



脚本之家
www.jb51.net



大咖网盟

日结/零扣量

万PV30~40元

广告



大资本联盟
dzbwm.com

实力海



熊掌联运
xiongzhangad.com

满百结算

不限行业

不限内容

九亿曝光

9CCMS.NET 暴力引流 快速变现 全新选择

同价收购 各种量各种技术等 QQ7959995

正版Windows 10 家庭/专业版 新年特价 248
OK资源站提供高清电影采集
★★★搜乐资源网 同步更新最快 采集补贴★★★
正版Windows 10 家庭/专业版 新年特价 248
5M独享云主机599/年
美港数据 高端香港服务器租用
动态拨号IP云主机, 电信ADSL独享20M
免备vps20/百独799/双线350/45互联
机房直销 每月半价抢购

免费网站全球加速和防御[优优加速]
【流量卡】【注册卡】全国通用 无需实名
Office 专业增强版 2019 正版办公软件 398元
宿迁100G高防VPS大带宽300元/月|创梦网络
群英云服务器送10M带宽30G防御,49元起
★云服务器5折, 天天抽红包抵扣★
[香港双高防]无视CC+DDOS/堪比广东!
中原地区核心数据中心, 月付299元起
香港高防10m大带宽独服, 低至999元

Office 专业增强版 2019 正版办公软件 398元
高价收【网赚粉】等各类活粉Q6883741
港湾网络徐州百独800/月,100G高防云150
【阿里云双11】亿元补贴-爆款产品一折起
服务器租用/托管-域名空间/认准腾佑科技
移动BGP, 百兆独享带宽, 399元/月
畅游网络 百独服务器 包跑满 998元
群英网络 300G高防仅需599元
韩国香港美国站群服务器 巨牛网络

不限内容【海外】独立服务器租用
香港/美国/韩国等十五国云服务
正版Windows 10 家庭/专业版 新年特价 248
天翼云, 云主机, 云存储, 云安全
华为云4核8g限时免费送另有代
高防ip 高防服务器租用
bgp多线机房、大带宽
浦东数据中心上海电信4星云主
热热资源、稳定更新、采集送料

免费服务器

无需实名流量卡

云服务器 仅售8元

上海真略 APP定制专家
规划.设计.开发 一站式APP定制服务

仅售



独家资源 免费采集

免费采集送福利
最适合小白站长的采集站

超清资源 每日更新

极速无弹窗 最全伦理

采集交流 QQ

python爬取网页转换为PDF文件

更新时间: 2018年06月07日 15:02:19 作者: moluchase 我要评论

这篇文章主要为大家详细介绍了python爬取网页转换为PDF文件, 具有一定的参考价值, 感兴趣的小伙伴们可以参考一下



图表数据分析



python课程



全网舆情监测



数据分析图表



python开发



语音转换文字

速度提升90%
精准度提升80%

MX MASTER 3
无线鼠标
全新电磁滚轮



logitech 罗技

站长推荐

正版 Windows 10
正版Windows 10 家庭/专业版 新年特价 248
操作系统限时抢购[¥248]

阿里云

云服务器ECS 8折
高性能云服务器低至
限量抢购

腾亿网络-佛山50G-400G
网站专用 老

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。于是便开始想着将Android的官方手册爬下来。

全篇的实现思路

- 分析网页
- 学会使用BeautifulSoup库
- 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

大家感兴趣的内容

- Python入门教程 超详细
- Python 列表(List)操作方
- Python 元组(Tuple)操作
- Python 字典(Dictionary)
- pycharm 使用心得 (一)

