秦川小道士

Python|网页转PDF,PDF转图片爬取校园课表~

```
import pdfkit
import requests
from bs4 import BeautifulSoup
from PIL import Image
from pdf2image import convert_from_path


def main():
    header={
    "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8,application/signed-exchange;v=b3",
        "Referer": "http://192.168.10.10/kb/",
        "Accept-Language": "zh-CN,zh;q=0.9",
        "Content-Type": "application/x-www-form-urlencoded",
        "Accept-Encoding": "gzip, deflate",
        "Connection": "Keep-Alive",
        "User-Agent":"Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/74.0.3729.169 Safari/537.36",
        "Accept-Encoding": "gzip, deflate",
        "Origin": "http://192.168.10.10",
        "Upgrade-Insecure-Requests": "1",
        "Cache-Control": "max-age=0",
        "Content-Length": "113"
    }

    url = 'http://192.168.10.10/kb/index.php/kcb/kcb/submit' #这是所在学校的课表查询响应页面，外网不可访问!

    yx = ["1院信息工程学院", "2院智能制造与控制术学院","3院外国语学院","4院经济与管理学院","5院艺术与设计学院" ]
    ulist = []
    n = 0

#自动获取班号
    kburl = 'http://192.168.10.10/kb/'#这是所在学校的课表查询查询页面，外网不可访问!
    r = requests.get(kburl)
    r.encoding = r.apparent_encoding
    soup2 = BeautifulSoup(r.text, 'html.parser')
    script = soup2.find('script', {'language': "JavaScript", 'type': "text/javascript"})  # 获取js片段
    bjhs = script.text[13:-287].split(',\r\n\r\n')   # 截取js需求区间，以空格符号为界,此处对嵌入式js处理!
    bjh = []
    for bjhx in range(5):
        a = bjhs[bjhx][1:-1].replace('"', '')  # 删除多余引号
        bjh.append(a.split(','))  # 追加新数组，字符串转化为数组

#以下开始爬取课表
    path = input('请粘贴存储地址：')  #手动输入文件保存地址
    for i ,j in zip(yx,bjh):#以学院进行外循环
        for g in range(len(j)):#以班号进行内循环
            data = {"province": i,
                    "bjh": j[g],
                    "Submit": "查 询"}#post查询提交参数

            Gg = path + r'\\'+ str(j[g]) + '.html'            #爬取网页暂存地址
```

昵称： 秦川小道士
园龄： 10个月
粉丝： 0
关注： 6
+加关注

< 2020年2月 >

日	一	二	三	四	五	六
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
1	2	3	4	5	6	7

搜索

找找看

谷歌搜索

常用链接

我的随笔
我的评论
我的参与
最新评论
我的标签

我的标签

Python(4)
爬虫(4)
微信(2)
智慧树(1)
wkhtmltopdf(1)
Django(1)
poppler(1)

随笔分类

Error(1)
Python随笔(4)

随笔档案

2019年8月(1)
2019年7月(3)
2019年6月(1)

阅读排行榜

1. 解决Django:UnicodeDecodeError: 'utf-8' codec can't decode byte 0xcb in position 325(1100)
2. Python|智慧树答案爬取(272)
3. Python|网页转PDF,PDF转图片爬取校园课表~(124)

```
                Pp = path + r'\\'+ str(j[g]) + '.pdf'              #网页转pdf暂存地址
                Pu = path + r'\\'+ str(j[g]) + '.jpeg'            #pdf转图片暂存地址
                r = requests.post(url,data=data,headers=header)     #发起查询请求，获取响应页面
                soup = BeautifulSoup(r.content,'html.parser')      #解析网页格式
                body = soup.find_all(name='body')               #爬取响应内容的课表部分
                html = str(body)                         #转换内容格式，方便后续操作。（此处为调错添加）
                with open(Gg,'w',encoding='utf-8') as f: #保存爬取到的课表，html格式
                    f.write(html)
                    f.close()

                #以上过程，课表爬取结束，初始爬取结果为html。以下为格式处理过程(html-pdf,pdf-.jpg)
                Pppath_wk = r'D:\wkhtmltopdf\bin\wkhtmltopdf.exe'# wkhtmltopdf安装位置
                #Pupath_wk = r'D:\wkhtmltopdf\bin\wkhtmltoimage.exe'  #这里原准备用它pdf来转图片
                Ppconfig = pdfkit.configuration(wkhtmltopdf=Pppath_wk) #设置调用程序路径位置（环境变量）
                #Puconfig = pdfkit.configuration(wkhtmltopdf=Pupath_wk)


                options1 = {
                    'page-size':'Letter',
                    'encoding':'UTF-8',
                    'custom-header': [('Accept-Encoding', 'gzip')]
                } #options1为设置保存pdf的格式
                '''options2 = {
                    'page-size': 'Letter',
                    'encoding': 'base64',
                    'custom-header': [('Accept-Encoding', 'gzip')]
                }'''#options2为设置保存图片的格式，未使用到，注释以便后续研究
                pdfkit.from_file(Gg,Pp,options=options1,configuration=Ppconfig)#转换html文件为pdf
                #pdfkit.from_file(Gg,Pu,options=options2,configuration=Puconfig)

                try:
                    convert_from_path(Pp, 300, path, fmt="JPEG", output_file=str(j[g]), thread_count=1) #pdf转为图片格式，此处注意保存路径的设置！

                except(OSError, NameError):
                    pass

                n+=1
                print('正在打印第%s张课表! ' % n)
        print("*" * 100)
        print('%s打印完毕! '% str(i))



main()

'''
**********第一版本需手动输入班级列表格式（供参考）************
    bjh = [
["10111501","10111502","10111503","10111504","10121501","10121502","10121503","10131501","10141501","10111503","10111504","10121503","ZB0111501","ZB0131501","ZB0141501","10111601","10111602","10111603","10121601","10121602","10131601","10141601","10161601","ZB0111601","ZB0121601","ZB0131601","10111701","10111702","10111703","10111704","10111705","10121701","10121702","10121703","10131701","10141701","10161701","ZB0111701","10211501","10211502","10211503","10211504","10211505","10221501","10221502","10221503","10231501","10231502","10241501","10241502","ZB0211501","ZB0221501","10211601","10211602","10221601","10231601","10241601","ZB0211601","ZB0221601","ZB0231601","10211701","10211702","10221701","10231701","10241701","ZB0211701","101011801","101011802","101011803","101011804","101021801","101021802","101021803","101031801","101041801","101051801","101051802","101061801","101071801","201011801","201051801"],

["10611501","10611502","10611503","10611504","10621501","10641501","10641502","10641503","ZB0641501","ZB0611501","10611601","10611602","10611603","10621601","10641601","10641602","ZB0611601","ZB0641601","10611701","10611702","10621701","10641701","10641702","ZB0611701","10911501","10911502","10921501","10921502","10931501","10931502","ZB0911501","ZB0921501","10911601","10921601","10931601","10911701","10931701","102011801","102011802","102021801","102031801","102041801","102041802","102051801","202011801","202051801"],


