

Python爬取网页转为PDF

原创 moluchase 最后发布于2017-08-23 17:01:12 阅读量 2583 ☆ 收藏

你的浏览器目前处于缩放状态，页面可能会出现错位现象，建议100%大小显示。

爬虫的起因

官方文档或手册虽然可以查阅，但是如果变成纸质版的岂不是更容易翻阅与记忆。如果简单的复制粘贴，不知道何时能够完成。于是想着将Anc册爬下来。

全篇的实现思路

- 1. 分析网页
- 2. 学会使用BeautifulSoup库
- 3. 爬取并导出

参考资料：
* 把廖雪峰的教程转换为PDF电子书
* Requests文档
* Beautiful Soup文档

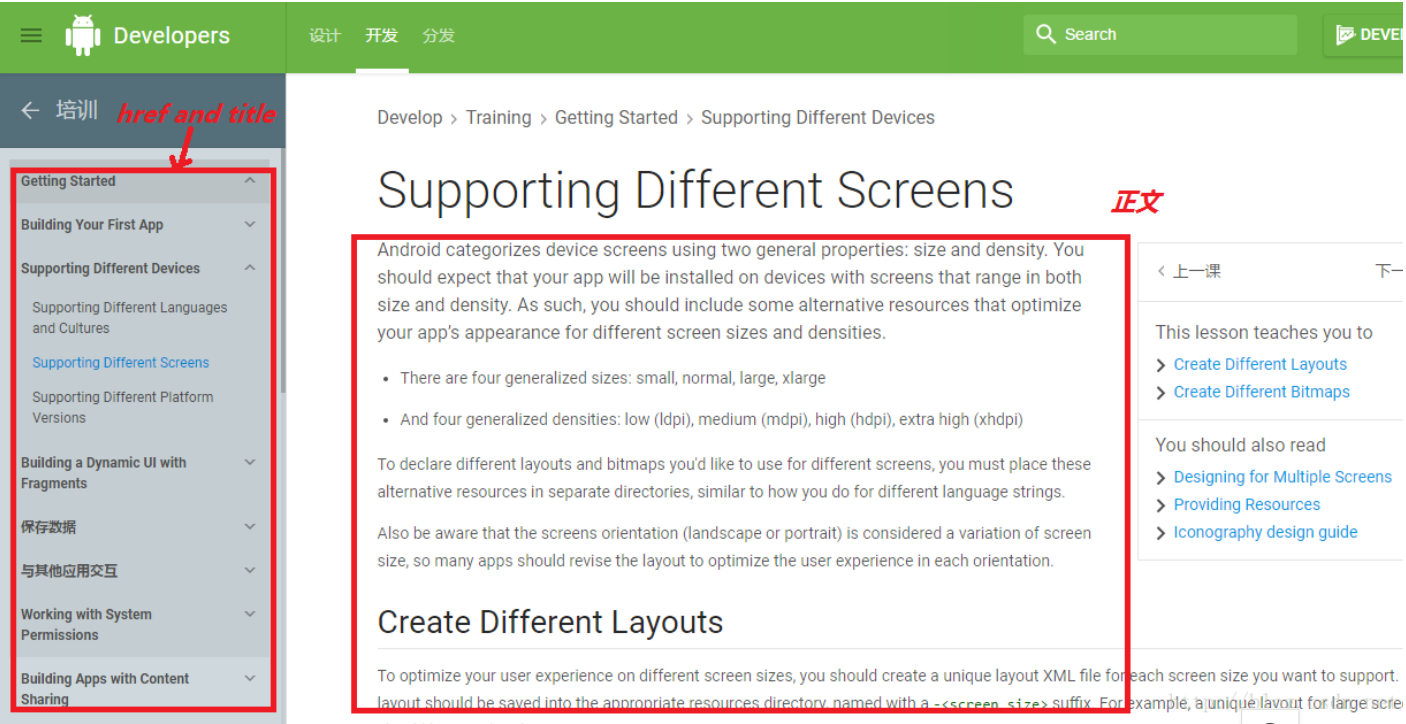
配置

在Ubuntu下使用Pycharm运行成功
转PDF需要下载wkhtmltopdf

具体过程

网页分析

这样一个网页<https://developer.android.com/training/basics/supporting-devices/screens.html>要做的是获取该网页的正文和标题，以及左边导航地址
如下所示：



接下来的工作就是找到这些标签喽...

