爬虫面试题整理

1.Q:爬虫爬取数据的流程？

A:

1.明确数据采集需求；

2.分析要采集数据的url和相关参数；

3.编码实现功能，1)获取url,2)对url进行筛选，找到自己想要的部分，3)入库，4)对数据去重；注意反爬虫的规则：1.验证码的识别；2.使用代理；3.httpclient头信息。

2.Q:如何抓取动态页面？

A:动态网页指几种可能：1）需要用户交互，如常见的登录操作；2）网页通过JS / AJAX动态生成，如一个html里有<div id="test"></div>，通过JS生成<divid="test"><span>aaa</span></div>；3）点击输入关键字后进行查询，而浏览器url地址不变

第一种方法:直接url法

（1）认真分析页面结构，查看js响应的动作；

（2）借助于firfox的firebug分析js点击动作所发出的请求url；

（3）将此异步请求的url作为scrapy的start\_url或yield request再次进行抓取。

第二种方法:借助于selenium

Selenium基于Javascript 并结合其WebDriver来模拟用户的真实操作，它有很好的处理Ajax的能力，并且支持多种浏览器（Safari，IE，Firefox，Chrome），可以运行在多种操作系统上面，Selenium可以调用浏览器的API接口，selenium 会打开一个浏览器，然后在新打开的浏览器中执行程序中模拟的动作。

第三种方法：scrapyjs

Splash是一个由 Scrapy 鼻祖 ScrapingHub 所创建的Javascript渲染服务。它是一个实现了HTTP API的轻量级浏览器，Splash是用Python实现的，同时使用Twisted和QT。Twisted（QT）用来让服务具有异步处理能力，以发挥webkit的并发能力，能够并行处理多个网页的轻量级网页浏览器，它可以执行自定义 JavaScript 代码并利用关闭图片或广告屏蔽的功能来提升渲染速度。

3.Q:验证码是如何解决的？

A:如果直接从生成验证码的页面把验证码下载到本地后识别，再构造表单数据发送的话，会有一个验证码同步的问题，即请求了两次验证码，而识别出来的验证码并不是实际需要发送的验证码。方法1.用session；方法2.用cookie；方法3.selenium+手动构造cookie,该方法无需识别验证码,难点在于确定该网站是用cookie中的什么key值来表示“用户名”和“密码”的。而且好像有些cookie是加密过的。可以先用get\_cookies()进行观察。

最有效的解决办法就是直接输出，手动输入验证码

4.Q:怎么模拟登陆？

A:方法一：直接使用已知的cookie访问

方法二：使用无头浏览器访问

5.Q:如何实现ajex请求的数据获取？

A:方法一：尝试用BeautifulSoup抓取

方法二：用selenium + PhantomJS抓取

6.Q:遇到过什么反扒机制，怎么解决的？

A:比如下拉加载数据的页面，可以使用selenium + PhantomJS模拟用户登录状态进行抓取，比如使用urllib2的request请求隐藏登录信息，设置User-Augent模拟浏览器，比如需要验证码登录，可以用cookic保存登录信息

7.Q:怎么使用私密ip代理 ，代理哪里找的，量是多少，怎么判断是否能用？

A:使用urllib2.ProxyHandler()设置IP代理信息，使用urllib2.build\_opener()构建自定义opener,可以在西刺代理的网站上找到私密代理，高匿IP的服务器，一般包年2万多，每天有3000个代理IP可以用，可以用IP代理池多线程ping百度在把记录状态保存在本地

8.Q:爬虫携带cookie，cookie过期怎么办？

A:加载cookies文件，使用requests库爬取数据并动态更新cookies，可以使cookies不失效，或者就是重新登录

9.Q:url如何去重？数据入库前如何去重？

A:第一种方法你可以在保存到数据库的时候，，创建表的时候链接加上unique，那么相同的链接就不会重复加了。第二种方法，可以用集合去重，，就是先把网址的列表转成set，然后再list转回来。。就会自动过滤掉重复网址，但是使用set方法会导致处理完的数据会被打乱顺序，如果是考虑有序性的数据，不建议用set去重

10.Q:如何利用scrapy框架发送post请求？

A:可以使用 yield scrapy.FormRequest(url, formdata, callback)方法发送POST请求。

如果希望程序执行一开始就发送POST请求，可以重写Spider类的start\_requests(self) 方法，并且不再调用start\_urls里的url。

11.Q:如果我要爬取的淘宝某个人的历史消费信息（登陆需要账号、密码、验证码），你会如何操作？

A:使用内置浏览器，可以直接使用账号密码进行登录，验证码可以保存到本地手动输入验证码，或者可以使用多线程连接打码网站进行解决

12.Q:接上题，你觉得除了账号、密码、验证码外，还需要携带哪些信息呢？

A:基本上除了解压缩的参数，其他都要携带上，以防止反爬机制用到其他参数

13.Q:每天爬取的数据量是多少，增量是多少，总量是多少，怎么知道增量是否重复？

A:具体要先分析抓取网站的数据量，在分时间段进行抓取，控制每天的抓取量。增量重复要看具体的去重策略，可以使用

14.Q:如何实现对不同用户获取同一页面不同内容的抓取？

A:设置cookic池，模拟多用户登录，用线程随机分配任务

15.Q:分布式爬取是如何设置的？

A:分点配置服务器，使用scrapy-reids：在Scrapy的基础上添加了一套以Redis数据库为核心的组件，让Scrapy框架支持分布式的功能，主要在Redis里做请求指纹去重，请求分配，数据临时存储。

16.Q:爬虫一共几台服务器在跑？

A:分点放置在不同的地方，最少3台

17.Q:项目中遇到哪些困难？

A:爬取知乎的时候，知乎的反爬机制是封账号，不是封ＩＰ，爬取腾讯的子网站，必须是ＱＱ号登录，而且cookie只有６分钟寿命，每天只能登录130次，超过就封号。

Q:

A: