



山东工商学院

Python 编程技术

学院： 信息与工程学院

班级： 电子信息专硕 2102 班

姓名： 董超

学号： 2021420073

Python 编程技术

任务描述

编写网络爬虫程序，爬取山东工商学院 2015 年 1 月 1 日之后的所有新闻，统计每年所有新闻中出现次数最多的 10 个词语并分别绘制柱状图显示。

代码

Getadata.py

```
1. import requests
2. from lxml import etree
3. from datetime import datetime
4. '''
5. 说明：
6.     本部分完成的是数据的获取,通过 endtime 控制爬取数据的截止日期,将每条新闻分别保存
   到 path 目录下
7. '''
8. if __name__ == '__main__':
9.     url = "https://www.sdtbu.edu.cn/index/ssyw.htm"
10.    path = "result/"
11.    endtime = "2015-01-01"
12.    headers = {
13.        'User-
14.        Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML,
15.        like Gecko) Chrome/89.0.4389.82 Safari/537.36'
16.    }
17.    TOPURL = "https://www.sdtbu.edu.cn/"
18.    statue = {
19.        200: "OK",
20.        404: "Not Found"
21.    }
22.    i = 1
23.    # 爬虫部分
24.    while True:
25.        try:
```

```

25.         # 新闻列表页进行请求
26.         response = requests.get(url=url, headers=headers)
27.         response.encoding = response.apparent_encoding
28.         print(url, f"第{i}页:的爬取状态:", statue[response.status_code])
29.
30.         # 创建解析器
31.         html = response.text
32.         parse = etree.HTML(html)
33.
34.         # 解析页面中的新闻详情页 url
35.         a_path = "//div[4]/div[2]/ul/li/a/@href" # 获取新闻详情页 url
36.         a_lst = parse.xpath(a_path)
37.         t_path = '//div[4]/div[2]/ul/li/a/@title' # 获取新闻 title
38.         t_lst = parse.xpath(t_path)
39.         ti_path = '//div[4]/div[2]/ul/li/p/text()' # 获取时间 time
40.         ti_lst = parse.xpath(ti_path)
41.
42.         # 对每个详情页面进行请求
43.         for a, t, ti in zip(a_lst, t_lst, ti_lst):
44.             newurl = TOPURL + a[2:]
45.             newresponse = requests.get(newurl, headers=headers)
46.             newresponse.encoding = newresponse.apparent_encoding
47.             # print("\t", newurl, "详情页:", newresponse.status_code)
48.             newhtml = newresponse.text
49.
50.             # 创建解析器
51.             newparse = etree.HTML(newhtml)
52.
53.             # 文字
54.             infopath = '//*[@id="vsb_content"]//text()'
55.             info = newparse.xpath(infopath)
56.
57.             text = ""
58.             # 去除转义字符
59.             for in_ in info:
60.                 if in_ == "\r\n" or in_ == "\r\n " or in_ == "\uffff" or in_ == "\xa0\xa0":
61.                     pass
62.                 else:
63.                     text += in_
64.             t = t.replace(r'"', r'""')
65.
66.             y = datetime.strptime(ti, '%Y-%m-%d') # 新闻发布的时间

```

```

67.             z = datetime.strptime(endtime, '%Y-%m-%d') # 截止日期
68.
69.             diff = z - y
70.             chazhi = diff.days * 86400 + diff.seconds # 时间差
71.
72.             # 设置时间限制
73.             if chazhi <= 0:
74.                 # print("y = ", y, "diff = ", diff)
75.                 filepath = f"{path}" + ti + " " + t + ".txt"
76.
77.                 with open(filepath, encoding=newresponse.encoding, mode=
                    "w") as fp:
78.                     fp.write(text)
79.             else:
80.                 exit(0)
81.
82.             # 获取"下页"的 url
83.             nextpath = '//table//div/a[@class="Next"]/@href'
84.             nexturl = parse.xpath(nextpath)
85.             url = TOPURL + "index/ssyw/" + nexturl[0][-7:]
86.
87.         except Exception as e:
88.             print(e)
89.
90.         i += 1

```

haufen.py

```

1. import os
2. '''
3. 说明:
4.     将爬取到的新闻按照年份汇总,保存到 yearpath 目录,每个文件文件名为年份
5. '''
6. if __name__ == '__main__':
7.     path = "result/"
8.     yearpath = "resultByyear/"
9.     # 按年进行新闻合并
10.    every_year_lst = {}
11.    # print(os.listdir("./result"))
12.    text = ''
13.    for in_ in os.listdir(path):
14.        fp = open(path + in_, mode="r", encoding='utf-8')
15.
16.        # 比较时间

```

```
17.         if int(in_[:4]) == 2022:
18.             f = fp.read().replace("\n", "")
19.             f = f.replace(" ", "")
20.             text += f
21.             every_year_lst[2022] = text
22.
23.         if int(in_[:4]) == 2021:
24.             f = fp.read().replace("\n", "")
25.             f = f.replace(" ", "")
26.             text += f
27.             every_year_lst[2021] = text
28.
29.         if int(in_[:4]) == 2020:
30.             f = fp.read().replace("\n", "")
31.             f = f.replace(" ", "")
32.             text += f
33.             every_year_lst[2020] = text
34.
35.         if int(in_[:4]) == 2019:
36.             f = fp.read().replace("\n", "")
37.             f = f.replace(" ", "")
38.             text += f
39.             every_year_lst[2019] = text
40.
41.         if int(in_[:4]) == 2018:
42.             f = fp.read().replace("\n", "")
43.             f = f.replace(" ", "")
44.             text += f
45.             every_year_lst[2018] = text
46.
47.         if int(in_[:4]) == 2017:
48.             f = fp.read().replace("\n", "")
49.             f = f.replace(" ", "")
50.             text += f
51.             every_year_lst[2017] = text
52.
53.         if int(in_[:4]) == 2016:
54.             f = fp.read().replace("\n", "")
55.             f = f.replace(" ", "")
56.             text += f
57.             every_year_lst[2016] = text
58.
59.         if int(in_[:4]) == 2015:
60.             f = fp.read().replace("\n", "")
```

```

61.         f = f.replace(" ", "")
62.         text += f
63.         every_year_lst[2015] = text
64.
65.     # 分年份保存
66.     for x, y in every_year_lst.items():
67.         filepath = yearpath + str(x) + ".txt"
68.         with open(filepath, encoding="utf-8", mode="w", ) as fp:
69.             print(x, "年的保存完成！")
70.             fp.write(y)

```

fenci.py

```

1. import os
2. import jieba
3. import matplotlib.pyplot as plt
4.
5. plt.rcParams['font.sans-serif'] = ['SimHei'] # 指定默认字体 SimHei 为黑体
6. plt.rcParams['axes.unicode_minus'] = False # 用来正常显示负
7.
8. '''
9. 说明:
10. 使用结巴分词,对每一年的新闻进行分词,并完成词频统计,截取词频前十的词语,绘制柱状
    图
11. '''
12. if __name__ == '__main__':
13.     yearpath = "resultByyear/"
14.     pic = "pic/"
15.     for in_ in os.listdir(yearpath):
16.         timelst = []
17.         wordlst = []
18.         # print("年份: ",in_[:4])
19.         fp = open(yearpath + in_, mode="r", encoding='utf-8')
20.         text = fp.read()
21.
22.         # 结巴分词
23.         cut_text = (jieba.lcut(text))
24.
25.         dic = {}
26.         for word in cut_text:
27.             if word not in dic:
28.                 dic[word] = 1
29.             else:
30.                 dic[word] += 1

```

```

31.
32.     # 统计每个词出现次数，从高到低排序
33.     swd = sorted(list(dic.items()), key=lambda lst: lst[1], reverse=True
    )
34.
35.     # 排除那些虚词，连词，标点符号等
36.     f1 = open('中文虚词列表.txt', encoding="utf-8")
37.     stop_wds = f1.read()
38.     f1.close()
39.
40.     count = 0
41.     for kword, times in swd:
42.         if kword not in stop_wds and count <= 10: # 当前词未包含在排除的
            那些词里面，就输出出现次数
43.             count += 1
44.             timelst.append(times)
45.             wordlst.append(kword)
46.             # print(kword, times)
47.
48.     p1 = plt.figure(figsize=(8, 6), dpi=80) # 确定画布大小
49.     plt.title(f'{in_[:4]}年份词频统计') # 设置标题
50.     plt.bar(range(len(timelst)), timelst, fc='blue') # 绘制柱状图
51.
52.     for a, b in zip(list(range(len(timelst))), timelst):
53.         plt.text(a, b, '%.1f' % b, ha='center', va='bottom', fontsize=10
    ) # 添加数据标签
54.     plt.xticks(range(len(timelst)), wordlst)
55.
56.     plt.savefig(f'{pic + in_[:4]}年份词频统计.png')
57.     # plt.show()
58.
59.     fp.close()
60.     # break

