Python爬取网页转为PDF

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

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

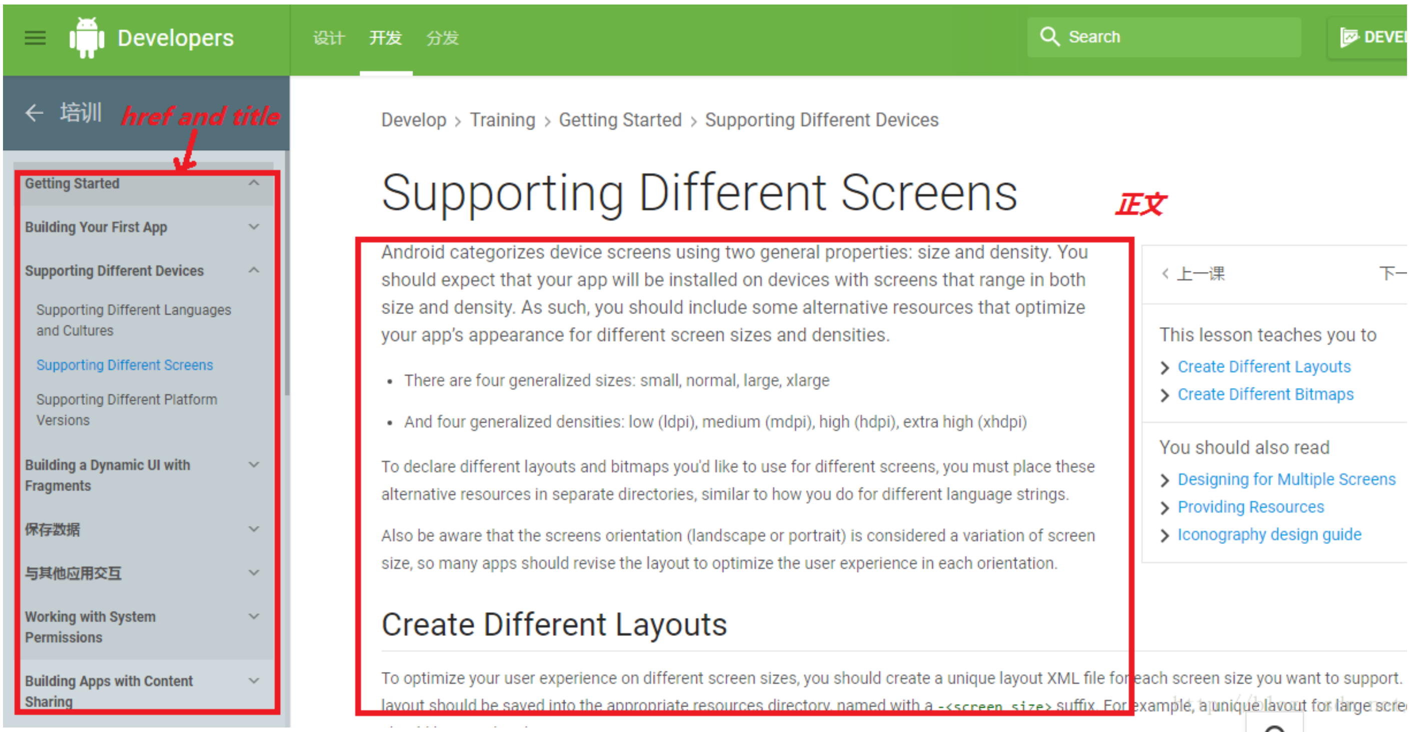
配置

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 址
如下所示：



Developers
设计
开发
分发
Search
DEVE
培训
href and title
Getting Started
Building Your First App
Supporting Different Devices
Supporting Different Languages and Cultures
Supporting Different Screens
Supporting Different Platform Versions
Building a Dynamic UI with Fragments
保存数据
与其他应用交互
Working with System Permissions
Building Apps with Content Sharing
Develop > Training > Getting Started > Supporting Different Devices
正文
< 上一课
下一
This lesson teaches you to
> Create Different Layouts
> Create Different Bitmaps
You should also read
> Designing for Multiple Screens
> Providing Resources
> Iconography design guide

Supporting Different Screens

Android categorizes device screens using two general properties: size and density. You should expect that your app will be installed on devices with screens that range in both size and density. As such, you should include some alternative resources that optimize your app's appearance for different screen sizes and densities.

