# [Python学习] 简单爬取CSDN下载资源信息



这是一篇Python爬取CSDN下载资源信息的例子，主要是通过urllib2获取CSDN某个人所有资源的资源URL、资源名称、下载次数、分数等信息；写这篇文章的原因是我想获取自己的资源所有的评论信息，但是由于评论采用JS临时加载，所以这篇文章先简单介绍如何人工分析HTML页面爬取信息。

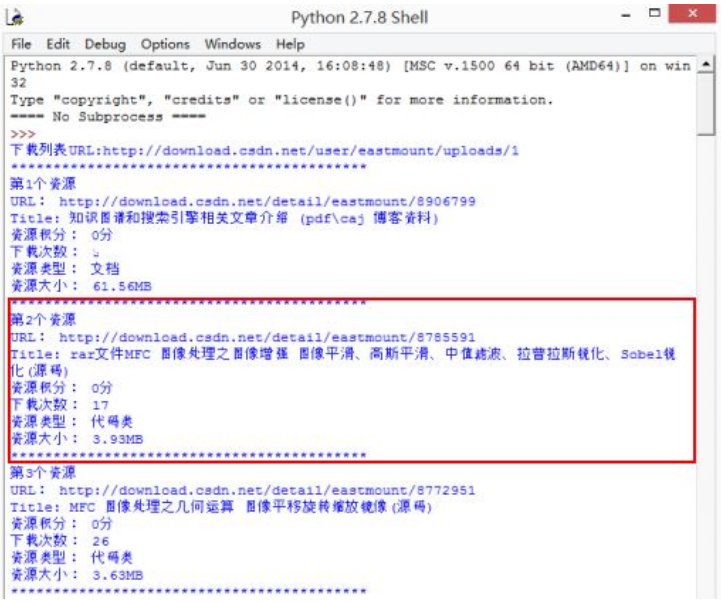
## 源代码

```
# coding=utf-8
import urllib
import time
import re
import os

#************************************************
#第一步 遍历获取每页对应主题的URL
#http://download.csdn.net/user/eastmount/uploads/1
#http://download.csdn.net/user/eastmount/uploads/8
#************************************************

num=1 #记录资源总数 共46个资源
number=1 #记录列表总数1-8
fileurl=open('csdn_url.txt','w+')
fileurl.write('***************获取资源URL************\n\n')

while number<9:
    url='http://download.csdn.net/user/eastmount/uploads/' + str(number)
    fileurl.write('下载列表URL:'+url+'\n\n')
    print unicode('下载列表URL:'+url,'utf-8')
    content=urllib.urlopen(url).read()
    open('csdn.html','w+').write(content)

    #获取包含URL块内容 匹配需要计算</div>个数
    start=content.find(r'<div class="list-container mb-bg">')
    end=content.find(r'<div class="page_nav">')
    cutcontent=content[start:end]
    #print cutcontent

    #获取块内容中URL
    #形如<dt><div><img 图标></div><h3><a href>标题</a></h3></dt>
    res_dt = r'<dt>(.*?)</dt>'
    m_dt =  re.findall(res_dt,cutcontent,re.S|re.M)
    for obj in m_dt:
        #记录URL数量
        print '*****************************************'
        print '第'+str(num)+'个资源'
        fileurl.write('*****************************************\n')
        fileurl.write('第'+str(num)+'个资源\n')
        num = num +1
        #获取具体URL
        url_list = re.findall(r"(?<=href=\").+?(?=\")|(?<=href=\').+?(?=\')", obj)
```

```
        for url in url_list:
                                        url_load='http://download.csdn.net'+url
            print 'URL:  '+url_load
            fileurl.write('URL:  http://download.csdn.net'+url+'\n')
        #获取资源标题
        #<a href="/detail/eastmount/8757243">MFC显示BMP图片</a>
        res_title = r'<a href=.*?>(.*?)</a>'
        title = re.findall(res_title,obj,re.S|re.M)
        for t in title:
            print unicode('Title: ' + t,'utf-8')
            fileurl.write('Title:  ' + t +'\n')


        #************************************************
        #第二步 遍历具体资源的内容及评论
        #http://download.csdn.net/detail/eastmount/8785591
        #************************************************

        #定位指定结构化信息盒Infobox
        resources = urllib.urlopen(url_load).read()
        open('resource.html','w+').write(resources)
        start_res=resources.find(r'<div class="wraper-info">')
        end_res=resources.find(r'<div class="enter-link">')
        infobox=resources[start_res:end_res]

        #获取资源积分、下载次数、资源类型、资源大小(前4个<span></span>)
        res_span = r'<span>(.*?)</span>'
        m_span = re.findall(res_span,infobox,re.S|re.M)
        print '资源积分:  '+m_span[0]
        fileurl.write('资源积分: ' + m_span[0] +'\n')
        print '下载次数:  '+m_span[1]
        fileurl.write('下载次数: ' + m_span[1] +'\n')
        print '资源类型:  '+m_span[2]
        fileurl.write('资源类型: ' + m_span[2] +'\n')
        print '资源大小:  '+m_span[3]
        fileurl.write('资源大小: ' + m_span[3] +'\n')


        #************************************************
        #第三步 如何获取评论
        #http://jeanphix.me/Ghost.py/
        #http://segmentfault.com/q/1010000000143340
        #http://casperjs.org/
        #************************************************


    else:
        fileurl.write('*****************************************\n\n')
        print '*****************************************\n'
        print 'Load Next List\n'
        number = number+1 #列表加1
#退出所有循环
else:
    fileurl.close()
```

