2021年10月21日 21:46

# 实例：小型网站新闻爬虫词云图

2021年10月20日
19:13

## 一，目标分析

### 1选取目标

华东交大“交大要闻”新闻页面
https://xw.ecjtu.edu.cn/
https://xw.ecjtu.edu.cn/jdyw/684.htm


### 2网页结构分析

（1）页面目录

a,文章目录

第1页：https://xw.ecjtu.edu.cn/jdyw.htm

第2页：https://xw.ecjtu.edu.cn/jdyw/684.htm

第3页：https://xw.ecjtu.edu.cn/jdyw/683.htm

第4页：https://xw.ecjtu.edu.cn/jdyw/682.htm

第5页：https://xw.ecjtu.edu.cn/jdyw/681.htm

……

第685页：https://xw.ecjtu.edu.cn/jdyw/1.htm

b,文章页面

经过总结如右侧：https://xw.ecjtu.edu.cn/info/1041/xxxxx.htm

（2）同页结构

```
    <div class="list fl">
|
<script language="javascript" src="/system/resource/js/centerCutImg.js"></script><script language="javascript" src="/system/resource/js/ajax.js"></script><ul>

            <li><a href="http://www.ecjtu.edu.cn/info/1053/30806.htm" target="_blank" title="【我为师生办实事】多举措营造良好引人留人环境">【我为师生办实事】多举措营造良好引人留人环境</a><i>[2021-10-12]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30805.htm" target="_blank" title="学校召开机械工程学科建设研讨会">学校召开机械工程学科建设研讨会</a><i>[2021-10-11]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30802.htm" target="_blank" title="广大师生热议纪念辛亥革命110周年大会">广大师生热议纪念辛亥革命110周年大会</a><i>[2021-10-09]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30800.htm" target="_blank" title="我校获批3项2021年度国家社科基金项目">我校获批3项2021年度国家社科基金项目</a><i>[2021-10-05]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30797.htm" target="_blank" title="我校在第十七届“挑战杯”全国大学生课外学术科技作品竞赛中获奖">我校在第十七届“挑战杯”全国大学生课外学术科技作品竞赛中获奖</a><i>[2021-10-03]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30793.htm" target="_blank" title="缅怀革命先烈 传承红色基因 我校开展烈士纪念日主题教育活动">缅怀革命先烈 传承红色基因 我校开展烈士纪念日主题教育活动</a><i>[2021-09-30]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30791.htm" target="_blank" title="学校召开党委会专题学习会">学校召开党委会专题学习会</a><i>[2021-09-30]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30788.htm" target="_blank" title="校党委理论学习中心组举行2021年度第十三次集体学习会">校党委理论学习中心组举行2021年度第十三次集体学习会</a><i>[2021-09-30]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30787.htm" target="_blank" title="学校召开50周年校庆总结大会">学校召开50周年校庆总结大会</a><i>[2021-09-30]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30785.htm" target="_blank" title="“互联网+”国赛总决赛备赛动员宣传与培训会议召开">“互联网+”国赛总决赛备赛动员宣传与培训会议召开</a><i>[2021-09-29]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30784.htm" target="_blank" title="专家组反馈进校实地考查机械设计制造及其自动化、测绘工程专业工程教育认证情况">专家组反馈进校实地考查机械设计制造及其自动化、测绘工程专业工程教育认证情况</a><i>[2021-09-29]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30783.htm" target="_blank" title="专家组来校对机械设计制造及其自动化、测绘工程专业进行工程教育认证">专家组来校对机械设计制造及其自动化、测绘工程专业进行工程教育认证</a><i>[2021-09-27]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30776.htm" target="_blank" title="我校2021级“联合培养专升本”新生开学典礼举行">我校2021级“联合培养专升本”新生开学典礼举行</a><i>[2021-09-27]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30768.htm" target="_blank" title="2021年度江西省自然科学基金项目评审结果公示，我校名列前茅">2021年度江西省自然科学基金项目评审结果公示，我校名列前茅</a><i>[2021-09-26]</i></li>
            <li><a href="http://www.ecjtu.edu.cn/info/1053/30759.htm" target="_blank" title="农工党华东交通大学支部委员会举行换届选举大会">农工党华东交通大学支部委员会举行换届选举大会</a><i>[2021-09-24]</i></li>

        </ul>

    <link rel="stylesheet" Content-type="text/css" href="/system/resource/css/pagedown/sys.css"><div class="pb_sys_common pb_sys_normal pb_sys_style3" style="margin-top:10px;text-align:center;"><span class="p_pages"><span class="p_first p_fun"><a href="../jdyw.htm">首页</
  </div>
 </div>
</div>
```

结构在list fl，<a>标签下

（3）文章页面结构

```
<div class="content-con ">
    <div id="vsb_content"><div class="v_news_content">
<span class="size12">p>我校学生教学信息员制度已实施一年。日前已评出首批优秀学生教学信息员35名。</p>  <p>该制度的实施充分发挥了学生参与教学管理和自我管理的主体作用，加强了我校教风
</div> <p align="right">责任编辑：admin</p></div><div id="div_vote_id"></div>
```

**结构在content-con，<p>标签下**

其他情况下存在渲染等非预期字符，故需要考虑白名单过滤汉字。

```
<div class="content-con ">
    <div id="vsb_content"><div class="v_news_content">
<span class="size12"><p><span style="font-size: 9pt; font-family: 宋体; mso-ascii-font-family: 'Times New Roman'; mso-hansi-font-family: 'Times
</div> <p align="right">责任编辑：admin</p></div><div id="div_vote_id"></div>
```

## 二，爬虫关键代码

### 1调用bs4库

from urllib.request import urlopen

```
from bs4 import BeautifulSoup

html = urlopen('url')
bs = BeautifulSoup(html.read(), 'html.parser')
#print(bs)
```

### 2爬取并截获文章内容

```
#爬取content-title的标题
nameList=bs.findAll('div',{'class':'content-title'})
h3_txt=[]
for i in nameList:
    h3_txt.append(str(i.findAll('h3')))

#爬取文章内容(p)
nameList=bs.findAll('div',{'class':'v_news_content'})
p_txt=[]
for i in nameList:#由于tag的属性get_text需要针对每一个变量
    for j in i.findAll('span'):
        p_txt.append(str(j.get_text()))
```

### 3处理文章内容，并保存到列表

```
nameList=bs.findAll('h3')#标签爬取

strlist=[]
pattern = re.compile(r'[\u4e00-\u9fa5]+')#白名单过滤中文字符
for i in nameList:
    p = re.findall(pattern, str(i))
    if str(p)!='[]':#过滤为空的情况
        strlist.append(p)
    else:
        continue
print(strlist)
```

### 4爬取第一页url

```
import re
from urllib.request import urlopen
from bs4 import BeautifulSoup

html = urlopen('https://xw.ecjtu.edu.cn/jdyw.htm')#每页15篇文章内容
bs = BeautifulSoup(html.read(), 'html.parser')

#爬取第一页url
url_txt=[]
nameList=bs.findAll('div',{'class':'list fl'})
for i in nameList:
    for link in i.findAll('a'):
        if 'href' and 'target' in link.attrs:#根据基本格式抓取文章URL
            url_txt.append(link.attrs['href'])#保存到列表

print(url_txt)
```

## 三，词云关键代码

```
def img_grearte():
    mask = imread("xin.jpeg")
    with open("txt_save.txt", "r") as file:
        txt = file.read()
    word = WordCloud(background_color="white",
                     width=800,
                     height=800,
                     font_path='simhei.ttf',
                     mask=mask,
```

```
stopwords=['的','是','和','要','了','有','与','也','将','就','对','在','到','也','人','为','等','我']
                ).generate(txt)
    word.to_file('test.png')
    print("词云图片已保存")
```

## 四，其他记录

### 1python requirement

需要安装以下库

Bs4,Re,wordcloud,imageio

### 2遇到安装wordcloud库的问题

解决方式： https://blog.csdn.net/DCclient/article/details/89818315

### 3遇到from scipy.misc import imread 报错

使用词云脚本的时候发生的错误

解决方式： https://www.cnblogs.com/Yanjy-OnlyOne/p/12032030.html

### 4处理../的意外情况

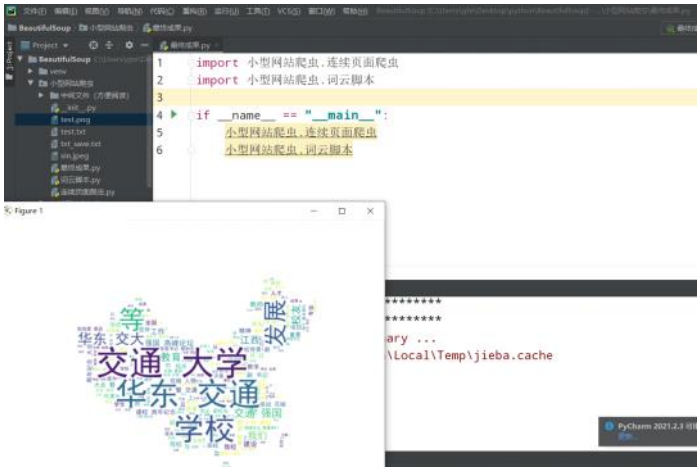
```
for i in nameList:
    for link in i.findAll('a'):
        if 'href' and 'target' in link.attrs:  # 根据基本格式抓取文章URL
            if '..' in link.attrs['href']:#处理../的意外情况
                url=link.attrs['href'].replace('..','https://xw.ecjtu.edu.cn')
                url_txt.append(url)# 保存到列表
            else:
                url_txt.append(link.attrs['href'])
```

### 5最终结果

简单调用同目录下的两个脚本

```
import小型网站爬虫.连续页面爬虫
import小型网站爬虫.词云脚本

if__name__=="__main__":
小型网站爬虫.连续页面爬虫
小型网站爬虫.词云脚本
```




## 五，脚本分析

### 1连续页面爬取.py

修改爬取页数：

```
62    for i in range(5):#指定页数为5页
63        url="https://xw.ecjtu.edu.cn
64        f = open("C:\\Users\\qte\\Des
65        #print(url)
66        f.write(page_go(url))
```

**2词语脚本.py**

修改禁用词：

```
62    for i in range(5):#指定页数为5页
63        url="https://xw.ecjtu.edu.cn
64        f = open("C:\\Users\\qte\\Des
65        #print(url)
66        f.write(page_go(url))
```