Day01

2017年7月6日 星期四

11:00

在Notepad中将多行合并为一行并加上分号

替换---"\r\n"---";"

\*\*\*Python爬取美团视频链接

美团美食主页面--->http://bj.meituan.com/category/meishi

1.先抓北京的所有区http://bj.meituan.com/category/meishi

2.抓每个区下的所有商圈http://bj.meituan.com/category/meishi/**chaoyangqu**

3.获取指定商圈下面的所有商店http://bj.meituan.com/category/meishi/**anzhen**

4.最后在指定的商店下获取商品http://bj.meituan.com/shop/**5673702**

问题：在获取指定商圈下面的所有商店时只获取到两个商店，与实际页面不符

猜测1：ajax请求，滑轮滚动到时，发送一次ajax请求，每次获取两个商店数据

解决：然而并不是

猜测2：动态加载

解决：使用动态代理

猜测3：以上都不是

解决：

分析：打印Jsoup.connction("http://bj.meituan.com/category/meishi/beijingdongzhan"")

.get()//拿到的后面动态加载的商店的<a></a >被转义成了&lt;a&gt;……&lt;/a&gt;

  用.excute().body()//拿到的没有上述问题，但是这个返回值是String类型的，

  而我们要的是Document对象，通过这个对象才能去select()

  方法：

  html()和outerHtml()的区别只是有没有包含本层而已，最终底层方法是一样的，所以这里确实转义了。

 其实也说得通，因为textarea里的内容是“文本”，html里的纯文本内容，如果不进行转义是不安全的。

     当然这里你的需求是"保持原文"，“转义再反转”其实是无法保持原文的。而且Apache的StringEscapeUtils的转义范围比Jsoup要小，所以其实反转是不完整的。

     有个比较hack的方法：

     Document doc = Jsoup.connect("http://sports.163.com/13/0830/22/97IFSI5I00051CD5.html").get();

    //清空jsoup的转义表，会使jsoup失去转义能力

Entities.EscapeMode.base.getMap().clear();

Elements elements = doc.select("textarea[id^=photoList]");

for(Element e:elements){

System.out.println(e.html());

}

猜测4：问题还是得不到解决

最终解决方案：指定商圈下的所有商店id都以jsoup串的形式存放在div的一个属性当中

cdiv 
v (div data-component=" pal -list" 
mt-component--booted" 
v cdiv cf 90i-Iist--nodeaI Iag-mod-viewed" 
v;ediv cf 3 -hub" data-hubmodule="www-index" . PoiList" " 
5673722, 316757, 72222,2532369, 
N/index\/$ 
4339577 , 41839271, 264121, 1652165 , 42377617 , 362107 , 3248848 , , 1651402, 1162886, , 1532152, 327012, 298572, , 155791879, 5933039 
4175918e , , 72774125 , 42341182 , 274666 , 264073 , 369462, 9736483e , 80410485 , 6778236 , 66935442, 66842639, 282096 , 4523121, , 2629 
, 263772, 6592834 , 331797 , 288934, 275868 , 376623 , 1651404, 94534643 , , 5817912 , 1651a22, , 5967228 , 77141657 , 2746 
27522e 4848953 288218 287160 66247872 I': 
: V' 26451683 26452431\" 
, \ "acm\" : 
, 3865091 IN" , : "AcategoryC49226973e8781121688.41722ea . , : 
, " : \ " , V' : , N " : ) , : \ \ "acm\": 
A "AcetegoryC4922S97308781121688.2530369.30\ " , \ : \ " N" , \ con : X" thumbN " , N" \ " : N : \ / bil ling- 
. neituan . com\N X/ \ N N,'cpvN ; ; ; ; ; 
fit nc ; = ; ; bg ; slot= 50028&emp ; mg ; ; 
ssae8e3 .1499275022. a. a. ; 
ent i ; i ns sho ; c t ; 
; ; di ; " , "mtad_cli " : N "https : \ W X/ bil ling- 
. neituan . corrAN X/ \ N Wc pv\, 'bi II ing ; un ; bu ; = ; ; ; 
fit nc i ; category_ids = I&amp ; ; slot= 500288emp ; mg ; I&amp ; 
tntdpid=ab04e7sa4a8c1 ssae8e3 .1499275022. a. a. ; 
ent i ; i ns sho ; c t ; 
ad=2'34555188amp ; di _id=5ea288emp ; " } , { \ " " : 
"28911963 , 28911981, 28911375, , : "Acategoryc49226973e8781121S88.251188a.4N" , : , \ : 

