

```

1  """
2  brief:使用BeautifulSoup和request爬天堂网的图片
3  author:chenyijun
4  date:2020-02-15
5  """
6  import requests
7  from bs4 import BeautifulSoup
8  import re
9  import urllib
10 import urllib.request
11
12
13 def cbk(a, b, c):
14     '''回调函数
15     @a:已经下载的数据块
16     @b:数据块的大小
17     @c:远程文件的大小
18     '''
19     per=100.0*a*b/c
20     if per>100:
21         per=100
22     print ('%.2f%%' % per)
23     print(" ")
24
25 url = 'https://www.ivsky.com/tupian/meishishijie/' #取一个图片目录 美食世界
26 headers ={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/53.0.2785.104 Safari/537.36 Core/1.53.3427.400 QQBrowser/9.6.12513.400', 'Referer':'http://www.ivsky.com/tupian/qita/index_11.html'}
27 html = requests.get(url, headers = headers) #获取网页内容
28 #print(html) #太多了，打印不出来
29
30 soup = BeautifulSoup(html.text, 'html.parser')
31
32 def spidertupian():
33     for i in range(1, 12):
34         link = url + '/index_'+str(i)+'.html'
35         #print(link) #打印链接
36

```

```

37  headers = {
38      'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36
      (KHTML, like Gecko) Chrome/53.0.2785.104 Safari/537.36 Core/1.53.3427.400 Q
      QBrowser/9.6.12513.400',
39      'Referer': 'http://www.ivsky.com/tupian/qita/index_11.html'}
40  html = requests.get(link, headers=headers)
41  mess = BeautifulSoup(html.text, 'html.parser')
42  # 查找标签为'ul', class属性为'ali'的标签元素, 因为class是python的关键字, 所
    以这里需要加个下划线 '_'
43  for page in mess.find_all('ul', class_='ali'):
44      #print(page) #打印带<ul class="ali">标签的
45      #print("-----")
46      x = 0
47      for img in page.find_all('img'): #文件夹的url
48          #print(img)
49          imgurl = img.get('src') #获取src字段
50          save_path = "F:/github/qtforpython/spider/tiantang/" + str(i) + "_" + s
            tr(x) + ".jpg" #拼接图片保存路径
51          imghttp = 'https:' + imgurl #拼接图片的url路径
52          #print(imghttp)
53          urllib.request.urlretrieve(imghttp, save_path, cbk)
54          x += 1
55
56  if __name__ == '__main__':
57      spidertupian()
58
59  #html格式化文档网站
60  #https://tool.oschina.net/codeformat/
61
62  #参考文档
63  # https://blog.csdn.net/qq_27492735/article/details/78478750
64  #
65  # https://www.cnblogs.com/Deaseyy/p/11266742.html
66  # https://blog.csdn.net/nicholas_K/article/details/85275793
67  # https://blog.csdn.net/dayun555/article/details/79375841
68  # https://www.52pojie.cn/thread-1071469-1-1.html
69  # https://www.cnblogs.com/fwc1994/p/5878934.html
70

```

```
tiantangSpider ×
"D:\Program Files\Python38\python.exe" F:/qtforpython/startedPython/tiantangSpider.py
Traceback (most recent call last):
  File "F:/qtforpython/startedPython/tiantangSpider.py", line 55, in <module>
    urllib.urlretrieve(imgurl,work_path,cbk)
NameError: name 'urllib' is not defined

进程已结束，退出代码 1
```

`urllib.urlretrieve(imgurl, work_path, cbk)`

改成

`urllib.request.urlretrieve(imgurl, work_path, cbk)`

源码分析

1.首先我们找的是天堂网的图片网站，以美食世界这个目录为例

<https://www.ivsky.com/tupian/meishishijie/>

所有图片 自然风光 城市旅游 动物图片 植物花卉 海洋世界 人物图片 **美食世界** 物品物件 运动体育 交通运输 建筑环境 装饰装修
广告设计 卡通图片 节日图片 设计素材 艺术绘画 其他类别

小分类

中华美食 日式料理 韩国料理 牛排 西餐美食 节日美食 地方小吃 蛋糕 甜品甜点 面包 糖果 干果 水果 蔬菜 菌类 肉类 蛋类
牛奶 咖啡 饮料 果汁 茶 茶道 酒 海鲜 五谷杂粮 香料 调味料 川菜 粤菜 苏菜 湘菜 烧烤 香肠 腊肠 面点 面食
家庭美食 营养美食 健康美食 零食 其他美食

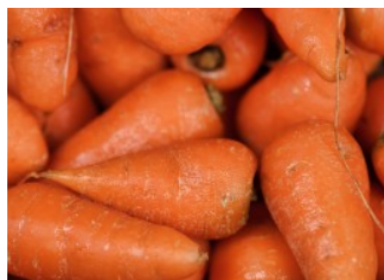
这些都图片目录



美味好吃的披萨图片(25张)



圆滚滚的土豆图片(12张)



营养健康清脆的胡萝卜图片(16张)



让人食欲大振的蛋挞图片(19张)



酸甜好吃又营养的西红柿图片(18张)



