链滴

HttpClient 入门

作者：fenng

原文链接：https://ld246.com/article/1497516877409

来源网站：链滴

java爬虫

java爬虫：

主要是用HttpClient模拟浏览器请求第三方站点url

然后响应，获取网页数据，

然后用Jsoup来提取我们需要的信息；

这里首先说一下HttpClient的用法，直接上代码:

```
CloseableHttpClient httpClient=HttpClients.createDefault(); // 创建httpClient实例
HttpGet httpget = new HttpGet("http://www.evafjs.cn/"); // 创建httpget实例
CloseableHttpResponse response=null;
 try {
        response = httpClient.execute(httpget);
    } catch (ClientProtocolException e) {  // http协议异常
        // TODO Auto-generated catch block
        e.printStackTrace();
    } catch (IOException e) { // io异常
        // TODO Auto-generated catch block
        e.printStackTrace();
    } // 执行get请求
 HttpEntity entity=response.getEntity(); // 获取返回实体
    try {
        System.out.println("网页内容：  "+EntityUtils.toString(entity, "utf-8"));
    } catch (ParseException e) {  // 解析异常
        // TODO Auto-generated catch block
        e.printStackTrace();
    } catch (IOException e) { // io异常
        // TODO Auto-generated catch block
        e.printStackTrace();
    } // 指定编码打印网页内容
    try {
        response.close();
    } catch (IOException e) {  // io异常
        // TODO Auto-generated catch block
        e.printStackTrace();
    } // 关闭流和释放系统资源
}
```

这种方法不是万能的，有的服务器会有反爬虫：

系统检测亲不是真人行为，因系统资源限制，我们只能拒绝你的请求。

这时我们就要模拟浏览器设置下User-Agemt头消息：

```
httpGet.setHeader("User-Agent", "Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:50.0) Gecko/2
100101 Firefox/50.0"); // 设置请求头消息User-Agent
```

HttpClient获取响应内容类型Content-Type

原文链接：HttpClient 入门

```
entity.getContentType().getValue()
```

运行输出：

```
Content-Type:text/html; charset=utf-8
```

当然Content-Type还有一堆，那这东西对于我们爬虫有啥用的，我们再>爬取网页的时候 ， 可以通过ontent-Type来提取我们需要爬取的网页 或者是爬取的时候 ,需要过滤掉的一些网页

HttpClient获取响应状态Status

我们HttpClient向服务器请求时，

正常情况 执行成功 返回200状态码，

不一定每次都会请求成功，

比如这个请求地址不存在 返回404

服务器内部报错 返回500

有些服务器有防采集，假如你频繁的采集数据，则返回403 拒绝你请求

```
response.getStatusLine().getStatusCode()
```

运行输出：

```
Status:200
```

代理IP

在爬取网页的时候，有的目标站点有反爬虫机制，对于频繁访问站点以及规则性访问站点的行为，会集屏蔽IP措施

这时候，代理ip就派上用场了；

分享几个代理ip的网站（可用：随缘）

建议使用高匿代理IP 国内代理IP 以及主干道网络大城市的代理IP 访问速度快

http://www.xicidaili.com/

http://www.66ip.cn/

```
HttpHost proxy=``new HttpHost(``"116.226.217.54"``, ``9999``);

RequestConfig requestConfig=RequestConfig.custom().setProxy(proxy).build();

httpGet.setConfig(requestConfig);
```

HttpClient连接超时及读取超时

httpClient在执行具体http请求时候 有一个连接的时间和读取内容的时间；

HttpClient连接时间

所谓连接的时候 是HttpClient发送请求的地方开始到连接上目标url主>机地址的时间，理论上是距离短越快，

线路越通畅越快，但是由于路由复杂交错，往往连接上的时间都不固定，>运气不好连不上，HttpClient的默认连接时间，据我测试，

默认是1分钟，假如超过1分钟 过一会继续尝试连接，这样会有一个问题 >假如遇到一个url老是连不，会影响其他线程的线程进去，说难听点，

就是蹲着茅坑不拉屎。所以我们有必要进行特殊设置，比如设置10秒钟 假如10秒钟没有连接上 我们报错，这样我们就可以进行业务上的处理，

比如我们业务上控制 过会再连接试试看。并且这个特殊url写到log4j日>志里去。方便管理员查看。

HttpClient读取时间

所谓读取的时间 是HttpClient已经连接到了目标服务器，然后进行内容>数据的获取，一般情况 读取据都是很快速的，

但是假如读取的数据量大，或者是目标服务器本身的问题（比如读取数据>库速度慢，并发量大等等..也会影响读取时间。

同上，我们还是需要来特殊设置下，比如设置10秒钟 假如10秒钟还没读>取完，就报错，同上，我可以业务上处理。

HttpClient给我们提供了一个RequestConfig类 专门用于配置参数比如连接时间，读取时间以及前面解的代理IP等。

```
HttpGet httpGet=new HttpGet("http://central.maven.org/maven2/"); // 创建httpget实例
 RequestConfig config=RequestConfig.custom()
.setConnectTimeout(5000)
 .setSocketTimeout(5000)
 .build();
 httpGet.setConfig(config);
```