过程：通过谷歌自带的网站开发者工具（F12）中的NetWork来分析网页请求：

1.输入网址：http://bj.meituan.com/category/meishi/anzhen(这就北京朝阳区安贞商圈下的所有的商店呈现页面)

X * faststoneca X 
C t bj.meituan.com/category/meishi/anzh@ 
IOA 
*fifiiä] 
fitß[Xiä] 
Gift card 
E Network 
#ZRE-äNetWorK 
lements Console Sources Network 
ne Profiles Applicatdn 
Security Audits 
Preserve log O Dis cac 
data URLs XHR 
css 
500 ms 
Font Doc WS 
2500 ms 
fe.co... 
Manifest Other 
3000 ms 
Time 
(from cac e 
(from cache) 
(from cache) 
(from cache) 
(from cache) 
(from cache) 
(from cache) 
(from cache) 
(from cache) 
(from cache) 
(from cache) 
3500 ms 
4000 ms 
4500 ms 
2.00 s 
5000 ms 
5500 ms 
400 s 
6000 ms 
Name 
— require m 
analytics.js 
loader/ mt-s... 
?f=fewww:vmn.v/Js/tips.js@gcb1008 
IL / 54 requests 08 / 32.1 transferred Finish: 
Status 
200 
200 
200 
200 
200 
200 
5.55 s 
imeline — Start Time 
3.00 s 
5.00 s 
script 
script 
script 
script 
script 
script 
sc ript 
script 
script 
script 
anzhen:60 
anzhen:60 
anzhen:36 
2 ms 
O ms 
3 ms 
2 ms 
I ms 
2 ms 
4 ms 
3 ms 
I ms 
g ms 
521 ms 2336: ms 

2滑动滑块到一定位置，网页发起动态加载请求，具体如下：

Name 
L] poilist 
64Dw_2e_. 
15 requests 3.1 KB transferred 
poilist 
5 requests 3.1 KB transferred 
X Headers Preview Response Cookies Timing 
enera 
Request URL: http: //b .meituan.con/_async . gif?r 
goryc1a9714ø8185173a68944 399287 B2CAcategoryC1e9714e8185173es 
44.27725e 
oryc1e9714e8185173a68944 6sese1 12%2CAcategoryC1e9714e8185173e68944.279sas.1 
51730689 &813459.1 2CAcategoryC1e9714e81851 3268944 4984473 
or-ycleg 
185173068 
goryc1e9714e8 
6%2CAcategory 
097140818 
54243.17%2CAcategoryC1a9714a81 
25, 7 "2 
nei bf . me it u 
ego ryc1a9714a8185173e68944.41839271.19 
i%2Fc ha Oya ngq u*3F mttS3DI . indexS252Fdefau 
2Fdefau1t%2Fpoi 
Request Method: GET 
Status Code. 200 OK 
Remote Address: 103.37.152.58:8 
ponse ea ers 
Request Headers (8) 
x 
recv: Acat 
0581. 
cle9714a8185173e68944 393287 e,Acategoryc109714e8185173a€8944 77 zse 1,Acategory 
ategoryc1a9714ø8185173a68944.27905.13 , AcategoryC1e9714081851730689 
59.14, Acateg 
X 
984273 
77.18 or cle9714e818s173e68944.41839271.19 
u : (category,'melshl/anzhen 
Headers Preview Response Cookies Timing 
Host: bf.meituan.com 
Origin: http://bj . meituan . com 
Referer: http://bi . meituan . . .ø. .34rvnesw 
User-Agent: Mozi11a/5.e (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/52.e.2743.116 safari/537.36 
X-Requested-With: XMLHttpRequest 
poiidList: 5754243, , 41839271] 
poiData: "acm" : "AcategoryC1e9714e8185173e68944.43 
tad icon" : 
7250.11" , : 
"45924464 
", "acm": "Acategor 
"mtad icon":"" 
485178" , "acm" : "Acategoryc1e9714e8185173a68944.16sase1.12" , "dealacm 
" "mtad icon": " 
, "mtad click I 
"mtad show I 
og_url 
og_url 
tad icon": 
" {"dealid" : "28396152", : "Acategoryc109714e8185173a68944.2796e5.13 , 
" "dealacm" : 
"mtad icon": 
"45647814" , "acm": "AcategoryC1e9714e8185173e68944.4813459.14", "dealacm" " 
h {"dealid": "30938019", "AcategoryC1e9714e8185173e68944.49 • 