- There are four generalized sizes: small, normal, large, xlarge
- And four generalized densities: low (ldpi), medium (mdpi), high (hdpi), extra high (xhdpi)

To declare different layouts and bitmaps you'd like to use for different screens, you must place these alternative resources in separate directories, similar to how you do for different language strings.

Also be aware that the screens orientation (landscape or portrait) is considered a variation of screen size, so many apps should revise the layout to optimize the user experience in each orientation.

Create Different Layouts

To optimize your user experience on different screen sizes, you should create a unique layout XML file for each screen size you want to support. layout should be saved into the appropriate resources directory, named with a `-<screen_size>` suffix. For example, a unique layout for large scre

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）

```
        "https":"https://vpn的IP:port",
    }

response=requests.get(url,proxies=proxies)
```

Beautiful Soup的使用

参考资料里面有Beautiful Soup文档，将其看完后，可以知道就讲了两件事：一个是查找标签，一个是修改标签。
本文需要做的是：

1. 获取标题和所有的网址,涉及到的是查找标签

```
#对标签进行判断，一个标签含有href而不含有description，则返回true
#而我希望获取的是含有href属性而不含有description属性的<a>标签，（且只有a标签含有href）
def has_href_but_no_des(tag):
    return tag.has_attr('href') and not tag.has_attr('description')

#网页分析，获取网址和标题
def parse_url_to_html(url):

    response=requests.get(url,proxies=proxies)
    soup=BeautifulSoup(response.content,"html.parser")
    s=[]#获取所有的网址
    title=[]#获取对应的标题
    tag=soup.find(id="nav")#获取第一个id为"nav"的标签，这个里面包含了网址和标题
    for i in tag.find_all(has_href_but_no_des):
        s.append(i['href'])
        title.append(i.text)

    #获取的只是标签集，需要加html前缀
    htmls = "<html><head><meta charset='UTF-8'></head><body>"
    with open("android_training_3.html",'a') as f:
        f.write(htmls)
```

1. 对上面获取的网址分析，获取正文，并将图片取出存于本地;涉及到的是查找标签和修改属性

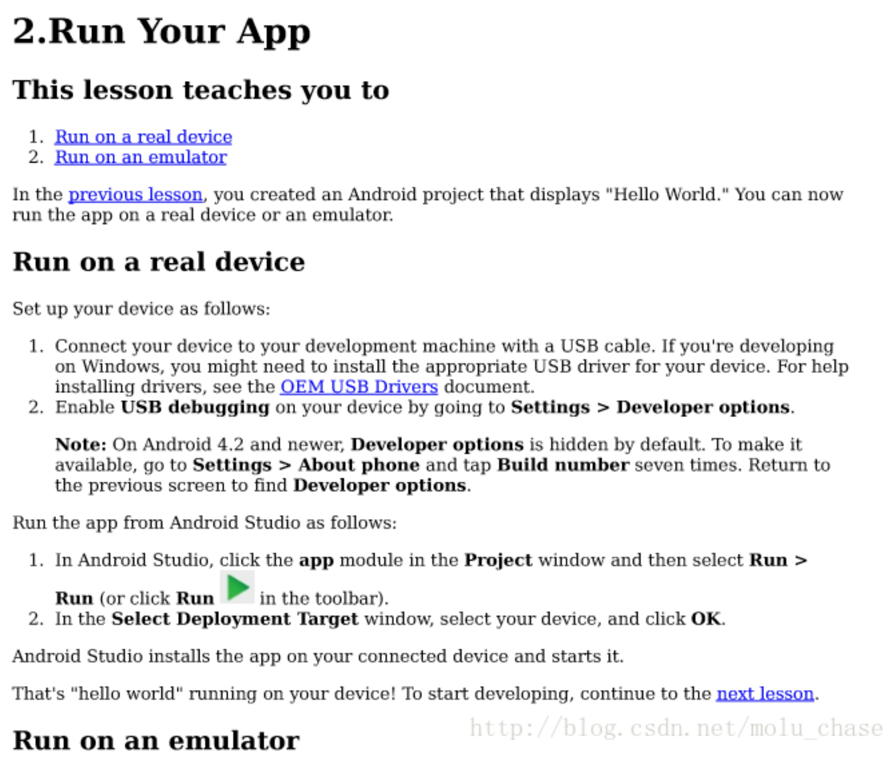
```
#网页操作，获取正文及图片
def get_htmls(urls,title):

    for i in range(len(urls)):
        response=requests.get(urls[i],proxies=proxies)
        soup=BeautifulSoup(response.content,"html.parser")
        htmls="<div><h1>"+str(i)+"."+title[i]+"</h1></div>"
        tag=soup.find(class_='jd-descr')
        #为image添加相对路径，并下载图片
        for img in tag.find_all('img'):
            im = requests.get(img['src'], proxies=proxies)
            filename = os.path.split(img['src'])[1]
            with open('image/' + filename, 'wb') as f:
                f.write(im.content)
            img['src']='image/'+filename
        htmls=htmls+str(tag)
        with open("android_training_3.html",'a') as f:
            f.write(htmls)
        print(" (%s) [%s] download end"%(i,title[i]))
    htmls="</body></html>"
    with open("android_training_3.html",'a') as f:
        f.write(htmls)
```

