[python爬虫] BeautifulSoup爬取+CSV存储贵州农产品数据

原创 Eastmount 最后发布于2017-10-29 23:29:31 阅读数 7049 ☆ 收藏 展开

在学习使用正则表达式、BeautifulSoup技术或Selenium技术爬取网络数据过程中，通常会将爬取的数据存储至TXT文件中，前面也讲述过海量数据存储至本地MySQL数据库中，这里主要补充BeautifulSoup爬取贵州农产品数据的过程，并存储至本地的CSV文件。

核心内容包括以下几点：

1.如何调用BeautifulSoup爬取网页数据。
2.如何存储数据至CSV文件。
3.如何解决中文字符存储的乱码问题，UTF8编码设置。
4.如何定时设置爬取任务，定时截图保存。

基础文章希望对大家有所帮助，尤其是刚入门学习BeautifulSoup爬虫知识，或者是遇到如何将中文数据存储至CSV文件下的同学。如果文章中存在错误或不足之处，还请海涵~

一. Python操作CSV库

CSV（Comma-Separated Values）是常用的存储文件，逗号分隔符，值与值之间用分号分隔。Python中导入CSV扩展包即可使用，包括写入文件和读取文件。

1.写入CSV文件

```
# -*- coding: utf-8 -*-
import csv
c = open("test-01.csv", "wb")  #写文件
writer = csv.writer(c)
writer.writerow(['序号','姓名','年龄'])

tlist = []
tlist.append("1")
tlist.append("小明")
tlist.append("10")
writer.writerow(tlist)
print tlist,type(tlist)

del tlist[:]  #清空
tlist.append("2")
tlist.append("小红")
tlist.append("9")
writer.writerow(tlist)
print tlist,type(tlist)

c.close()
```

其中writerow用于写文件，这里增加列表list写入。输出如下图所示：

	A	B	C	D
1	序号	姓名	年龄	
2	1	小明	10	
3	2	小红	9	
4				

2.读取CSV文件

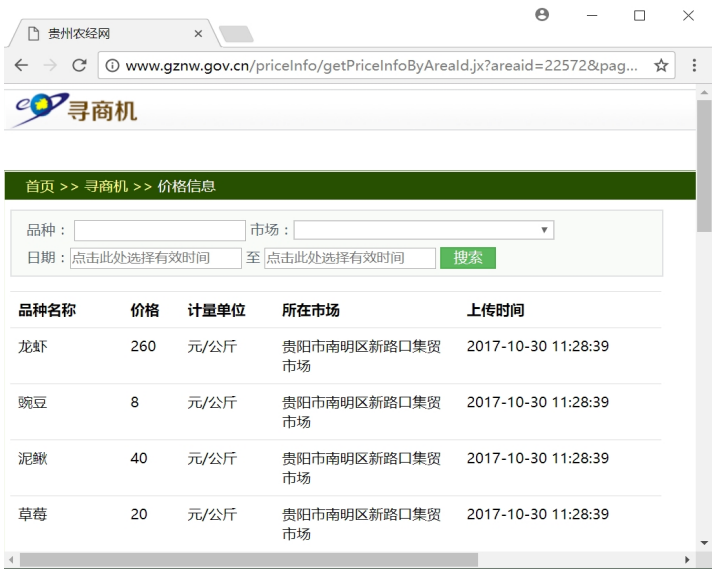
```
# -*- coding: utf-8 -*-
import csv
c = open("test-01.csv", "rb")  #读文件
reader = csv.reader(c)
for line in reader:
    print line[0],line[1],line[2]
c.close()
```

输出结果如下：

```
序号 姓名 年龄
1 小明 10
2 小红 9
```

二. BeautifulSoup爬取贵州农经网

打开贵州农经网可以看到每天各类农产品的价格及产地更新数据，如下图所示。
网址：http://www.gznw.gov.cn/priceInfo/getPriceInfoByAreaId.jx?areaid=22572&page=1



现在需要通过BeautifulSoup获取 产品、价格、单位、产地和发布时间五个信息，通过浏览器“审查元素”可以看到每行数据都位于<tr>节点下，其class属性为“odd gradeX”，调用find_all函数即可获取。

爬取第一页的代码如下：

```
# -*- coding: utf-8 -*-
import urllib
from bs4 import BeautifulSoup

i = 1
while i<5:
    print "爬取第" + str(i) + "页"
    url = "http://www.gznw.gov.cn/priceInfo/getPriceInfoByAreaId.jx?areaid=22572&page=" + str(i)
    print url
    content = urllib.urlopen(url).read()
    soup = BeautifulSoup(content, "html.parser")
    print soup.title.get_text()

    num = soup.find_all("tr",class_="odd gradeX")
    for n in num:
        con = n.get_text()
        num = con.splitlines()
        print num[1],num[2],num[3],num[4],num[5]

    i = i + 1
```

注意以下几点：

1.由于网页涉及到翻页，通过分析URL发现，与"page=xxx"相关，则定义循环爬取1-5页的内容；

2.BeautifulSoup相当于将爬取的网页解析成树状结构，调用soup.title可以得到<title>xxxx</title>内容，再调用get_text()函数获取值；