https://www.jb51.net/article/141646.htm

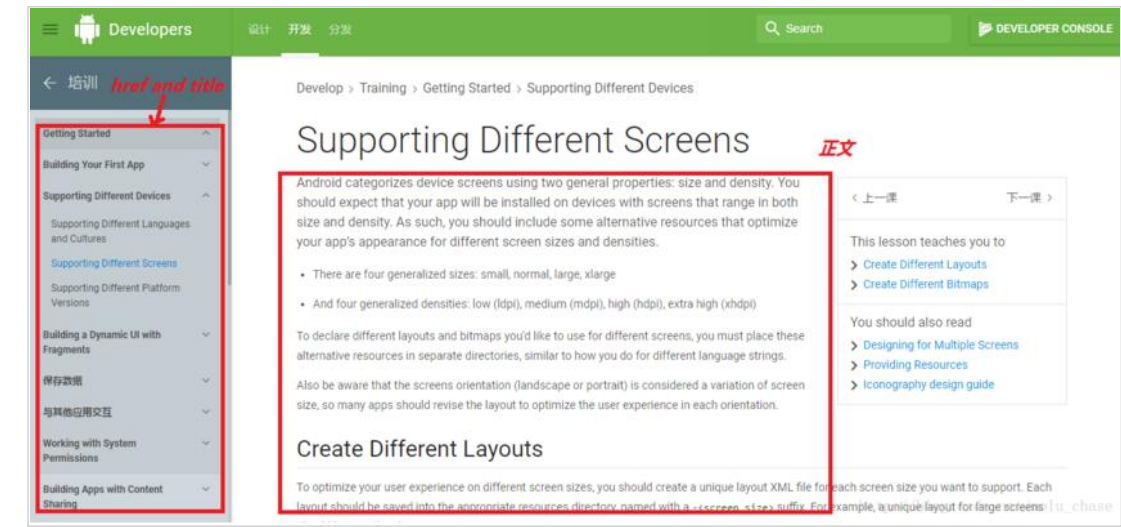
1/5

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

如下所示的一个网页,要做的是获取该网页的正文和标题, 以及左边导航条的所有网址



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)

```
1 proxies={
2     "http":"http://vpn的IP:port",
3     "https":"https://vpn的IP:port",
4 }
5
6 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,涉及到的的是查找标签

```
1 #对标签进行判断, 一个标签含有href而不含有description, 则返回true
2 #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)
3 def has_href_but_no_des(tag):
4     return tag.has_attr('href') and not tag.has_attr('description')
5
6 #网页分析, 获取网址和标题
7 def parse_url_to_html(url):
8
9     response=requests.get(url,proxies=proxies)
10    soup=BeautifulSoup(response.content,"html.parser")
11    s=[]#获取所有的网址
12    title=[]#获取对应的标题
13    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
14    for i in tag.find_all(has_href_but_no_des):
15        s.append(i['href'])
16        title.append(i.text)
17
18    #获取的只是标签集, 需要加html前缀
19    htmls = "<html><head><meta charset='UTF-8'></head><body>"
20    with open("android_training_3.html",'a') as f:
21        f.write(htmls)
```

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

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

- 6 python strip()函数 介绍
- 7 python 中文乱码问题深
- 8 Python科学计算环境推
- 9 python逐行读取文件内
- 10 python中使用xlrd、xlw

精准.高能.
掌握全局.
罗技大师系列

logitech 罗技

创梦

宿迁100G+

移动8GP900元起 40

- 最近更新的内容
- 在Django的视图中使用fo
 - Python遍历numpy数组的
 - 在python3中pyqt5和may
 - Python openpyxl 遍历所
 - PyCharm代码提示忽略大
 - TensorFlow实现打印每一
 - python解析模块(ConfigP
 - Python List cmp()知识点
 - Python实现二叉搜索树
 - Python 字符串定义

在线工具
代码格式化等

高防主机
600G 防护

- 常用在线小工具
- CSS代码工具
 - JavaScript代码格式化工具
 - 在线XML格式化/压缩工具
 - php代码在线格式化美化工
 - sql代码在线格式化美化工
 - 在线HTML转义/反转义工
 - 在线JSON代码检验/检验/
 - JavaScript正则在线测试工
 - 在线生成二维码工具(加强
 - 更多在线工具

百度云SEO专

收录快排名高 CDN加

亿恩云服务器 1核/2G/4

年付4折起, 49

```
3
4 for i in range(len(urls)):
5     response=requests.get(urls[i],proxies=proxies)
6     soup=BeautifulSoup(response.content,"html.parser")
7     htmls="<div><h1>"+str(i)+"."+title[i]+"</h1></div>"
8     tag=soup.find(class_='jd-descr')
9     #为image添加相对路径,并下载图片
10    for img in tag.find_all('img'):
11        im = requests.get(img['src'], proxies=proxies)
12        filename = os.path.split(img['src'])[1]
13        with open('image/' + filename, 'wb') as f:
14            f.write(im.content)
15        img['src']='image/'+filename
16    htmls=htmls+str(tag)
17    with open("android_training_3.html",'a') as f:
18        f.write(htmls)
19    print(" (%s) [%s] download end"%(i,title[i]))
20    htmls="</body></html>"
21    with open("android_training_3.html",'a') as f:
22        f.write(htmls)
```

Pytho
爬取余
详

Py
登
_p

Py
静
详

2.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