美味好吃的意大利面图片(22张)



右击属性找到个图片的url

<https://img.ivsky.com/img/tupian/li/201908/23/pisa-008.jpg>
<https://img.ivsky.com/img/tupian/li/201908/21/tudou-001.jpg>
<https://img.ivsky.com/img/tupian/li/201908/21/huluobo-019.jpg>
<https://img.ivsky.com/img/tupian/li/201908/21/danta-007.jpg>
<https://img.ivsky.com/img/tupian/li/201908/21/xihongshi-021.jpg>
<https://img.ivsky.com/img/tupian/li/201908/21/pasta-010.jpg>
<https://img.ivsky.com/img/tupian/li/201908/20/shiliu-004.jpg>
<https://img.ivsky.com/img/tupian/li/201908/18/youyu-009.jpg>
<https://img.ivsky.com/img/tupian/li/201908/18/qiaokeli-006.jpg>

这是前面9个目录图片对应的url。

拉到最下面，有网页的分页



美味好吃的汉堡图片(12张)



补充维生素C的柠檬图片(12张)



金黄的玉米图片(13张)



绿色的香菜图片(13张)



营养补铁的海带图片(11张)



辛辣驱寒的生姜图片(18张)



免费图片素材网

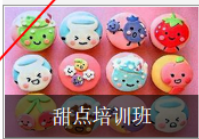


早餐小吃培训

- 国外服务器
- 小吃
- 甜点培训
- 加盟小吃
- 留学中介
- 国外租车



甜点



甜点培训班

- 国外旅游
- 留学留学中介
- 国外婚礼
- 小吃加盟网
- 小吃加盟排行榜
- 留学机构



小吃培训



小吃培训班

查看对应页的url

<https://www.ivsky.com/tupian/meishishijie/>
https://www.ivsky.com/tupian/meishishijie/index_2.html
https://www.ivsky.com/tupian/meishishijie/index_3.html
https://www.ivsky.com/tupian/meishishijie/index_4.html
https://www.ivsky.com/tupian/meishishijie/index_5.html
https://www.ivsky.com/tupian/meishishijie/index_6.html

https://www.ivsky.com/tupian/meishishijie/index_7.html
https://www.ivsky.com/tupian/meishishijie/index_8.html
https://www.ivsky.com/tupian/meishishijie/index_9.html
https://www.ivsky.com/tupian/meishishijie/index_10.html
https://www.ivsky.com/tupian/meishishijie/index_11.html

2.分析网页的源码

把源码放到<https://tool.oschina.net/codeformat/>这个网站上查看结构

在线代码格式化

HTML格式化XML格式化CSS格式化JSON格式化JavaScript格式化Java格式化SQL格式化

HTML格式化采用 Jsoup

待格式化HTML:
图片">西瓜图片汤圆图片生日蛋糕图片大白菜图片水饺图片<a href="/tupian

格式化复制格式化代码

格式化HTML:

</div><p>美味好吃的披萨图片(25张)</p>
<div class="il_img">

</div><p>圆滚滚的土豆图片(12张)</p>
<div class="il_img">

</div><p>营养健康清脆的胡萝卜图片(16张)</p>

与前面的对比发现，图片的路径与url的区别就是少了一个https:

```
for page in mess.find_all('ul', class_='ali'):
    for img in page.find_all('img'):
        imgre = re.compile(r'src="(.*?\.\jpg)" alt')
        imglist = re.findall(imgre,html.text)
        imgurl = img.get('src')
        print(imgurl) #打印每张图的路径
```

用代码打印输出


```
//img.ivsky.com/img/tupian/li/201908/23/pisa-008.jpg
//img.ivsky.com/img/tupian/li/201908/21/tudou-001.jpg
//img.ivsky.com/img/tupian/li/201908/21/huluobo-019.jpg
//img.ivsky.com/img/tupian/li/201908/21/danta-007.jpg
//img.ivsky.com/img/tupian/li/201908/21/xihongshi-021.jpg
//img.ivsky.com/img/tupian/li/201908/21/pasta-010.jpg
//img.ivsky.com/img/tupian/li/201908/20/shiliu-004.jpg
//img.ivsky.com/img/tupian/li/201908/20/xigua-009.jpg
//img.ivsky.com/img/tupian/li/201908/19/guiji-004.jpg
//img.ivsky.com/img/tupian/li/201908/19/shengri_dangao-016.jpg
//img.ivsky.com/img/tupian/li/201908/18/youyu-009.jpg
//img.ivsky.com/img/tupian/li/201908/18/qiaokeli-006.jpg
//img.ivsky.com/img/tupian/li/201908/18/hanbao-001.jpg
//img.ivsky.com/img/tupian/li/201908/16/ningmeng-007.jpg
//img.ivsky.com/img/tupian/li/201908/16/yumi-001.jpg
//img.ivsky.com/img/tupian/li/201908/15/xiangcai-013.jpg
//img.ivsky.com/img/tupian/li/201908/15/haidai-007.jpg
//img.ivsky.com/img/tupian/li/201908/15/shengjiang-002.jpg
```

