



分享

Python爬取网站博客教程并制作成PDF

文章来源：企鹅号 - 嗨学Python

要把教程变成PDF有三步：

- 1、先生成空html，爬取每一篇教程放进一个新生成的div，这样就生成了包含所有教程的html文件(BeautifulSoup)
- 2、将html转换成pdf(wkhtmltopdf)
- 3、如果有反爬，在爬取的过程中还需要代理ip

BeautifulSoup

Beautiful Soup 是一个可以从HTML或XML文件中提取数据的Python库.它能够通过你喜欢的转换器实现惯用的文档导航,查找,修改文档的方式.Beautiful Soup会帮你节省数小时甚至数天的工作时间.

安装

```
pip3 install BeautifulSoup4
```

开始使用

将一段文档传入 BeautifulSoup 的构造方法,就能得到一个文档的对象, 可以传入一段字符串或一个文件句柄.

如下所示:

```
from bs4 import BeautifulSoup
```

```
soup = BeautifulSoup(open("index.html"))
```

```
soup = BeautifulSoup("data")
```

首先,文档被转换成Unicode,并且HTML的实例都被转换成Unicode编码.

然后,Beautiful Soup选择最合适的解析器来解析这段文档,如果手动指定解析器那么Beautiful Soup会选择指定的解析器来解析文档.

对象的种类

Beautiful Soup 将复杂 HTML 文档转换成一个复杂的树形结构,每个节点都是 Python 对象,所有对象可以归纳为 4 种: Tag , NavigableString , BeautifulSoup , Comment .

- 1、Tag: 通俗点讲就是 HTML 中的一个标签, 类似 div , p.
- 2、NavigableString: 获取标签内部的文字, 如, soup.p.string。
- 3、BeautifulSoup: 表示一个文档的全部内容。
- 4、Comment: Comment 对象是一个特殊类型的 NavigableString 对象, 其输出的内容不包括注释符号.

Tag

Tag就是html中的一个标签, 用BeautifulSoup就能解析出来Tag的具体内容, 具体的格式为soup.name,其中name是html下的标签, 具体实例如下:

```
print soup.title输出title标签下的内容, 包括此标签, 这个将会输出
```

```
The Dormouse's storyThe Dormouse's story
```

```
The Dormouse's story
```

```
print soup.head输出head标签下的内容
```

```
The Dormouse's story
```

如果 Tag 对象要获取的标签有多多个的话, 它只会返回所以内容中第一个符合要求的标签。

Tag 属性

每个 Tag 有两个重要的属性 name 和 attrs:

name: 对于Tag, 它的name就是其本身, 如soup.p.name就是p

attrs是一个字典类型的，对应的是属性-值，如print soup.p.attrs,输出的就是{'class': ['title'], 'name': 'dromouse'},当然你也可以得到具体的值，如print soup.p.attrs['class'],输出的就是[title]是一个列表的类型，因为一个属性可能对应多个值,当然你也可以通过get方法得到属性的，如： print soup.p.get('class')。还可以直接使用print soup.p['class']

wkhtmltopdf

wkhtmltopdf主要用于HTML生成PDF。

pdfkit是基于wkhtmltopdf的python封装，支持URL，本地文件，文本内容到PDF的转换，其最终还是调用wkhtmltopdf命令。

安装

先安装wkhtmltopdf，再安装pdfkit。

<https://wkhtmltopdf.org/downloads.html>

pdfkit

shell pip3 install pdfkit

转换url/file/string

完整代码

运行过程截图：

生成的效果图：

The Dormouse's story

发表于: 2019-07-25

原文链接：https://kuaibao.qq.com/s/20190725A0UC2L00?refer=cp_1026

腾讯「云+社区」是腾讯内容开放平台帐号（企鹅号）传播渠道之一，根据《[腾讯内容开放平台服务协议](#)》转载发布内容。

- 上一篇：[UME链交所携万人演唱会冲击亚太](#)
- 下一篇：[如何手动写一个Python脚本自动爬取Bilibili小视频](#)

同媒体快讯

Python爬取加密的m3u8视频流的小电影并转换成mp4	2020-02-24
Python多进程方式抓取基金网站内容的方法分析	2020-02-24
Python爬取代理时遇到反爬的解决措施	2020-02-24
Python实现抓取斗鱼实时弹幕	2020-02-24
python一键生成属于QQ历史报告，看看你对QQ了解多深？	2020-02-24
24式加速你的Python	2020-02-24

社区

专栏文章

互动问答

技术沙龙

技术快讯

团队主页

开发者手册

智能钛AI

分享

活动

原创分享计划

自媒体分享计划

资源

在线学习中心

技术周刊

社区标签


开发者实验室

关于

社区规范

免责声明

联系我们



扫码关注云+社区

领取腾讯云代金券

热门产品

域名注册

云存储

热门推广

人脸识别

SSL 证书

更多推广

数据安全

网站监控

云服务器

宿主机

网站备案

语音识别

学生机

域名备案

区块链技术

数据可视化

短信群发平台

消息队列

CDN 加速

文字识别

网络加速

视频转码

视频点播

关系型数据库

图片文字识别

数据安全审计

域名解析

MySQL 数据库

小程序开发

Copyright © 2013 - 2020 Tencent Cloud. All Rights Reserved. 腾讯云 版权所有 京ICP备11018762号京公网安备 11010802020287