```

4. Python|每日定时对指定好友发送随机消息~(72)
5. Python|早安日报，微信批量发送消息~(38)

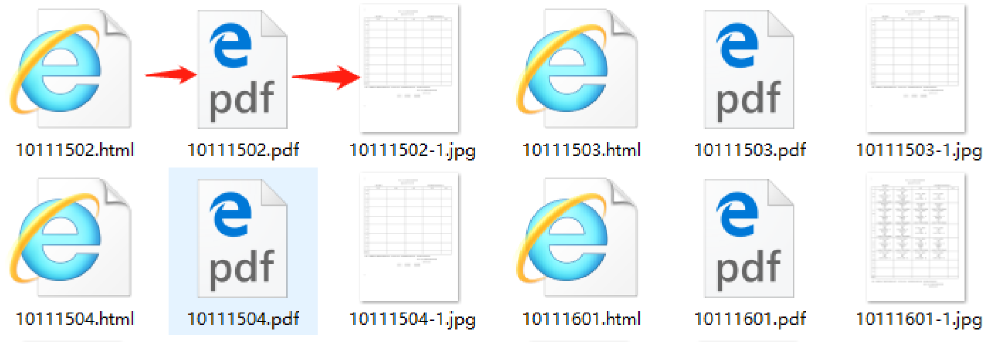
```
["10311501","10311502","10311503","10331501","10341501","ZB0311501","10311601","10311602","10311603","10311604","10311605","10311606","10321501","10321601","10331601","10331602","10341601","10351601","ZB0311601","10311701","10311702","10311703","10311704","10311705","10311706","10311707","10321701","10331701","10331702","10341701","10351701","ZB0311701","SX0341701","103011801","103011802","103011803","103011804","103011805","103011806","103011807","103011808","103011809","103031801","103031802","103041801","103051801","203011801"],

["10411501","10411502","10421501","10451501","10451502","10451503","10451504","10451505","10451506","ZB0451501","ZB0411501","10411601","10411602","10421601","10451601","10451602","10451603","10451604","10451605","ZB0411601","ZB0451601","10411701","10411702","10421701","10451701","10451702","10451703","ZB0411701","ZB0451701","ZB0451702","SX0411701","10711501","10731501","10731502","10731503","10731504","10731505","10731506","10731507","10731508","10731509","ZB0711501","ZB0731501","10711601","10731601","10731602","10731603","10731604","10731605","10731606","10731607","10731608","10731609","10731610","10731611","10731612","10741601","10741602","ZB0711601","ZB0731601","ZB0731602","ZB0731603","10711701","10731701","10731702","10731703","10731704","10731705","10731706","10731707","10741701","10741702","ZB0711701","ZB0731701","ZB0731702","ZB0731703","SX0711701","104011801","104011802","104021801","104021802","104021803","104031801","104031802","104041801","104051801","104051802","104051803","104051804","104051805","104051806","104051807","104051808","104051809","104061801","104061802","204021801","204021802","204031801","204041801","204051801","204051802","204051803","204051804"],

["10511501","10511502","10521501","10521502","10521503","10531501","10531502","10531503","10541501","10541502","10541503","ZB0521501","ZB0521502","ZB0511501","10511601","10511602","10511603","10521601","10521602","10521603","10521604","10531601","10531602","10531603","10531604","10541601","ZB0511601","ZB0521601","10511701","10511702","10521701","10521702","10521703","10521704","10531701","10531702","10531703","10531704","10541701","ZB0511701","ZB0521701","105011801","105011802","105011803","105021801","105021802","105021803","105021804","105021805","105031801","105031802","105031803","105031804","105031805","105041801","205011801","205021801"]
]

**********制作人：秦小道************
**********版本号：第二版************
********发布日期：2019.6.21*********
'''
```

爬取结果预览图：



爬取过程中碰到了许多错误，比如poppler与wkhtmltopdf为引入软件，需要将其bin目录添加至环境变量path中；

整个脚本只写了主函数~，习惯有大问题，得慢慢纠正！

整个脚本都做了注释，其中爬取地址为局域网址，如需参考，请按需求更改~

打包为.exe文件使用的是pyinstaller,但文件打包后仍需将poppler与wkhtmltopdf文件手动加入，到新的环境需手动设置这俩个应用的环境变量。

分类：Python随笔

标签：Python，poppler，wkhtmltopdf，爬虫

好文要顶 关注我 收藏该文

秦川小道士
关注 - 6
粉丝 - 0
+加关注

0 0

» 下一篇： Python|智慧树答案爬取

posted @ 2019-06-30 10:15 秦川小道士 阅读(124) 评论(0) 编辑 收藏

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册， 访问 网站首页。

相关博文：

- .Net 把网页Html转PDF文件
- 批量图片转PDF
- C#简单实现office转pdf、pdf转图片
- DOC转PDF工具
- python网页转pdf

» 更多推荐...

精品问答：大数据常见问题之 flink 五十问

最新 IT 新闻：

- 果粉缅怀苹果教父：今天是乔布斯65岁生日！
- Windows 10 v2004驱动升级有变：允许自动更新
- 卧底“传销式”算命网站，骂医务、骗宝妈
- 莫斯科公交司机人工识别中国乘客 人脸识别系统也启用
- 病毒来袭：人类战胜病毒的“终极拐点” 到底在哪里?

» 更多新闻...