页码输出

```
for i in range(0, 12):
    link = url + '/index ' + str(i) + '.html'
    print(link) #打印链接
```

https://www.ivsky.com/tupian/meishishijie//index_0.html
https://www.ivsky.com/tupian/meishishijie//index_1.html
https://www.ivsky.com/tupian/meishishijie//index_2.html
https://www.ivsky.com/tupian/meishishijie//index_3.html
https://www.ivsky.com/tupian/meishishijie//index_4.html
https://www.ivsky.com/tupian/meishishijie//index_5.html
https://www.ivsky.com/tupian/meishishijie//index_6.html
https://www.ivsky.com/tupian/meishishijie//index_7.html
https://www.ivsky.com/tupian/meishishijie//index_8.html
https://www.ivsky.com/tupian/meishishijie//index_9.html
https://www.ivsky.com/tupian/meishishijie//index_10.html
https://www.ivsky.com/tupian/meishishijie//index_11.html

查看网站上的源码结构

在线代码格式化

Feedback

HTML格式化 XML格式化 CSS格式化 JSON格式化 JavaScript格式化 Java格式化 SQL格式化

HTML格式化采用 Jsoup

待格式化HTML:

图片">西瓜图片汤圆图片生日蛋糕图片大白菜图片水饺图片<a href="/tupian

格式化

复制格式化代码

格式化HTML:

```
</div>
<div class="pagelist">
  <span class="page-cur">1</span>
  <a href="/tupian/meishishijie/index_2.html">2</a>
  <a href="/tupian/meishishijie/index_3.html">3</a>
  <a href="/tupian/meishishijie/index_4.html">4</a>
  <a href="/tupian/meishishijie/index_5.html">5</a>
  <a href="/tupian/meishishijie/index_6.html">6</a>
  <a href="/tupian/meishishijie/index_7.html">7</a>
  <a href="/tupian/meishishijie/index_8.html">8</a>
  <a href="/tupian/meishishijie/index_9.html">9</a>
  <a href="/tupian/meishishijie/index_10.html">10</a>
  <a href="/tupian/meishishijie/index_11.html">11</a>
  <a class="page-next" href="/tupian/meishishijie/index_2.html">下一页</a>
```

```
# 查找标签为'ul', class属性为'ali'的标签元素, 因为class是python的关键字, 所以这里需要加个下划线'_'
for page in mess.find_all('ul', class_='ali'):
    print(page) #打印带<ul class="ali">标签的
```

查找这段html的结构

待格式化HTML：

```
/02/shuiguo_dangao-009.jpg"/></a></div><p><a href="/tupian/shuiguo_dangao_v57549/" target="_blank" title="美味的水果蛋糕图片">美味的水果蛋糕图片(15张)</a></p></li><li><div class="il_img"><a href="/tupian/xiangzi_v57544/" target="_blank" title="好吃的橙子坚果图片">好吃的橙子坚果图片(8张)</a></a></div><p><a href="/tupian/xiangzi_v57544/" target="_blank" title="好吃的橙子坚果图片">好吃的橙子坚果图片(8张)</a></p></li></div><div class="il_img"><a href="/tupian/xuehua_niurou_v57536/" target="_blank" title="香喷喷的雪花秀牛肉图片">香喷喷的雪花秀牛肉图片(11张)</a></a></div><p><a href="/tupian/xuehua_niurou_v57536/" target="_blank" title="香喷喷的雪花秀牛肉图片">香喷喷的雪花秀牛肉图片(11张)</a></p></li><li><div class="il_img"><a href="/tupian/xiangjiao_v57527/" target="_blank" title="好吃有营养的香蕉图片">好吃有营养的香蕉图片(10张)</a></a></div><p><a href="/tupian/xiangjiao_v57527/" target="_blank" title="好吃有营养的香蕉图片">好吃有营养的香蕉图片(10张)</a></p></li></div><div class="il_img"><a href="/tupian/mianhuatang_reyin_v57534/" target="_blank" title="好吃的棉花糖图片">好吃的棉花糖图片(16张)</a></a></div><p><a href="/tupian/mianhuatang_reyin_v57534/" target="_blank" title="好吃的棉花糖图片">好吃的棉花糖图片(16张)</a></p></li></ul>
```

格式化 复制格式化代码

格式化HTML：

这段标签下的结构

```
<html>
<head></head>
<body>
<ul class="ali">
<li>
<div class="il_img">
<a href="/tupian/chengzi_v57765/" target="_blank" title="酸甜可口的橙子图片">酸甜可口的橙子图片(17张)</a></a>
</div><p><a href="/tupian/chengzi_v57765/" target="_blank" title="酸甜可口的橙子图片">酸甜可口的橙子图片(17张)</a></p></li>
<div class="il_img">
```

拼接url图片路径和下载保存路径，下载图片

```
1 save_path = "F:/github/qtforpython/spider/tiantang/" + str(i) + "_" + str(x) + ".jpg" #拼接图片保存路径
2 imghttp = 'https:' + imgurl #拼接图片的url路径
3 urllib.request.urlretrieve(imghttp, save_path, cbk)
```

运行下载图片：