关于Requests的使用

详细参考文档，这里只是简单的使用Requests获取html以及使用代理翻墙（网站无法直接访问，需要VPN）

6 举报 呈

CSDN

首页 博客 学院 下载 论坛 问答 活动 专题 招聘 APP VIP会员 续费8折 博客之星 Python工程师

搜索

点击上方“何俊林”，马上关注，每天早上8:50准时推送真爱，请置顶或星标本文转载自公号Python攻城狮，作者：... 博文 来自： 突围的鱼

Python爬取网页并存储为pdf

起因是最近准备学习TensorFlow，找了个网页教程，质量感觉挺好，但是页面广告巨多，不小心就能中雷，就想用...

屏蔽百家号 -(baijiahao)

快过年了，回家了，发个非技术博客吧。最近被百家号恶心到不行，搜了下屏蔽方法，在家懒得翻墙用谷歌，又懒得...

怎么自动批量把网页保存成PDF?

公司内部有一个通用模板，大家在上面提需求，经领导审批后需要我保存成PDF的形式。请问有没有什么可以自动批量操作...

将网页内容保存为PDF及为PDF创建多级书签

当你觉得某网页上的内容很不错，想保存下来，另存为下来的时候，存的是 html 页面，存的内容比较多，不好。下...

python爬虫修改版.pdf

第一章 爬虫和数据。第二章 Requests 模块。第三章 正则表达式。第四章 XPATH 提取数据。第五章 动态 HTML 处理...

一种在windows下利用python中保存网页为pdf的方法

系统：win10 64 位python版本：Python 3.6.4目标：把某一个网页保存为pdf工具：pdfkit首先我尝试了一种直接的...

[286]python将html转化为pdf

前言前面我们对博客园的文章进行了爬取，结果比较令人满意，可以一下下载某个博主的所有文章了。但是，我们...

1

出现错误

阅读数

☆

手机阅读

阅读数

赏

下载

阅读数 44

博文 来自： pikapika_chu的博客

阅读数 5329

博文 来自： 周小董

阅读数 6484

博文 来自： tijos803的博客

阅读数 1892

博文 来自： JosephPai的博客

阅读数 40

博文 来自： banfan0440的博客

阅读数 2380

博文 来自： C与Python实战

阅读数 123

博文 来自： weixin_34214500...



近视眼做激光手术利与弊

去近视眼手术

阅读数 3968

博文 来自： albertzyz的博客

07-17

下载

火狐网页保存为mht(UnMHT) v7.2.0 官方版.zip

火狐网页保存为mht(UnMHT)是一款针对firefox浏览器的辅助工具，用来将网页内容保存为mht本地文件格式。火狐网页...

将网页转换成pdf文档的方法

工具：wkhtmltopdf Adobe Acrobat 7.0 ProfessionalTeleport Pro (V1.69 Portable版本) (使用过的最好的整站拷...

浅谈利用python保存整个网站页面

阅读数 5010

博文 来自： gorquan的博客

CSDN

首页 博客 学院 下载 论坛 问答 活动 专题 招聘 APP VIP会员 续费8折 博客之星 Python工程师

目标, 要兼容所有浏览器, 让文件名和文件内的中英显示正常。首先, 文件下载, 肯定要有个文件名\$filename\$enc... 博文 来自: 风继续吹



近视眼做激光手术利与弊

去近视眼手术

Html页面保存为PDF

考试报名要打印个材料, html存不好, 找了半天才知道怎么保存成PDF, 其实就是选目标打印机时选择存为PDF即可... 博文 来自: Tyler Yang的博客

Python爬虫下载PDF文件

requests库defget_file_content(date,files):time=date[0:4]+date[5:7]file_name=files[0][1]suburl=homepag... 博文 来自: sinat_38944741

另类爬虫: 从PDF文件中爬取表格数据

简介 本文将展示一个稍微不一样点的爬虫。 以往我们的爬虫都是从网络上爬取数据, 因为网页一般用HTML,... 博文 来自: weixin_337540

python 爬取网页内容并保存为pdf格式

利用Python爬取网页中的图片内容, 并将其转换为pdf格式的文件。

读秀破解使用的下载必备工具SSLIBDTXZ1.3下载

SSLIBDTXZ1.3 这个在很多地方是找不到的 找到了也很贵 读秀破解使用的下载必备工具SSLIBDTXZ1.3 相关下载链接: //d... 论坛



百强微商团队评选排行榜

微商品牌排行

python将网页上的教程爬取下来存成pdf

首先: pip install webpage2pdfpip install pypdf2如果没有安装pyqt5, 则需要安装pyqt5, 高本版或报错, 可以装... 博文 来自: weixin_42296333...