```

pyechart.py

```

1. from pyecharts.charts import Bar
2. from pyecharts import options as opts
3. import os
4. import jieba
5. from pyecharts.globals import ThemeType
6.
7. if __name__ == '__main__':
8.     yearpath = "resultByyear/"

```

```
9.     pic = "pic/"
10.    for in_ in os.listdir(yearpath):
11.        timelst = []
12.        wordlst = []
13.        # print("年份: ",in_[:4])
14.        fp = open(yearpath + in_, mode="r", encoding='utf-8')
15.        text = fp.read()
16.        cut_text = (jieba.lcut(text))
17.
18.        dic = {}
19.        for word in cut_text:
20.            if word not in dic:
21.                dic[word] = 1
22.            else:
23.                dic[word] += 1
24.
25.        swd = sorted(list(dic.items()), key=lambda lst: lst[1], reverse=True
    ) # 统计每个词出现次数，从高到低排序
26.
27.        f1 = open('中文虚词列表.txt', encoding="utf-8") # 排除那些虚词，连词，
    标点符号等
28.        stop_wds = f1.read()
29.        f1.close()
30.        count = 0
31.        for kword, times in swd:
32.            if kword not in stop_wds and count <= 10: # 当前词未包含在排除的
    那些词里面，就输出出现次数
33.                count += 1
34.                timelst.append(times)
35.                wordlst.append(kword)
36.        # 使用 pyechart 绘制柱状图
37.        bar = (
38.            Bar(init_opts=opts.InitOpts(theme=ThemeType.WALDEN))
39.                .add_xaxis(wordlst)
40.                .add_yaxis(str(in_[:4]), timelst)
41.                .set_global_opts(title_opts=opts.TitleOpts(title=f'{in_[:4]}
    年份词频统计'))
42.        )
43.        bar.render(f'{pic + in_[:4]}年份词频统计.html')
44.        # break
```


运行结果





爬虫部分运行结果

result/2021-11-04 排查安全隐患 创建平安校园 确保师生安全——学校领导班子成员深入一线开展校园安全专项检查活动.txt
result/2021-11-03 我校一课程入选全省干部教育培训研修课程.txt
result/2021-11-02 深秋送暖：白光昭走访慰问学校退休老同志.txt
result/2021-11-01 【图文】强化责任 夯实基础 重点突破——学校召开党政领导班子特色建设务虚会.txt
result/2021-11-01 加强高水平项目研究 助推高质量发展进程——学校19项成果获批山东省研究生教育质量提升计划和教育创新计划项目.txt
result/2021-11-01 山海乘风二十载，初心弥坚展新篇——数学与信息科学学院（大数据学院）举行建院20周年庆典.txt
result/2021-10-31 构建完善教学质量保障体系——学校召开本学期教学督导工作会议.txt
result/2021-10-30 【图文】搭建平台促交流 携手同心共发展——山东工商学院校友会临沂代表处成立.txt
result/2021-10-30 【图文】推动校地深度融合 赋能烟台经济社会高质量发展——学校与烟台市政府举行校地合作座谈会.txt
result/2021-10-30 【图文】强化担当意识，积极主动作为 奋力推动学校科研工作高质量发展——学校召开国家“两金”项目申报总结及动员部署会.txt
result/2021-10-29 【图文】原山东省副省长、人大常委会副主任贾耕来校考察调研.txt
result/2021-10-29 【图文】加强校际交流 深化开放办学——山东建筑大学校长于德湖一行来校访问交流.txt
result/2021-10-29 学习贯彻习近平法治思想，提升全省政府立法工作质量——学校承办全省政府立法业务培训班.txt
result/2021-10-29 【图文】学习贯彻习近平总书记重要讲话精神 努力开创学校特色发展新局面——学校党委理论学习中心组召开学习会.txt
result/2021-10-28 【图文】学校荣获“高校教育新闻宣传先进单位”.txt
result/2021-10-28 提升新闻舆论传播力 凝聚事业发展精气神——学校召开新闻宣传工作专题会议.txt
result/2021-10-28 【图文】加强特色教材建设 夯实财商教育基础——学校召开财商教育特色教材建设推进会.txt

