链滴

# [python 爬虫] 从百度贴吧抓取数据导入到 wordpress

作者：xuwangcheng14

原文链接：https://ld246.com/article/1484994489007

来源网站：链滴

    这是在某人的爬虫脚本的基础上修改而来的，再次感谢!

    使用方法：

    运行下面的python脚本,需要输入几个参数,看提示就明白了,注意在 def getContent里有个键的地方,其中变量dataX是打出的标准贴吧格式的内容演示,sqlStr是导入wordpress的数据库评论表w_comments中(需要你自己改动的地方是comment_post_ID,这个需要自己先写篇文章,再到数据库的w_posts表中查看)。如果你想插入到其它的系统(比如discuz!)中,只要了解下对应系统的数据库模型就了。

<pre>

<pre class="prettyprint lang-py"># -*- coding:utf-8 -*-
import urllib
import urllib2
import re


class Tool:
    removeImg = re.compile('<img.*?>| {7}|')
    removeAddr = re.compile('<a.*?>|</a>')
    replaceLine = re.compile('<tr>|<div>|</div>|</p>')
    replaceTD = re.compile('<td>')
    replacePara = re.compile('<p.*?>')
    replaceBR = re.compile('<br><brr>|<br>')
    removeExtraTag = re.compile('<.*?>')
    removeSpace = re.compile(' ')
    def replace(self,x):
        x = re.sub(self.removeImg,"",x)
        x = re.sub(self.removeAddr,"",x)
        x = re.sub(self.replaceLine,"\n",x)
        x = re.sub(self.replaceTD,"\t",x)
        x = re.sub(self.replacePara,"\n    ",x)
        x = re.sub(self.replaceBR,"\n",x)
        x = re.sub(self.removeExtraTag,"",x)
        x = re.sub(self.removeSpace," ",x)
        return x.strip()



class BDTB:
    def __init__(self,baseURL,seeLZ,floorTag):
        self.baseURL = baseURL
        self.seeLZ = '?see_lz='+str(seeLZ)
        self.tool = Tool()
        self.file = None
```
```

```
        self.floor = 1
        self.defaultTitle = u"百度贴吧" #默认的标题，如果没有成功获取到标题的话则会用这个标题
        self.floorTag = floorTag

    def getPage(self,pageNum):
        try:
            url = self.baseURL+self.seeLZ + '&amp;pn=' + str(pageNum)
            request = urllib2.Request(url)
            response = urllib2.urlopen(request)
            return response.read().decode('utf-8')

        except urllib2.URLError,e:
            if hasattr(e,"reason"):
                print "fail to connect,reason:",e.reason
                return None
    def getTitle(self,page):
        pattern = re.compile('class="core_title_txt pull-left text-overflow  " title="(.*?)" style="',re
S)
        result = re.search(pattern,page)
        if result:
            return result.group(1).strip()
        else:
            return None

    def getPageNum(self,page):
        pattern = re.compile('&lt;li class="l_reply_num".*?&gt;*?&lt;/span&gt;.*?&lt;span.*?&gt;(.
?)&lt;/span&gt;')
        result = re.search(pattern,page)
        if result:
            return result.group(1).strip()
        else:
            return None

    def getContent(self,page):
        pattern = re.compile(ur'&lt;div id="post_content_.*?&gt;(.*?)&lt;/div&gt;.*?&lt;span class
"tail-info"&gt;.*?[\u697c]&lt;/span&gt;&lt;span class="tail-info"&gt;(.*?)&lt;/span&gt;&lt;/div
gt;',re.S)
        items = re.findall(pattern,page)
        pattern_author=re.compile('alog-group="p_author".*?target="_blank"&gt;(.*?)&lt;/a&gt;'

        authors=re.findall(pattern_author,page)
        contents = []
        p=0
        for item in items:
            content = self.tool.replace(item[0])
            date=item[1]
            dataX=content+date+authors[p]
            sqlstr="INSERT INTO `wp_comments` (`comment_ID`, `comment_post_ID`, `comment_
uthor`, `comment_author_email`, `comment_author_url`, `comment_author_IP`, `comment_date
, `comment_date_gmt`, `comment_content`, `comment_karma`, `comment_approved`, `comme
t_agent`, `comment_type`, `comment_parent`, `user_id`) VALUES (null, 51, '"+authors[p]+"', '', '',
'36.32.28.204', '"+date+":00', '"+date+":00', '"+content+"', 0, '1', '', '', 0, 0);\n"
            contents.append(sqlstr.encode('utf-8'))
            p=p+1
```

```
        return contents


    def setFileTitle(self,title):
        if title is not None:
            self.file = open(title + ".txt","w+")
        else:
            self.file = open(self.defaultTitle + ".txt","w+")


    def writeData(self,contents):
        for item in contents:
            if self.floorTag == '1':
                floorLine = "\n" + str(self.floor) + u"-----------------------------------------------------
-----------------------------------\n"
                self.file.write(floorLine)
            self.file.write(item)
            self.floor += 1


    def start(self):
        indexPage = self.getPage(1)
        pageNum = self.getPageNum(indexPage)
        title = self.getTitle(indexPage)
        self.setFileTitle(title)
        if pageNum == None:
            print "URL已失效，请重试"
            return
        try:
            print "该帖子共有" + str(pageNum) + "页"
            for i in range(1,int(pageNum)+1):
                print "正在写入第" + str(i) + "页数据"
                page = self.getPage(i)
                contents = self.getContent(page)
                self.writeData(contents)
        except IOError,e:
            print "写入异常，原因：" + e.message
        finally:
            print "写入任务完成！！！"



print u"请输入帖子代号"
baseURL = 'http://tieba.baidu.com/p/' + str(raw_input(u'http://tieba.baidu.com/p/'))
seeLZ = raw_input("是否只获取楼主发言，是输入1，否输入0\n")
floorTag = raw_input("是否写入楼层信息，是输入1，否输入0\n")
bdtb = BDTB(baseURL,seeLZ,floorTag)
bdtb.start()</pre>

<br />
</pre>
```