```
"""
brief:使用BeautifulSoup和request爬天堂网的图片
author:chenyijun
date:2020-02-15
"""
import requests
from bs4 import BeautifulSoup
import re
import urllib
 import urllib.request


 def cbk(a, b, c):
  '''''回调函数
  @a:已经下载的数据块
  @b:数据块的大小
  @c:远程文件的大小
  '''
  per=100.0*a*b/c
  if per>100:
  per=100
  print ('%.2f%%' % per)
  print(" ")

 url = 'https://www.ivsky.com/tupian/meishishijie/' #取一个图片目录 美食世界
 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'}
 html = requests.get(url, headers = headers) #获取网页内容
 #print(html) #太多了，打印不出来

 soup = BeautifulSoup(html.text,'html.parser')

 def spidertupian():
  for i in range(1, 12):
  link = url +'/index_'+str(i)+'.html'
  #print(link) #打印链接

```

```
    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'}
    html = requests.get(link, headers=headers)
    mess = BeautifulSoup(html.text, 'html.parser')
    # 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
    #print("--------------------------")
    x = 0
    for img in page.find_all('img'): #文件夹的url
    #print(img)
    imgurl = img.get('src') #获取src字段
    save_path = "F:/github/qtforpython/spider/tiantang/" + str(i) + "_" + str(x) + ".jpg" #拼接图片保存路径
    imghttp = 'https:' + imgurl #拼按图片的url路径
    #print(imghttp)
    urllib.request.urlretrieve(imghttp, save_path, cbk)
    x += 1

if __name__ == '__main__':
    spidertupian()

#html格式化文档网站
#https://tool.oschina.net/codeformat/

#参考文档
# https://blog.csdn.net/qq_27492735/article/details/78478750
#
# https://www.cnblogs.com/Deaseyy/p/11266742.html
# https://blog.csdn.net/nicholas_K/article/details/85275793
# https://blog.csdn.net/dayun555/article/details/79375841
# https://www.52pojie.cn/thread-1071469-1-1.html
# https://www.cnblogs.com/fwc1994/p/5878934.html