3.核心代码是通过num = soup.find_all("tr",class_="odd gradeX")找到内容；

4.通过splitlines()删除换行，并生成序列，再依次获取内容num[n]。

输出如下所示：

```
爬取第1页
http://www.gznw.gov.cn/priceInfo/getPriceInfoByAreaId.jx?areaid=22572&page=1
贵州农经网
龙虾 260 元/公斤 贵阳市南明区新路口集贸市场 2017-10-30 11:28:39
豌豆 8 元/公斤 贵阳市南明区新路口集贸市场 2017-10-30 11:28:39
泥鳅 40 元/公斤 贵阳市南明区新路口集贸市场 2017-10-30 11:28:39
```

```
··· 爬取第2页
http://www.gznw.gov.cn/priceInfo/getPriceInfoByAreaId.jx?areaid=22572&page=2
贵州农经网
石斑鱼 60 元/公斤 贵阳市南明区新路口集贸市场 2017-10-30 11:28:39
紫菜（干） 34 元/公斤 贵阳市南明区新路口集贸市场 2017-10-30 11:28:39
纽荷尔 14 元/公斤 贵阳市南明区新路口集贸市场 2017-10-30 11:28:39
黄鳝 70 元/公斤 贵阳市南明区新路口集贸市场 2017-10-30 11:28:39
...
```

接下来需要将内容存储至CSV文件中，这里最容易出现的错误是中文乱码的问题。一方面，需要将创建的CSV文件设置为UTF-8编码，另一方面需要调用encode('utf-8')函数转化为中文编码方式，写入文件。代码如下：

```
# -*- coding: utf-8 -*-
"""
Created on Fri Oct 20 20:07:36 2017

@author: eastmount CSDN 杨秀璋
"""

import urllib
from bs4 import BeautifulSoup
import csv
import codecs

c = open("test.csv","wb")    #创建文件
c.write(codecs.BOM_UTF8)     #防止乱码
writer = csv.writer(c)       #写入对象
writer.writerow(['产品','价格','单位','批发地','时间'])

i = 1
while i <= 4:
    print "爬取第" + str(i) + "页"
    url = "http://www.gznw.gov.cn/priceInfo/getPriceInfoByAreaId.jx?areaid=22572&page=" + str(i)
    content = urllib.urlopen(url).read()
    soup = BeautifulSoup(content,"html.parser")
    print soup.title.get_text()
    tt = soup.find_all("tr",class_="odd gradeX")
    for t in tt:
        content = t.get_text()
        num = content.splitlines()
        print num[0],num[1],num[2],num[3],num[4],num[5]
        #写入文件
        templist = []
        num[1] =  num[1].encode('utf-8')
        num[2] =  num[2].encode('utf-8')
        num[3] =  num[3].encode('utf-8')
        num[4] =  num[4].encode('utf-8')
        num[5] =  num[5].encode('utf-8')
        templist.append(num[1])
        templist.append(num[2])
        templist.append(num[3])
        templist.append(num[4])
        templist.append(num[5])
        #print templist
        writer.writerow(templist)
    i = i + 1

c.close()
```

输出如下所示，写入CSV文件下载前4页内容。

	A	B	C	D	E
1	产品	价格	单位	批发地	时间
2	龙虾	260	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
3	豌豆	8	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
4	泥鳅	40	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
5	草莓	20	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
6	火龙果	12	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
7	海带（干）	36	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
8	牛蛙	40	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
9	车厘子	140	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
10	墨鱼	70	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
11	雪梨	8	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
12	石斑鱼	60	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
13	紫菜（干）	34	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
14	纽荷尔	14	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
15	黄鳝	70	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
16	大黄鱼	40	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
17	绿豆芽	5	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
18	甲鱼	50	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
19	哈密瓜	6	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
20	木瓜	16	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
21	香瓜	7	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
22	干辣椒	28	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
23	洋鸡蛋	14	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28
24	鲈鱼	48	元/公斤	贵阳市南明区新路口集贸市场	2017/10/30 11:28

如果在增加一个定时机制，每天定时爬取就非常完美了。最后补充一下Anaconda制作的定时任务。

三. Python设置定时任务截图

通常可以使用系统的任务来实现，比如WPS、QQ等软件自动更新，或定时提醒，这里只需要每天定时执行爬虫代码即可，后面讲写一篇文章详细介绍。
参考：http://blog.csdn.net/wwy11/article/details/51100432

下面是每隔10秒钟打开网页，然后进行截屏的操作。作为在线备份的代码，仅供参考。

```
# -*- coding:utf-8 -*-
from PIL import ImageGrab
import webbrowser as web
import time
import os

#定时15分钟
def sleeptime(hour,min,sec):
    return hour*3600 + min*60 + sec

second = sleeptime(0,0,10)
j = 1
while j==1:
    time.sleep(second)
    #打开网页
    url = ["https://www.gzzzb.gov.cn/","http://www.gznw.gov.cn/"]
    i = 1
    while i<=len(url):
        web.open_new_tab(url[i-1])
        time.sleep(5)
        im = ImageGrab.grab()
        im.save(str(j)+'.jpg','JPEG')#图片保存
        i = i+1
        j = j+1
```

《勿忘心安》
勿要把酒倚寒窗，庭院枯叶已飞霜。
忘怀之前坎坷路，劝君一醉付流光。
心中愁苦漫翻滚，雪上寒鸦入画堂。
安知我辈庸庸过，双鬓飞白亦疏狂。

很喜欢这首诗，也享受在公交车上备课的日子，心很静很安，更享受和期待新装修的新家，人生漫漫，还是带着一丝微笑和她前行。接下来再忙还是挤点时间看看分布式爬虫和深度学习，十月这个节点终于结束啦。学生的笔记不错，有我的风范，大家也很认真。
Remember you are born to live. Don’ t live because you are born! Don't go the way life takes you.Take life the way you go! Follow my heart and nana's footsteps forever.

最后希望这篇文章对你有所帮助。
(By:Eastmount 2017-10-30 18:00 http://blog.csdn.net/eastmount/)