3.于是，就去大胆尝试了一下

3.1首先查看其中一个商店的ID

bj.meituan.com/category/meishi/anzhen?mtt=1.index%2Fdefault%2Fpoi.O.O.j4n.mesw 
o 
164 A 
Gin Caut 
14k 
Elements Console Sources Network Timeline Profiles Application Security Audits 
\ "AcategoryC1e9714e8185173e68944.66047872.80\" , \ "dealacm\" : \ " i" , \ : , \ " : 
\ , \ : \ "limit\" : 10, \ "biglmageMode\" : true}" , " limit" : 10, "initialLoad" : 10}" data-hubstamF 
(div nodeal log-acm-viewed" data-mtd- 
showlog data -mtd-clicklog 
J-mtad-link" href="http://bj.meituan.com/shop/5673702? 
acm=lhunyaiIsWIe9714e8185173e68944.5673702. . pz . I. :bdw" target= 
blank" data-mod-mtt= 
Styles Computed 
element. style { 
. poi -tile-nodeal 
. basic . link { 
float: left; 
Qent Listeners 
DOM Breakpoints Properties 
:hov 
. cls 
. poi -tile 
info 
(div 
(div cf" 
icon-shangiia' 
a class link f3 J -mtad-link 
: : after 
(div 
href 
http : /bj . . 
'56737m' 
target= " 
_blank A 
e e 32px; 
text-align: left; 
max-width: 22epx; 
overflow: hidden; 
text-overflow: ellipsis; 
white-space: nowrap; 
color: •#666; 
f3_b { 
font-size: 16px 
?f=fewww: /www/cmss/deallist.„:I 
?f=fewww: /nwWc 
3 16 3014 3 16 0 
1499311996470 2947 
16 0 
149931196470 3013 
div 
div 
a.Iink.f3J-mtad-Iink 

3.2在2中找一个ID按照3.1中的②的链接拼一个，例如：http://bj.meituan.com/shop/1650501，然后去浏览器访问

安 页 姜 直 一 北 京 贞 近 # × 蒂 人 马 车 一 人 均 价 一 营 业 
新 过 西 卉 甬 民 豆 矣 
c 》 bj.meituan.com 
/ shop ／ 1650501 
我 的 订 单 
美 团 」 諒 
团 ， 
谲 输 入 商 品 名 称 、 地 址 等 
忉 换 城 市 
meltl_」an.COm 
比 格 叶 多 美 虾 吃 虾 訣 析 烤 肉 
全 部 分 类 
北 京 团 美 食 团 新 疆 朝 阳 区 贞 
我 的 美 团 
最 近 
联 系 客 
扌 叟 索 
必 胜 客 
新 疆 
信 掘 
我 是 商 家 
评 1 飠 人 訕 911 
商 家 资 质 
代 理 商 招 菱 
不 满 意 免 单 
品 全 档 案 
更 多 
V 
苜 页 
今 日 新 单 
美 食 
新 疆 查 西 来 甫 民 会 厅 
新 疆 麦 西 来 甫 民 族 宴 会 厅 
朝 阝 日 区 和 平 街 青 年 沟 路 13 区 1 号 （ 煤 炭 大 厦 东 侧 ） 
010 一 56210099 
蛋 糕 东 来 颠 好 伦 
身 边 外 卖 
4 ． 7 分 
0 
囗 
回 《 
庸 囗 
下 载 美 团 手 机 
专 車 分 抵 1 囗 亓 
4 
已 1245 
双 人 餐 
已 g84 
亓 代 金 券 1 张 ， 可 叠 加 
截 止 到 2D17m7 · 22 
截 止 到 2D17m7 · 22 
截 止 到 2D17m7 · 28 
立 即 抢 
门 店 1 飠 
立 即 抢 
门 店 1 飠 料 
EPIJ 
立 即 抢 
口 《 、 VI nn 
内 存 會 
63 
囗 K/S 
囗 K/S 

