

【python爬虫】获取百度贴吧帖子的python脚本

作者: [xuwangcheng14](#)

原文链接: <https://ld246.com/article/1446637260304>

来源网站: [链滴](#)

许可协议: [署名-相同方式共享 4.0 国际 \(CC BY-SA 4.0\)](#)

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

```
class Tool:
```

```
    removeImg = re.compile('&lt;img.*?&gt;| {7}|')
    removeAddr = re.compile('&lt;a.*?&gt;|&lt;/a&gt;')
    replaceLine = re.compile('&lt;tr&gt;|&lt;div&gt;|&lt;/div&gt;|&lt;/p&gt;')
    replaceTD = re.compile('&lt;td&gt;')
    replacePara = re.compile('&lt;p.*?&gt;')
    replaceBR = re.compile('&lt;br&gt;&lt;brr&gt;|&lt;br&gt;')
    removeExtraTag = re.compile('&lt;.*?&gt;')
    removeSpace = re.compile('&nbsp;')
    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('<li class="l_reply_num".*?>*</span>.*?<span.*?>(</span>?)</span>')
    result = re.search(pattern,page)
    if result:
        return result.group(1).strip()
    else:
        return None

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

    authors=re.findall(pattern_author,page)
    contents = []
    p=0
    for item in items:
        content = "\n"+self.tool.replace(item[0])+"\n"
        date="\n"+item[1)+"\n"
        dataX=content+date+authors[p]
        contents.append(dataX.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>
```