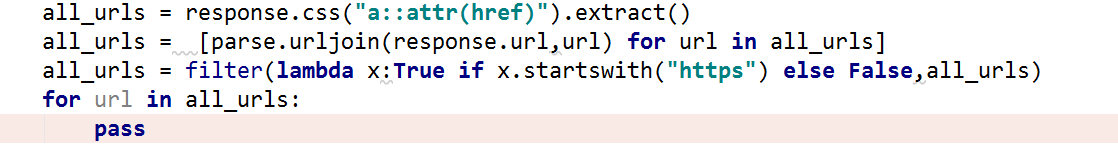
3.requests中要使用session需要用requests.session(),并用session.cookies和cookie绑定:

Session.cookies=cookielib.LWPCookieJar(filename=’file’)，file为已保存的cookie文件，cookie文件可以在session访问后用session.cookies.save()来保存；而scrapy则会自行保存cookie，并在访问时将cookie加入Request来访问

1. scrapy的Request类进行访问是异步进行的，所以要设置一个参数callback来表示对访问的url所返回的response进行下一步的操作，注意到我们访问网页的方式是直接创建了Request(...)这个对象，这就和新建一个线程很像，而线程中的也有一个处理函数，因此异步的处理过程都是很像的，给一个对象内部的参数定义好，只需将其新建出来后就不用管了
2. 关于scrapy中的yield和return:

首先要知道yield是用来生成一个generator（生成器），并将其作为返回值返回,这是一个可迭代对象，而scrapy中则需要在后台接受一个Request组成的可迭代对象来依次访问它们，所以用yield是可以达到这种效果的，同理return一个Request的数组也是可以的，但是对于有些时候元素过多时，用yield可以提高效率

1. scrapy shell是一个交互式的shell，需要先安装ipython;执行scrapy shell <url>就会使用Scrapy downloader下载指定url的页面数据,并且打印出可用的对象和函数列表 ;当想要给其添加header时，可以使用scrapy shell -s xxx=xxx <url>
2. 利用filter筛选url，案例如下：



要善于用lambda表达式来简化一些需要传入函数的地方

8.xpath的筛选常用格式如：**"//\*[class='job\_request']/p/span[2]/text()"**

注意其中span[2]代表p元素下的第二个元素，其序号下标是从1开始而非0开始

1. 关于itemloader：

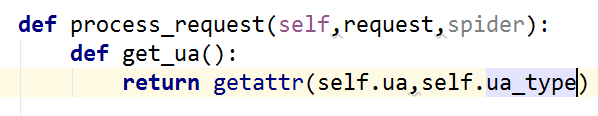
我们在spider中利用itemloader的css选择器或者xpath先获取网页中内容，而把之后的处理交给item.py中对应的XXXitem,在里面定义了要爬取的变量xxx=scrapy.Field(input\_processor=xxx,output\_processor=xxx),其中可以传入的有：

MapCompose()里面填一个函数对象(不带括号),这个函数需要能接受一个value值，表示是选择器选择下来的字段，然后在里面进行处理，再return一个值就是处理后的该变量的值。

而且MapCompose函数里面可以填多个函数对象，可以让函数对象依次执行，直到达到自己所预想的效果

Join(‘,’)可以将字符串数组之间用‘，’连接起来，TakeFirst()则是取数组中的第一个，如果没有元素则返回None。而input\_processor和output\_processor区别是前者作用于一个list对象中的每一个元素，后者直接作用于list对象

1. python等动态语言的特性：能在函数中继续定义函数



而getattr(a,b)可以获取a对象的b属性，其中b属性如果是一个变量，如

b=‘random’,则获取的是a的random属性,如果是一串字符串字面值则直接是该字面的属性；这种方法相比于直接的a.b等于是将要获取的属性写活了，要特别注意这种用法

1. 关于format的用法：

支持诸如str.format(a,b,c)：其中str中包含{0},{1},{2}，这时候就将a,b,c

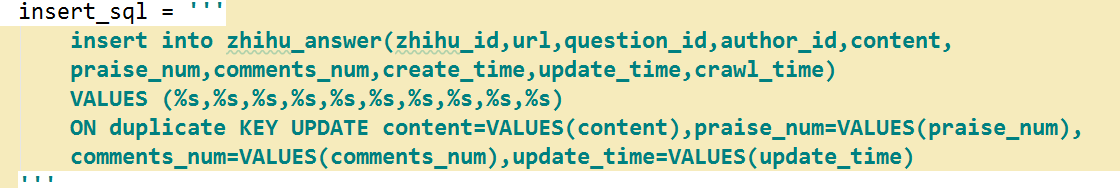
的值按照顺序分别赋给这三个位置

1. 从mysql中随机读取数据：

[select](http://www.111cn.net/tags.php/select/" \t "http://www.111cn.net/database/mysql/_blank) \* from 表名 order by rand( ) limit 1;  //此处的1就是取出数据的条数

1. 当想向mysql中插入数据又怕出现主键相同导致报错时可用结构：

On duplicate key update xxx=values(content)....



1. ip代理：scrapy-crawlera收费版ip代理，简单好用

Tor洋葱网络：需要VPN，稳定好用

1. 要为每个spider设置不同的settings，可以在class的开头新增一项custom\_settings的dict，

它的key可以去settings中复制，value则根据需求定制，这样scrapy底层会优先选用custom\_settings，当没有此项时再使用settings中的全局设置；其中可以定制的扩展项包括cookie的禁用，自动限速等，具体可去查看官方文档