1. 转为PDF
这一步需要下载wkhtmltopdf,在Windows下执行程序一直出错..Ubuntu下可以

```
def save_pdf(html):
    """
  把所有html文件转换成pdf文件
```

```
        'page-size': 'Letter',
        'encoding': "UTF-8",
        'custom-header': [
            ('Accept-Encoding', 'gzip')
        ]
    }
pdfkit.from_file(html, "android_training_3.pdf", options=options)
```

最后的效果图



2.Run Your App

This lesson teaches you to

1. Run on a real device
2. Run on an emulator

In the previous lesson, you created an Android project that displays "Hello World." You can now run the app on a real device or an emulator.

Run on a real device

Set up your device as follows:

1. Connect your device to your development machine with a USB cable. If you're developing on Windows, you might need to install the appropriate USB driver for your device. For help installing drivers, see the OEM USB Drivers document.
2. Enable **USB debugging** on your device by going to **Settings > Developer options**.

   Note: On Android 4.2 and newer, **Developer options** is hidden by default. To make it available, go to **Settings > About phone** and tap **Build number** seven times. Return to the previous screen to find **Developer options**.

Run the app from Android Studio as follows:

1. In Android Studio, click the **app** module in the **Project** window and then select **Run > Run** (or click **Run** ▶ in the toolbar).
2. In the **Select Deployment Target** window, select your device, and click **OK**.

Android Studio installs the app on your connected device and starts it.

That's "hello world" running on your device! To start developing, continue to the next lesson.

Run on an emulator

moluchase
发布了177 篇原创文章 · 获赞 69 · 访问量 36万+

利用Python把网页内容转换为pdf格式文件，批量下载到本地！ 阅读数 1080

使用Google浏览器的打印命令时，保存下来的pdf文件中包含网页中的所有内容（左右边框和广告等），想仅把当前... 博文 来自： Python达人

使用Python将HTML转成PDF 阅读数

主要使用的是wkhtmltopdf的Python封装——pdfkit安装1. Install python-pdfkit:$ pip install pdfkit2. Install wk... 博文 来自： weixin_34192

将廖雪峰的学习教程转换成PDF文件，代码只适合该网站，如果需要其他网站的教程，可靠需要进行稍微的修改。# c... 博文 来自： 一步一个脚印

Python抓取HTML网页并以PDF保存

一、前言 今天介绍将HTML网页抓取下来，然后以PDF保存，废话不多说直接进入教程。今天的例子以廖雪峰老师的

你的浏览器目前处于缩放状态，页面 出现错 位现象，建议100%大小显示。

python写扫雷小游戏（pygame） 阅读数

学了python后，在9月初开始比赛，比赛类容是在一个星期内（白天有课，其实只有星期一到星期五晚上和双休有时... 博文 来自： qq_42847252的

用python爬虫批量下载pdf 阅读数

今天遇到一个任务，给一个excel文件，里面有500多个pdf文件的下载链接，需要把这些文件全部下载下来。我知道... 博文 来自： yllifesong的博客

python之html网页转PDF 阅读数

接上一篇，博主目前所要做的任务，除了要将图片转成pdf外，可能还需要根据爬去站点的内容来合成一篇pdf格式文... 博文 来自： zuo19960618

爬取含有PDF的网页下载

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 相关下载链接：//download.csdn.net... 论坛

python3爬虫下载网页上的pdf 阅读数 1万+

#coding=UTF-8#爬取大学nlp课程的教学pdf文档课件http://ccl.pku.edu.cn/alcourse/nlp/importurllib.requesti... 博文 来自： jonathanzh的博客

编译原理课程总结--第七章：语义分析和中间代码的产生 阅读数 3836

第七章：语义分析和中间代码的产生首先是语义分析的任务：（1）审查每一个语法结构的静态语义，即验证语法正... 博文 来自： 飞菜博客

Python抓取网页并保存为PDF 阅读数 1万+

抓取HTML文档，转化成PDF文档 博文 来自： shenwanjiang111...