## 显示结果

显示内容包括资源URL、资源标题、资源积分、下载次数、资源类型和资源大小：




比如现在爬取郭霖大神的资源信息，其中页面链接如下：(共7页)

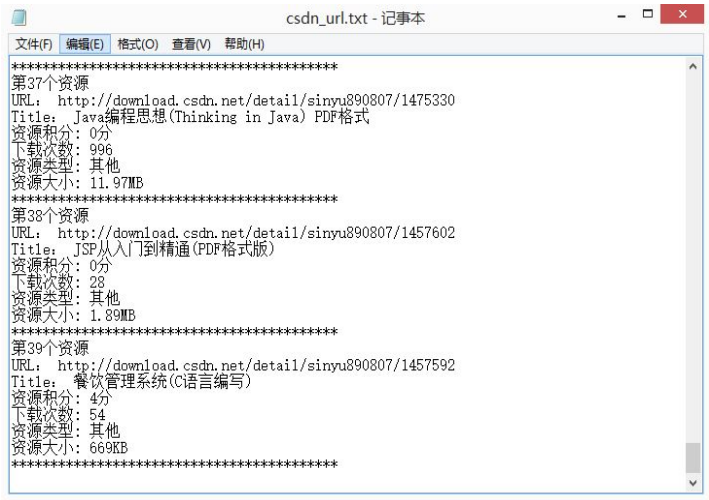
http://download.csdn.net/user/sinyu890807/uploads/1

http://download.csdn.net/user/sinyu890807/uploads/7

简单修改Python源代码URL后，下载页面如下图所示：


运行结果如下图所示：

```
>>>
下载列表URL:http://download.csdn.net/user/sinyu890807/uploads/1
****************************************
第1个资源
URL： http://download.csdn.net/detail/sinyu890807/7754669
Title: Android照片结合LruCache和DiskLruCache Demo
资源积分： 0分
下载次数： 4161
资源类型： 代码类
资源大小： 1.06MB
****************************************
第2个资源
URL： http://download.csdn.net/detail/sinyu890807/7749729
Title: 《第一行代码——Android》试读
资源积分： 0分
下载次数： 512
资源类型： 文档
资源大小： 1.55MB
****************************************
第3个资源
URL： http://download.csdn.net/detail/sinyu890807/7747691
Title: Android第一行代码源码
资源积分： 0分
下载次数： 7246
资源类型： 代码类
资源大小： 46.44MB
****************************************
```




## HTML分析

首先，获取每列中的所有资源的URL和标题，通过分析源代码。

```
<dt>
   <div class="icon"><img src="/images/minetype/rar.gif" title="rar文件"></div>
   <div class="btns"></div>
   <h3><a href="/detail/eastmount/8772951">
          MFC 图像处理之几何运算 图像平移旋转缩放镜像(源码)</a>
       <span class="points">0</span>
   </h3>
</dt>
<dd class="meta">上传者:
    <a class="user_name" href="/user/eastmount">eastmount</a>
         | 上传时间：2015-06-04
         | 下载26次
</dd>
<dd class="intro">
         该资源主要参考我的博客【数字图像处理】六.MFC空间几何变换之图像平移、镜像、旋转
         缩放详解，主要讲述基于VC++6.0 MFC图像处理的应用知识，要通过MFC单文档视图实现显
         示BMP图片。
</dd>
<dd class="tag">
     <a href="/tag/MFC">MFC</a>
     <a href="/tag/%E5%9B%BE%E5%83%8F%E5%A4%84%E7%90%86">图像处理</a><
```

```
</dd>
```

对应的HTML显示如下图所示：


然后通过URL去到具体的资源获取我自己称为像消息盒的信息：


对应审查元素的信息如下所示，获取0分即可：

最后我想做的事获取评论信息，但是它是通过JS实现的：

```
<div class="section-list panel panel-default">
   <div class="panel-heading">
      <h3 class="panel-title">资源评论</h3>
   </div>
   <!-- recommand -->
   <script language='JavaScript' defer type='text/javascript'

src='/js/comment.js'></script>
   <div class="recommand download_comment panel-body" sourceid="8772951"></div>
</div>
```

显示的JS页面部分如下：

```
var base_url= (window.location.host.substring(0,5)=='local') ? 'http://local.downloadv3.csdn.net' :
'http://download.csdn.net';
base_url = ""; $(document).ready(function(){

        CC_Comment.initConfig();
        CC_Comment.getContent(1);
});
var CC_Comment =
{
        sourceid:0,
        initConfig:function()
        {
                var sid = parseInt($(".download_comment").attr('sourceid'));
                if(isNaN(sid) || sid<=0)
                {
                        this.sourceid = 0;
                }else
                {
                        this.sourceid = sid;
                }

        }
....
}
```

最后希望文章对你有所帮助吧！下一篇准备分析下Python如何获取JS的评论信息，同时该篇文章可以给你提供一种简单的人工分析页面的例子；也可以获取某个人CSDN资源下载多、分数高的给你挑选。基础知识，仅供参考~

(By:Eastmount 2015-7-21 下午5点 http://blog.csdn.net/eastmount/)



Eastmount 博客专家
发布了444 篇原创文章 · 获赞 5918 · 访问量 484万+