result/2015-01-20 新生生活系列数学实验课开始顺利结束.txt
result/2015-01-20 【图文】刘新生、郭金创一行赴“第一书记”帮包村走访慰问.txt
result/2015-01-16 3671名学子喜获本学年国家奖励助学金.txt
result/2015-01-16 【图文】刘新生到基建处、丰岛经济研究院、浙商研究院调研.txt
result/2015-01-15 学生处携手烟台交运集团开展“交运助学暖心帮帮”活动.txt
result/2015-01-15 【图文】党委中心组本期第六次学习会议召开.txt
result/2015-01-13 刘新生一行到上海对外经贸大学调研.txt
result/2015-01-12 【图文】团省委来校部来我校调研共青团工作.txt
result/2015-01-09 山东大学一行来我校交流.txt
result/2015-01-09 【图文】谭秀森、隋松毅到统计学院社情民意调查中心调研.txt
result/2015-01-09 【图文】新一轮岗位设置与聘用工作启动.txt
result/2015-01-09 【图文】学校校友工作座谈会召开.txt
result/2015-01-08 校图书馆报委员会2014年度会议召开.txt
result/2015-01-08 资产管理与后勤保障系统工作调度会召开.txt
result/2015-01-06 校报编辑部推荐我校2014年十件大事.txt
result/2015-01-05 我校民主党派集体及个人获烟台市2014年度多项表彰.txt

将新闻按照年份进行分类保存结果

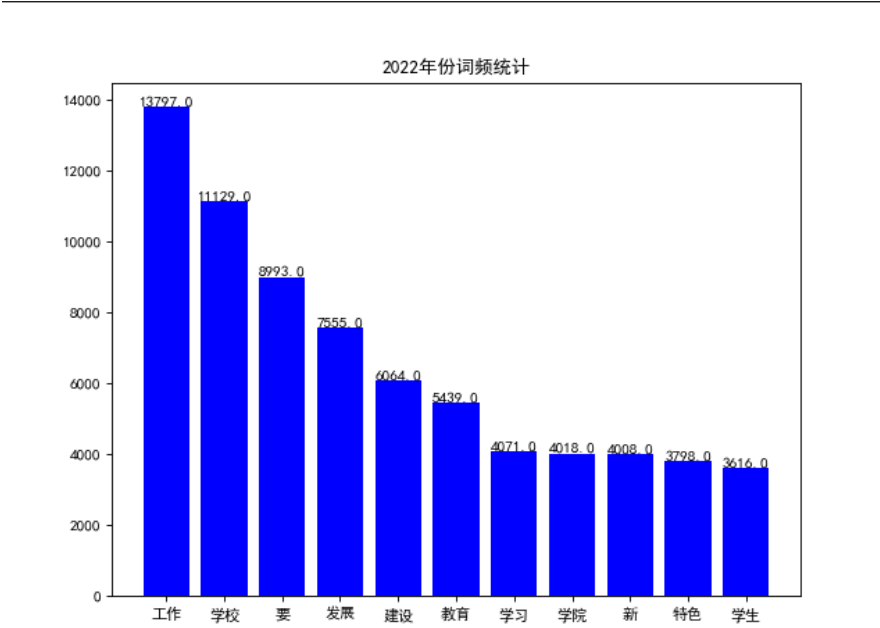
2015 年的保存完成！
2016 年的保存完成！
2017 年的保存完成！
2018 年的保存完成！
2019 年的保存完成！
2020 年的保存完成！
2021 年的保存完成！
2022 年的保存完成！

名称	修改日期	类型	大小
 2015.txt	2022/6/17 14:32	文本文档	908 KB
 2016.txt	2022/6/17 14:32	文本文档	1,800 KB
 2017.txt	2022/6/17 14:32	文本文档	2,404 KB
 2018.txt	2022/6/17 14:32	文本文档	3,329 KB
 2019.txt	2022/6/17 14:32	文本文档	4,336 KB
 2020.txt	2022/6/17 14:32	文本文档	5,296 KB
 2021.txt	2022/6/17 14:32	文本文档	6,187 KB
 2022.txt	2022/6/17 14:32	文本文档	6,427 KB

分词并统计词频部分

```
"D:\Program Files\Anaconda3\envs\pachong\python.exe" C:/Users/32243/Desktop/1
年份： 2015
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\32243\AppData\Local\Temp\jieba.cache
Loading model cost 0.991 seconds.
Prefix dict has been built successfully.
工作 1833
学校 1376
要 1129
发展 902
等 806
我校 770
建设 707
教育 688
学院 684
月 608
学生 545
年份： 2016
工作 3690
学校 2748
要 2507
发展 1819
等 1486
建设 1467
教育 1464
我校 1372
学院 1303
学习 1118
月 1089
年份： 2017
```

利用 matplotlib 制作柱状图



利用 pyechart 制作柱状图