```
1 def save_pdf(html):
2     """
3     把所有html文件转换成pdf文件
4     """
5     options = {
6         'page-size': 'Letter',
7         'encoding': "UTF-8",
8         'custom-header': [
9             ('Accept-Encoding', 'gzip')
10        ]
11    }
12    pdkit.from_file(html, "android_training_3.pdf", options=options)
```

最后的效果图



数据可视化方法

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

<http://blog.csdn.net/moluchase>

以上就是本文的全部内容，希望对大家的学习有所帮助，也希望大家多多支持脚本之家。

您可能感兴趣的文章:

Python实现合并同一个文件夹下所有PDF文件的方法示例


Python结合ImageMagick实现多张图片合并为一个pdf文件的方法

python使用pdfminer解析pdf文件的方法示例


浅谈python实现Google翻译PDF,解决换行的问题

python实现从pdf文件中提取文本,并自动翻译的方法


- python爬取网页内容转换为PDF文件
- python实现pdf转换成word/txt纯文本文件
- Python解析并读取PDF文件内容的方法
- Python多图片合并PDF的方法
- Python实现html转换为pdf报告(生成pdf报告)功能示例

如您对本文有所疑义或者对本文内容提供补充建议，请联系小编 **QQ交谈**，本站会保留修改者版权

原文链接：<https://blog.csdn.net/moluchase/article/details/77508260>



扫描右侧二维码
关注脚本之家



你与百万开发者在一起

微信公众号搜索“**脚本之家**”，选择关注
程序猿的那些事、送书等活动等着你

python 爬取 pdf




编程大牛的网盘资料

关注有福利




相关文章

- 


4个月零基础掌握人工智能Python编程

Python编程培训

1.2万阅读
- 


基于python的图片修复程序（实现水印去除）

这篇文章主要给大家介绍了关于python图片修复程序的相关资料，可以用于实现图片中水印去除，主要利用的是OpenCV这个框架实现的，文中通过示例代码介绍的非常详细，需...

2018-06-06
- 


python刷投票的脚本实现代码

这篇文章主要介绍了写了个python刷投票的脚本,需要的朋友可以参考下

2014-11-11
- 


python集合删除多种方法详解

这篇文章主要介绍了python集合删除多种方法详解,文中通过示例代码介绍的非常详细，对大家的学习或者工作具有一定的参考学习价值,需要的朋友可以参考下

2020-02-02
- 

Python中py文件引用另一个py文件变量的方法

下面小编就为大家分享一篇Python中py文件引用另一个py文件变量的方法，具有很好的参考价值，希望对大家有所帮助。一起跟随小编过来看看吧

2018-04-04
- 

利用标准库fractions模块让Python支持分数类型的方法详解

最近在工作中遇到了分数处理，查找相关的资料发现可以利用Fraction类来实现，所以下面这篇文章主要给大家介绍了关于利用标准库fractions模块让Python支持分数类型的相关资...

2017-08-08



Python实现银行账户资金交易管理系统

这篇文章主要介绍了Python银行账户资金交易管理系统，本文通过实例代码给大家介绍的非常详细，具有一定的参考借鉴价值,需要的朋友可以参考下

2020-01-01



java判断三位数的实例讲解

在本文里小编给大家整理了关于java怎么判断三位数的实例方法以及要点总结，需要的朋友们跟着学习下。

2019-06-06



如何基于python操作excel并获取内容

这篇文章主要介绍了如何基于python操作excel并获取内容,文中通过示例代码介绍的非常详细，对大家的学习或者工作具有一定的参考学习价值,需要的朋友可以参考下

2019-12-12



python re正则表达式模块(Regular Expression)

Python 的 re 模块（Regular Expression 正则表达式）提供各种正则表达式的匹配操作，在文本解析、复杂字符串分析和信息提取时是一个非常有用的工具。

2014-07-07



python映射列表实例分析

这篇文章主要介绍了python映射列表,实例分析了python映射列表遍历计算其中每一个元素的使用技巧,需要的朋友可以参考下

2015-01-01



羊奶粉排行榜10强有哪些

品牌羊奶粉排行榜

4.0万阅读

141646

最新评论