```

```
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)

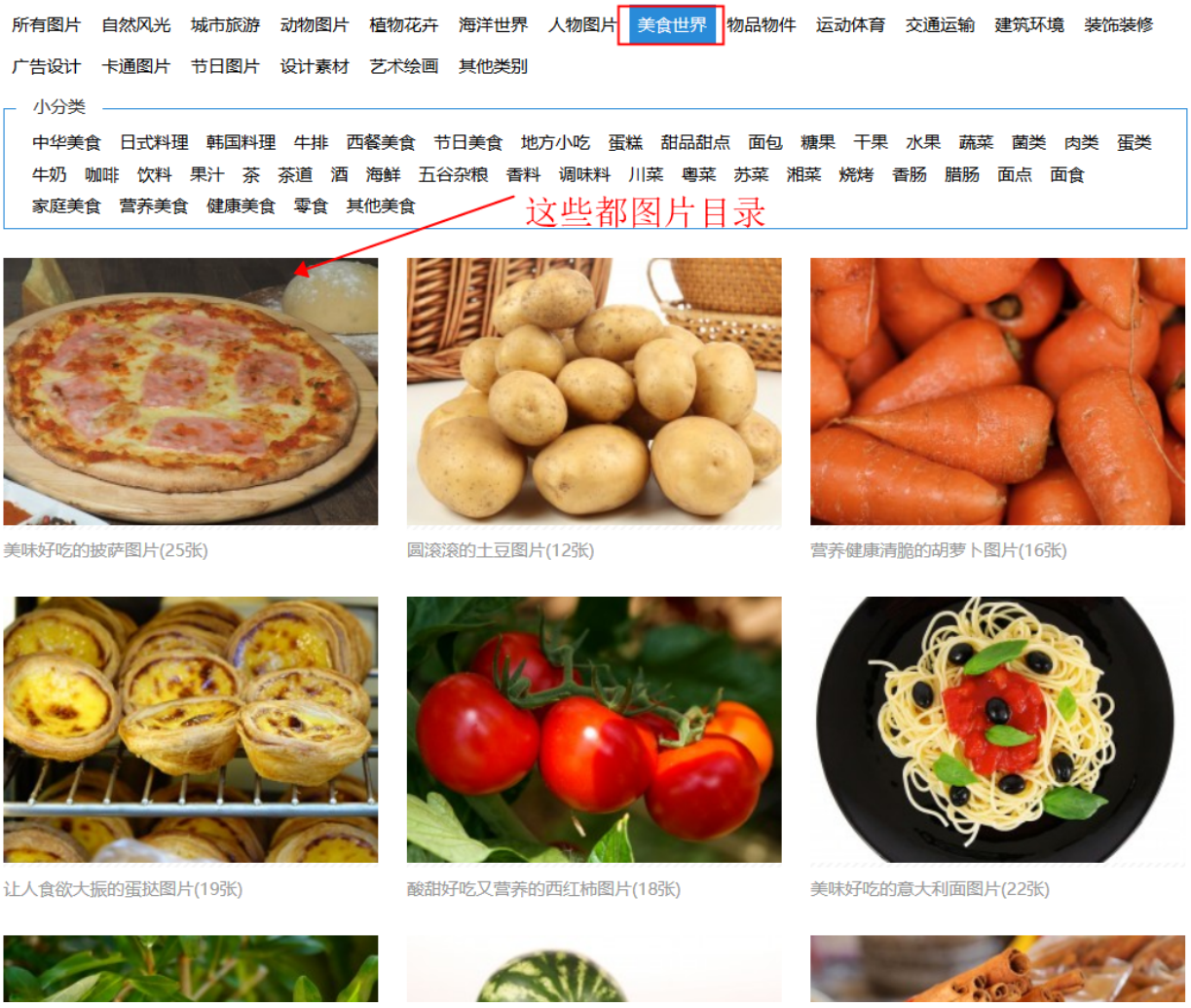
改成

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/



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

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

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

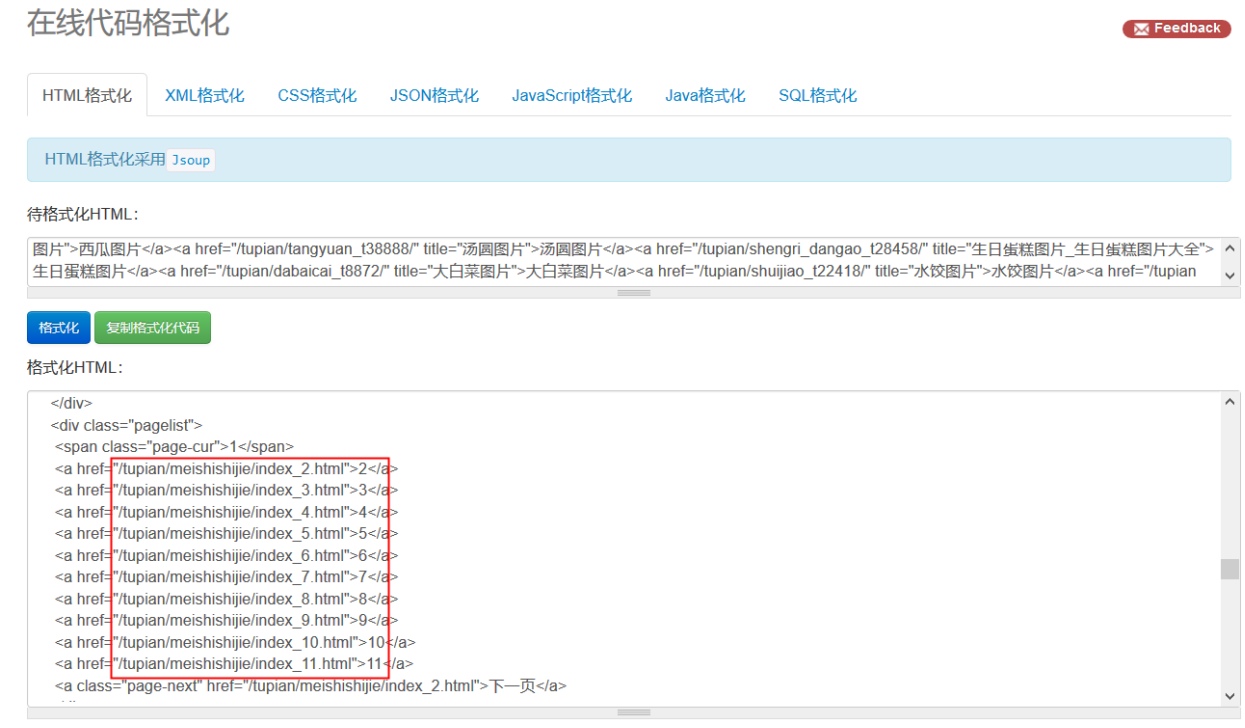
```
//img.ivsky.com/img/tupian/li/201908/23/pisa-008.jpg
//img.ivsky.com/img/tupian/li/201908/21/tudou-001.jpg
//img.ivsky.com/img/tupian/li/201908/21/huluobo-019.jpg
//img.ivsky.com/img/tupian/li/201908/21/danta-007.jpg
//img.ivsky.com/img/tupian/li/201908/21/xihongshi-021.jpg
//img.ivsky.com/img/tupian/li/201908/21/pasta-010.jpg
//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/guipi-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




```
# 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"/> 美味的水果蛋糕图片(15张)   好吃的橡子坚果图片(8张)   香喷喷的雪花秀牛肉图片(11张)   好吃有营养的香蕉图片(10张)   好吃的棉花糖图片(16张)

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(17张)"><img alt="酸甜可口的橙子图片" src="//img.ivsky.com/img/tupian/li/201908/15/chengzi-008.jpg" /></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>
   <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 #拼按图片的url路径
3 urllib.request.urlretrieve(imghttp, save_path, cbk)
```

运行下载图片：