3.3去安贞首页验证这家店是否存在

x 
C t bj.meituan.com/category/meishi/anzhen 
911 
ttB8Rfi 
*ÆttF 
2653Ai*ffi 
v 
X 
mg mun 
Elements Console Sources Network Timeline Profiles Application Security Audits 
mtd-showlog 
<div " 
clicklog 
> J-mtad-link" href="http://ww.meituan .com/shop/165e5aI 
<div 
icon-shangiia 
a class link f3 J-mtad-link nref http://wx.meituan.cm/shop A65ø5øI? 
index%2Fdefa. - . 
Styles Computed 
element. style { 
. poi -tile-nodeal 
Qent Listeners 
data-mtd-showlog data-mtd- 
h : Ode.'" 
target= 
DOM Breakpoints Properties 
: hov 
. cls 
. poi -tile 
blank" 
'span class:' 
: : after 
(div 
(div 
3 16 3525 Syui_3 16_0 3524 
data-mod-mtt= 
. basic . link:hover { 
text-decoration: none; 
•#333; 
color: 
. poi-tile-nodeal .poi-tile 
. basic .1ink { 
float: left; 
e e 32px; 
text-align: left; 
max-width: 22epx; 
overflow: hidden; 
text-overflow: ellipsis; 
info 
info 
?f=fewww: /www/cmss/deallist.„:I 
?f=fewww: /www,'c...ss/deallist...:l 

3.4那么问题来了既然ID都在这数组中，怎么样才能拿到？答：通过之前失败的案例中抓取的网页来查找关键词，最终找到。

获取出的json有点不对的地方，即json没有正确嵌套导致不能正确嵌套

ÉE.txt - 
{"acti on" : 
[99228003, 68344983, 110977328, 1170527 
17103436], 
9922 
og _ url : 
6834 
show log ur cl 
\ "4498S020, Ä159ö825, 41601468, 4602636 
al 685296, 4170 
1170 
show log ur cl 
\ "44626903, Ä421ö767, 44211238, 4421194 
\ "deal 
'AcategoryC5224215364 
og _ url \ " . 
'AcategoryC5224215364 
og _ url \ " . 
'29036921, 35191078, 44 
'AcategoryC5224215364 
\"mtad_click_l 
'AcategoryC5224215364 
\"mtad_click_l 
'AcategoryC5224215364 
og _ url \ " . 
'45692630, 45693728, 45 
'43886761, Ä3886126\", 
'big ImageMode\" : true]' 
data': 
, 112420850, 77159167, 117440044, 372213, 312612, 1539906, 6032267, 1 
39274460, 39274710, 44465539\", 
003. 1\", 
983. 2\", 
V', : 110977328. 3\", 
og _ url \ " : ck_l og _ url \ " : 
17, 42728812\", 
2767. 4\", 
V', : 112420850. 5\", 
og _ url \ " : ck_l og _ url \ " : 
699, 42442767, 38132361\", 
2426608. 77159167. 6\", 
43291334 \ 
2426608. 117440044. 7\", 
5105, 45930991, 35191472\", 
2426608. 372213. 8\", : : 
ur 
2426608. 312612. 9\", : : 
ur 
2426608. 1539906. 10\", 
8651 V', : 6032267. 11 V', \ "dealacm\" : 
acm\" : 117103436. 12\", \ "deal : 
\ "limi t \ " : 10, 
':10, "initial Load": 10] 
limi t. 

