链滴

一个简易的图片爬虫玩具

作者：Pleuvoir

原文链接：https://ld246.com/article/1522470618719

来源网站：链滴

准备工作

字符串保存为文件

```
def save(filename, contents):
    fh = open(filename, 'w', encoding='utf-8')
    fh.write(contents)
    fh.close()
```

从URL读取并保存二进制文件

```
def read_and_save_file(url,newFileName):
    r = requests.get(url)
    with open(newFileName,'wb') as file:
        file.write(r.content)
        file.close()
```

返回网页的干净文本

```
def getHtmlPlainText(url):
    return requests.get(url).text
```

获取网页中待下载图片的url

```
def getImagesUrl(url):
    text = getHtmlPlainText(url)
```

```
    #匹配图片的正则
    r = re.compile('<img pic_type="0" class="BDE_Image" src="(.*?)"')
    ImagesUrlList = r.findall(text)
    return ImagesUrlList
```

main

获取贴吧具体贴子内图片

```
def fetchPictures(max_page):

    print('抓取开始')
    startTime = time.time()
    page = 1
    # 创建文件夹
    exists = os.path.exists(os.getcwd() + '\pictures')
    if exists==True: print(os.getcwd() + '\pictures'+"路径已存在!")
    if exists == False:
        os.mkdir('pictures')

    # 切换到此目录下下载的文件将保存于此  os.getcwd()为当前工作目录
    os.chdir(os.path.join(os.getcwd(), 'pictures'))
    while page <= max_page :

        url = 'http://tieba.baidu.com/p/2460150866?pn=' + str(page)
        ImagesUrl = getImagesUrl(url)
        print(url)
        print('抓取第' + str(page) + '页')
        #批量下载
        for i in range(0,len(ImagesUrl)):
            #图片urli
            #print(ImagesUrl[i])
            print('抓取第' + str(i+1) + '张图片')
            read_and_save_file(ImagesUrl[i], str(time.time()) + '.jpg')
        page += 1

    print('抓取结束 耗时:' + str(time.time() - startTime)[0:4] + 's')
```

获取帝吧前N页的内容

```
def tieba_crawler(max_pages):
    print('抓取开始')
    startTime = time.time()
    content = ''
    page = 1
    pn = 0
    while page <= max_pages:
        print('抓取第' + str(page) + '页')
        content += '### 第' + str(page) + '页\n'
        url = "http://tieba.baidu.com/f?kw=%E6%9D%8E%E6%AF%85&ie=utf-8&pn=" + str(pn)
        plain_text = getHtmlPlainText(url)
        soup = BeautifulSoup(plain_text, 'html.parser')
```

```
        for link in soup.findAll('a', {'class', 'j_th_tit'}):
            href = 'http://tieba.baidu.com/' + link.get('href')
            title = link.get('title')
            content += '[' + title + ']' + '(' + href + ')' + '\n'
            # print(content)

        page += 1
        pn = (page - 1) * 50

    print('抓取结束 耗时:' + str(time.time() - startTime)[0:4] + 's')
    save('B:/python-crawler-space/tieba.md', content)
```

完整代码

```
#!/user/bin/env python
# -*- encoding:utf-8 -*-

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


def tieba_crawler(max_pages):

    print('抓取开始')
    startTime = time.time()
    content = ''
    page = 1
    pn = 0
    while page <= max_pages:
        print('抓取第' + str(page) + '页')
        content += '### 第' + str(page) + '页\n'
        url = "http://tieba.baidu.com/f?kw=%E6%9D%8E%E6%AF%85&ie=utf-8&pn=" + str(pn)
        plain_text = getHtmlPlainText(url)
        soup = BeautifulSoup(plain_text, 'html.parser')
        for link in soup.findAll('a', {'class', 'j_th_tit'}):
            href = 'http://tieba.baidu.com/' + link.get('href')
            title = link.get('title')
            content += '[' + title + ']' + '(' + href + ')' + '\n'
            # print(content)

        page += 1
        pn = (page - 1) * 50

    print('抓取结束 耗时:' + str(time.time() - startTime)[0:4] + 's')
    save('B:/python-crawler-space/tieba.md', content)



#字符串保存为文件
def save(filename, contents):
```

```
    fh = open(filename, 'w', encoding='utf-8')
    fh.write(contents)
    fh.close()



#从网上读取并保存二进制文件
def read_and_save_file(url,newFileName):

    r = requests.get(url)
    with open(newFileName,'wb') as file:
        file.write(r.content)
        file.close()



def fetchPictures(max_page):

    print('抓取开始')
    startTime = time.time()
    page = 1
    # 创建文件夹
    exists = os.path.exists(os.getcwd() + '\pictures')
    if exists==True: print(os.getcwd() + '\pictures'+"路径已存在!")
    if exists == False:
        os.mkdir('pictures')

    # 切换到此目录下下载的文件将保存于此  os.getcwd()为当前工作目录
    os.chdir(os.path.join(os.getcwd(), 'pictures'))
    while page <= max_page :

        url = 'http://tieba.baidu.com/p/2460150866?pn=' + str(page)
        ImagesUrl = getImagesUrl(url)
        print(url)
        print('抓取第' + str(page) + '页')
        #批量下载
        for i in range(0,len(ImagesUrl)):
            #图片urli
            #print(ImagesUrl[i])
            print('抓取第' + str(i+1) + '张图片')
            read_and_save_file(ImagesUrl[i], str(time.time()) + '.jpg')
        page += 1

    print('抓取结束 耗时:' + str(time.time() - startTime)[0:4] + 's')


# 获取网页中待下载图片的url
def getImagesUrl(url):

    text = getHtmlPlainText(url)
    # 匹配图片的正则
    r = re.compile('<img pic_type="0" class="BDE_Image" src="(.*?)"')
    ImagesUrlList = r.findall(text)
```

```
    return ImagesUrlList


#返回网页的干净文本
def getHtmlPlainText(url):

    return requests.get(url).text


#fetchPictures(70)
#tieba_crawler(2)
```