利用python3爬虫下载图片、pdf文档

环境语言环境: python3.6操作系统: Win10第三方库requests互联网上的资源大都是以二进制形式存储和运输的, ... 博文 来自: Face_to_sun

爬取含有PDF的网页

对含有PDF的网站, 爬取含有PDF的网页, java放入jar包以后直接输入网站执行即可 下载

Python-gitbook2pdf一个轻量级gitbook网页转pdf的小工具

gitbook2pdf: 一个轻量级gitbook网页转pdf的小工具 下载

动态规划入门到熟悉, 看不懂来打我啊

持续更新。。。。。。2.1斐波那契系列问题2.2矩阵系列问题2.3跳跃系列问题3.1 01背包3.2 完全背包3.3多重背包3.... 博文 来自: hebtu666



百强微商团队评选排行榜

微商品牌排行

Java学习的正确打开方式

在博主认为, 对于入门级学习java的最佳学习方法莫过于视频+博客+书籍+总结, 前三者博主将淋漓尽致地挥毫于这... 博文 来自: 程序员宜春的博客

python json java mysql pycharm android linux json格式 c# 数组类型 泛型约束 c#的赛狗日程程序 c# 传递数组 可变参数 c# 生成存储过程 c# list 补集 c#获得所有窗体 c# 当前秒数转成年月日 c#中的枚举 c# 计算校验和 连续随机数不重复c#

©2019 CSDN 皮肤主题: 编程工作室 设计师: CSDN官方博客



moluchase

TA的个人主页 >

原创 177

粉丝 58

获赞 69

评论 34

访问 36万+

等级: 博客 5

周排名: 7万+

积分: 5001

总排名: 1万+



举报

https://blog.csdn.net/molu_chase/article/details/77508260

4/6

CSDN

首页 博客 学院 下载 论坛 问答 活动 专题 招聘 APP VIP会员 续费8折 博客之星 Python工程师

关注

私信

最新文章

线性回归 最小二乘法 方差


mac安装LightGBM with Anaconda

关于np.newaxis的一点理解

正则化方法：L1和L2 regularization、数据集扩增、dropout


在mac上安装Xgboost Python库

分类专栏




数据结构

7篇




算法练习

7篇




蓝桥杯

7篇



leetcode

17篇



Java

33篇

展开

归档

2017年11月

4篇

2017年10月

9篇

2017年9月

1篇

2017年8月

10篇

2017年7月

4篇

2017年6月

5篇

2017年5月

5篇

2017年4月

3篇

展开

热门文章

Pycharm如何添加第三方库和插件

阅读数 32885

关于np.newaxis的一点理解

阅读数 22701

前端神器-sublime text3插件安装及使用

阅读数 19649

Android下设置drawableleft导入的图片大小

1

你的浏览器目前处于缩放状态，页面可能会出现错位现象，建议100%大小显示。

出现错误

赏

举报

用

阅读数 14450

最新评论

mac安装LightGBM wit...

Jancydc: [reply]jwy19900622[/reply] 请问你按照教程安装好了没有呀, 我按照教程在cmake ...

mac安装LightGBM wit...

jwy19900622: [reply]momo_mo520[/reply] 嗯, 已经搞定啦~ ~ 谢谢

mac安装LightGBM wit...

momo_mo520: [reply]jwy19900622[/reply] 删掉--without-multilib

mac安装LightGBM wit...

jwy19900622: 您好, 运行第二行代码提示: Error: invalid option: --without-multilib, 这个有 ...

ubuntu中ifconfig -...

wfh666: [reply]u014095069[/reply] 我也是这个问题, 感觉没找到门路。



程序人生



CSDN资讯

 QQ客服

 kefu@csdn.net

 客服论坛

 400-660-0108

工作时间 8:30-22:00

关于我们

招聘

广告服务

网站地图

京ICP备19004658号 经营性网站备案信息

 公安备案号 11010502030143

©1999-2020 北京创新乐知网络技术有限公司

网络110报警服务

北京互联网违法和不良信息举报中心

中国互联网举报中心 家长监护 版权申诉

1

你的浏览器目前处于缩放状态, 页面错位现象, 建议100%大小显示。

出现错



举报

https://blog.csdn.net/molu_chase/article/details/77508260

6/6