修正会后的json如下

{"action":"\/index\/poilist","data":{"poiidList":[99228003,68344983,110977328,117052767,112420850,77159167,117440044,372213,312612,1539906,6032267,117103436],"poiData":[{"dealid":"39274460,39274710,44465539","acm":"AcategoryC11508628254521165807.99228003.1","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"44000149","acm":"AcategoryC11508628254521165807.68344983.2","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"44983020,41599825,41601468,46025362","acm":"AcategoryC11508628254521165807.110977328.3","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"41732528,41685296,41703917,42728812","acm":"AcategoryC11508628254521165807.117052767.4","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"44620903,44210767,44211238,44211949","acm":"AcategoryC11508628254521165807.112420850.5","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"38132423,34590653,38401699,42442767,38132361","acm":"AcategoryC11508628254521165807.77159167.6","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"43291837,43291334","acm":"AcategoryC11508628254521165807.117440044.7","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"29036921,35191078,44795105,45930991,35191472","acm":"AcategoryC11508628254521165807.372213.8","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"26152701","acm":"AcategoryC11508628254521165807.312612.9","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"41477158","acm":"AcategoryC11508628254521165807.1539906.10","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"45692630,45693728,45688651","acm":"AcategoryC11508628254521165807.6032267.11","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""},{"dealid":"43886761,43886126","acm":"AcategoryC11508628254521165807.117103436.12","dealacm":"","mtad\_icon":"","mtad\_show\_log\_url":"","mtad\_click\_log\_url":""}],"limit":10,"bigImageMode":true},"limit":10,"initialLoad":10}

还有一个神奇的发现就是错区的json串粘贴到微软的OneNote，所有的错误居然都神奇被修正了

Day02问题总结

图片的抓取问题：按照原本的经验发现获取不到图片的地址

正常的思路就是定位到img标签过后，直接用jsoup提供的attr("src")，获取出来的src地址和F12检查出来的地址长相大相径庭，

通过Jsoup获取出来的：

通过F12检查出来的：http://p0.meituan.net/208.126/deal/a016d5c50702d3bd08736ea7f2e78630416805.jpg@0\_188\_2664\_1615a%7C388h\_640w\_2e\_90Q

那么问题来了：

1.Jsoup获取出来的是个什么鬼？！

百度了一下：将图片对应的二进制代码，通过base64转换成可以打印的字母数字的形式。图片类型可以为:images/gif,images/jpg,images/png等。data:图片类型;base64,base64转换成的数据

优点：

当图片很小时，但是占用一个HTTP请求显得太浪费时，可以用这个技术。

当图片是服务器端生成，并且每个用户都不一致的时候，例如验证码，有的网站的头像也可以用这种技术。

缺点：

base64转换过后的图片会比原来的图片大1/3

DataURL形式的图片不会被浏览器所缓存。

Tips:网上有在线的转换工具，可以将图片转换成dataurl格式

BUT：Jsoup从图片的src属性中获取出来的dataurl并不正确或是不完整，不知是不是美团做了手脚还是像网友所说Jsoup对src属性做了过滤

最后问题还是图片的地址怎么找？

既然Jsoup获取的src和F12检查的不一样，那么会不会他们的标签结构也不一样，于是我用Jsoup定位到图片位置使用.toString()方法打印出来了该条标签结构：

<img class="lazy-img J-lazy-img" src="" data-src="http://p0.meituan.net/208.126/deal/a016d5c50702d3bd08736ea7f2e78630416805.jpg@0\_188\_2664\_1615a%7C388h\_640w\_2e\_90Q" alt="小腰5串1份，包间免费-望京小腰烧烤（安贞店）-美团" width="117" height="70">

F12检查出来的img标签长这样

<img class="" src="http://p0.meituan.net/208.126/deal/a016d5c50702d3bd08736ea7f2e78630416805.jpg@0\_188\_2664\_1615a%7C388h\_640w\_2e\_90Q" alt="小腰5串1份，包间免费-望京小腰烧烤（安贞店）-美团" width="117" height="70" id="yui\_3\_16\_0\_1\_1499430105399\_1669">

通过比对不难发现，用Jsoup获取的img中data-src属性中才是我们想要的图片地址，它和F12检查的src属性中的地址是一模一样的，再拿它去访问一样，看看能不能出现图片

计算机生成了可选文字:
京 小 生 烤 一 人 均 价 0 × data url_ 百 度 雀 索 
× @〕到片的DatalJRLF*-a × 团 a01bd5C50702d 彐 《 × 
。 》 C 亡 PO ． meituan ． net / 208 ． 126/de引/a016d5C50702d3bd08736ea7f2e786引〕416805jpg@0一188 2664 1615a ％ 7C388h 640W 2e 一 90Q 