2.win10 下 python2 爬虫美女图片逐步优化

作者：jackssybin
原文链接：https://ld246.com/article/1564974784845
来源网站：链滴

```
#coding:utf-8
#完成通用爬虫，抓取一个页面队列中所有图片

import requests
import re
import time
from bs4 import BeautifulSoup
import uuid
import urllib
import os

headers={ 'User-Agent':'Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gec
o) Chrome/52.0.2743.116 Safari/537.36' }
base_url='http://m.17786.com'
base_img_url='https://wcyouxi.sc601.com/file/p/20190803/21/ordurlsatqe.jpg'
save_path='E:\\360MoveData\Users\Administrator\Pictures\Camera Roll\crawl\meizi\\'

def download_detail_url_soup(url,folerName,num):
    print url,folerName
    html=requests.get(url, headers=headers, timeout=50)
    html.encoding = 'utf-8'
    soup = BeautifulSoup(html.text,"lxml" ,from_encoding='utf-8')
    print soup.title
    try:
        imgs=soup.select('.contimglist')[0].find_all("img", {"src": re.compile(".*\.jpg")})
        for img in imgs:
            # 获取图片下载地址，并下载图片
            # dizhi = li.find('img')['src']
            base_path=save_path + folerName
            if not os.path.isdir(base_path):
                os.mkdir(base_path)
            # 图片保存地址
            bc = base_path +"\\"+str(num)+ str(uuid.uuid4()) + ".jpg"
             # 下载图片并保存
            urllib.urlretrieve(img.get("src"), bc)
    except:
        print "下载失败"
        pass

    return -1

def getDetailPageInfo(url):
    # < div class ="page" >
    # < a  href = "17409_2.html" class ="linkpage shpage" > 上一页 < / a >
    # < a href="17409.html" > 首页 < / a >
    # < a name="allpage" class ="allpage" > < span class ="nowpage" > 3 < / span > / 18 < /
 >
    # < a href="/rihan/17409_18.html" > 尾页 < / a >
    # < a href="/rihan/17409_4.html" class ="linkpage" > 下一页 < / a >
    # < / div >
    html = requests.get(url, headers=headers, timeout=50)
    html.encoding = 'utf-8'
    soup = BeautifulSoup(html.text, "lxml", from_encoding='utf-8')
    print soup.title
```

```
    print soup.select('.page a')
    pageInfo={}
    for page in soup.select('.page a'):
        if page.text.find("/") != -1:
            pageInfo['total'] = str(page.text)[page.text.find("/")+1:]
    print pageInfo

    url.rfind("/")

    baseUrl=url[0:url.rindex("/")+1]
    base_detail_url=url[url.rindex("/")+1:]
    base_detail_url=base_detail_url[0:base_detail_url.rfind(".")]
    if base_detail_url.rfind("_")>0:
        base_detail_url=base_detail_url[0:base_detail_url.rfind("_")]
    pageInfo['first']=base_detail_url;
    folderName=soup.title
    for num in range(int(pageInfo['total'])):
        if num ==0:
            detail_html_url= baseUrl+pageInfo['first']+".html"
        else:
            detail_html_url= baseUrl+pageInfo['first']+"_"+str(num+1)+".html"
        download_detail_url_soup(detail_html_url,folderName.string,num+1)




def getAllPageInfo(url):
    # < div class ="page" >
    # < a  href = "17409_2.html" class ="linkpage shpage" > 上一页 < / a >
    # < a href="17409.html" > 首页 < / a >
    # < a name="allpage" class ="allpage" > < span class ="nowpage" > 3 < / span > / 18 < /
 >
    # < a href="/rihan/17409_18.html" > 尾页 < / a >
    # < a href="/rihan/17409_4.html" class ="linkpage" > 下一页 < / a >
    # < / div >
    html = requests.get(url, headers=headers, timeout=50)
    html.encoding = 'utf-8'
    soup = BeautifulSoup(html.text, "lxml", from_encoding='utf-8')
    print soup.title
    print soup.nav
    print soup.find_all("ul", class_="tag",limit=12)[0]
    allIndexList=[]
    allIndexList.append('/meinv/');
    allIndexList.append('/mxmn/');
    for index in  soup.find_all("ul", class_="tag",limit=12)[0].find_all("li"):
        allIndexList.append(index.find("a").get("href"))

    allDetailPageInfos=[]
    for detailCrawUrl in allIndexList:
        detail_crawl_url=base_url+detailCrawUrl
        print detail_crawl_url
        htmla = requests.get(detail_crawl_url, headers=headers, timeout=50)
        htmla.encoding = 'utf-8'
        soupa = BeautifulSoup(htmla.text, "lxml", from_encoding='utf-8')
```

```
        print soupa.title
        for detailPageInfo in soupa.find("div",attrs={'id':'list'}).find_all("li"):
            getDetailPageInfo(base_url+detailPageInfo.find("a").get("href"))

if __name__ == '__main__':
    print "begin"
    # push_redis_list(6952)#开启则加任务队列.其中的值请限制在5400以内。不过是用于计算页码的

    # push_index_all_url_to_redis_list()#//从首页爬取所有url
    # detail_url='http://m.772586.com/mxmn/17313.html'
    # detail_url='http://m.772586.com/rihan/17409_5.html'
    # download_detail_url_soup(detail_url,"test")
    # getDetailPageInfo(detail_url)

    detail_url = 'http://m.772586.com/'
    getAllPageInfo(detail_url);


    #get_big_img_url()#开启则运行爬取任务
```