PYTHON
编程基础

爬虫程序示例

爬虫程序示例

```
1   import re
2   import requests
3   from urllib.parse import quote #导入quote方法对URL中的字符进行编码
4   class BaiduNewsCrawler: #定义BaiduNewsCrawler类
5       headersParameters = { #发送HTTP请求时的HEAD信息
6           'Connection': 'Keep-Alive',
7           'Accept': 'text/html, application/xhtml+xml, */*',
8           'Accept-Language':
                    'en-US,en;q=0.8,zh-Hans-CN;q=0.5,zh-Hans;q=0.3',
9           'Accept-Encoding': 'gzip, deflate',
10          'User-Agent':
                  'Mozilla/6.1 (Windows NT 6.3; WOW64; Trident/7.0; rv:11.0) like Gecko'
11      }
```

爬虫程序示例

```
12    def __init__(self, keyword, timeout): #定义构造方法
13        self.url='http://news.baidu.com/ns?word='
    + quote(keyword) + '&tn=news&from=news&cl=2&rn=20&ct=1' #要爬取的新闻网址
14        self.timeout=timeout #连接超时时间设置（单位：秒）
15    def GetHtml(self): #定义GetHtml方法
16        request=requests.get(self.url, timeout=self.timeout,
                  headers=self.headersParameters)     #根据指定网址爬取网页
17        self.html=request.text #获取新闻网页内容
18    def GetTitles(self): #定义GetTitles方法
19        self.titles = re.findall(r'<h3 class="c-title">([\s\S]*?)</h3>',self.html)   #匹配新闻标题
20        for i in range(len(self.titles)):      #对于每一个标题
21            self.titles[i]=re.sub(r'<[^>]+>',",self.titles[i])     #去除所有HTML标记，即<...>
22            self.titles[i]=self.titles[i].strip()      #将标题两边的空白符去掉
```

爬虫程序示例

```
23      def PrintTitles(self): #定义PrintTitle方法
24          no=1
25          for title in self.titles: #输出标题
26              print(str(no)+':'+title)
27              no+=1
28  if __name__ == '__main__':
29      bnc = BaiduNewsCrawler('南开大学',30) #创建BaiduNewsCrawler类对象
30      bnc.GetHtml() #获取新闻网页的内容
31      bnc.GetTitles() #获取新闻标题
32      bnc.PrintTitles() #输出新闻标题
```

爬虫程序示例

1:南开大学MBA项目2019招生信息
2:南开大学“二次选拔”受欢迎 近八成新生参加考试“自选”成才之路
3:南开大学计算机系到演员 张桐回顾“不安分”的青春
4:南开大学一批教师和集体获得天津市荣誉表彰
5:南开大学校长曹雪涛寄语新生:
6:1亿元!南开大学校友张文中捐巨资支持母校发展
……

爬虫程序示例

提示

因为程序输出的是实时从百度新闻上抓取的新闻标题，所以实际运行该程序时会看到不同的输出结果。

requests模块在使用前需要先安装，可以在系统控制台下输入如下命令完成requests模块的下载和安装：

```
pip install requests -i http://pypi.douban.com/simple --trusted-host=pypi.douban.com
```

爬虫程序示例

用正则表达式对爬取的网页进行分析前，可以先在浏览器下访问网页，按键盘上的F12功能键出现浏览器的调试工具查看页面上的元素；然后查看要获取元素的HTML代码，并根据HTML代码书写正则表达式进行元素匹配。例如，在第19行代码中：

```
<h3 class="c-title">([\s\S]*?)</h3>
```

即是每条新闻标题对应的HTML代码格式。