燕大侠v
156篇文章
关注 排名:7000+

weixin_34192993
4616篇文章
关注 排名:千里之外

007与狼共舞
98篇文章
关注 排名:千里之外

Limerence
63篇文章
关注 排名:千里之外

Python爬取喜欢的博客，并将博客转成PDF 阅读数 114

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 阅读数 59

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... 博文 来自： weixin_44521703...

屏蔽百家号 -(baijiahao) 阅读数 2万+

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... 博文 来自： 慢慢积累

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... 问答

将网页内容保存为PDF及为PDF创建多级书签 阅读数 1723

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... 博文 来自： 鹏鹏的博客

python爬虫修改版.pdf 07-09

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... 下载

一种在windows下利用python中保存网页为pdf的方法 阅读数 44

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... 博文 来自： pikapika_chu的博客

[286]python将html转化为pdf 阅读数

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，可以一下子下载某个博主的所有文章了。但是，我们... 博文 来自： 周小董

举报

物联网大致可分为感知层、网络层、设备管理层、应用层等四个层次。其中最能体现物联网特征的，就是物联网的感... 博文 来自： tijos803的博客

1

你的浏览器目前处于缩放状态，页面 出现错 位现象，建议100%大小显示。

Python爬虫：抓取Python教程保存为PDF电子书

Github传送门：https://github.com/JosephPai/PythonCrawler-Html2Pdf 欢迎点赞~环境python3.6准备工具爬...

python下载网页转化成pdf

阅读

最近在学习一个网站补充一下cg基础。但是前几天网站突然访问不了了，同学推荐了waybackmachine这个网站，它... 博文 来自： banfan0440的博

Python爬虫：爬取在线教程生成pdf

阅读数

作为一名程序员，经常要搜一些教程，有的教程是在线的，不提供离线版本，这就有些局限了。那么同样作为一名程... 博文 来自： C与Python实战

python转html页面为pdf

阅读数

python转html页面为pdf：安装wkhtmltopdf略apt-get install python-pippip install pdfkitvi aa.py#!/usr/bin/p... 博文 来自： weixin_342145

赏

初试Python爬虫下载pdf

阅读数 3968

最近刚学完Boyd的ConvexOptimization，真是对Boyd神佩服得五体投地。在他的lectureslides末尾发现原来还有... 博文 来自： albertyzy的博客

火狐网页保存为mht(UnMHT) v7.2.0 官方版.zip

07-17

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... 下载

将网页转换成pdf文档的方法

阅读数 2678

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

浅谈利用python保存整个网站页面

阅读数 5010

空闲的时候随便找了一个网站练习一下爬虫，总结一下自己写爬虫遇到的知识点实现的功能抓取全站URL获取CSS，J... 博文 来自： gorquan的博客

web页面生成TXT文件供另存为下载

阅读数 23

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

Html页面保存为PDF

阅读数 3058

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

阅读数 7637

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_38944746的...

另类爬虫：从PDF文件中爬取表格数据

阅读数 145

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_33754065...

python 爬取网页内容并保存为pdf格式

09-16

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。 下载

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

python将网页上的教程爬取下来存成pdf

阅读

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_422963

举报

利用python3爬虫下载图片、pdf文档

阅读数

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

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

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

你的浏览器目前处于缩放状态，页面 出现错 位现象，建议100%大小显示。 ×

动态规划入门到熟悉，看不懂来打我啊

阅读数

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

Java学习的正确打开方式

阅读数 2

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

python json java mysql pycharm android linux json格式 c# 数组类型 泛型约束 c#的赛狗日程序 c# 传递数组 可变参数 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#

1

赏

©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	粉丝	获赞	评论	访问
177	58	69	34	36万+

等级: 博客 6 周排名: 7万+

积分: 5001 总排名: 1万+

勋章: 恒

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

举报

Java	33篇

展开

你的浏览器目前处于缩放状态，页面 出现错位现象，建议100%大小显示。

1

赏

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
阅读数 17487

关于setOnCheckedChangeListener的使用
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。

举报

你的浏览器目前处于缩放状态，页面 出现错位现象，建议100%大小显示。

1

赏

举报