

哈尔滨工业大学

<<信息检索>>

实验报告

(2022 年度春季学期)

姓名:	王艺丹
学号:	1190201303
学院:	计算机学院
教师:	张宇

实验一 网页文本预处理

一、实验目的

对信息检索中网页文本预处理的流程和涉及的技术有一个全面的了解,包括抓去网页、网页正文提取、分词处理、停用词处理等环节。

二、实验内容

(一)网页的抓取和正文提取

爬取网页(至少 1000 个,其中包含附件的网页不少于 100 个,多线程实现爬虫可加分),然后提取网页标题和网页正文,以及网页中的附件并保存附件到本地 `craw.json` 文件中。

(二)分词处理、去停用词处理

将提取的网页文本进行分词和去停用词处理,并将结果前十行保存至本地 `preprocessed.json` 文件中

三、实验过程及结果

综述:

`utils.py`: 实现各功能函数复用

- `get_stop_dic(file_path)`: 构建停用词字典树
- `tokenize(content:str)`: 分词及去停用词处理
- `pages_to_json(data, save_path)`: 将网页数据存储在 json 格式,每一行对应一个网页的 json 数据
- `read_pages(file_path)`: 读入路径下的离线网页数据,转为字典格式

`PageCraw.py`: 类 `Page()`, 实现单独网页爬取

- `load_url(url, att_dir, tokenizer=False)`: 静态方法,无需实例化,实现从 url 进行数据爬取,并将附件存储在 `att_dir` 路径下的功能, `tokenizer` 参数控制是否对标题及内容进行分词
- `show(self)`: 将数据转为字典格式输出,保护数据完整性

`MultiCraw.py`: 类 `Craw()`, 实现多线程网页数据爬取

- `multi_craw(att_dir, urls)`: 多线程爬取 `urls` 对应网站的网页数据,并将附件下载在 `att_dir` 目录下
- `bfs(url)`: 对当前 url 下根地址进行 bfs, 获取相关的其余连接
- `multi_geturls(urls)`: 对 `urls` 列表中的 url 多线程进行 bfs 广搜, 获取大量不同的 urls 列表
- `show(self)`: 将 `pages` 中的 `page` 以字典形式输出, 保护数据

craw.py: 进行多线程数据爬取

segment.py: 对数据进行分词以及去停用词处理

闪光点:

✧ 多线程数据爬取:

利用 concurrent 函数库中的 futures 实现

```
1. def multi_craw(self, att_dir, urls = []):
2.     executor = futures.ThreadPoolExecutor(max_workers=10)
3.     start = time.time()
4.     fs = []
5.     urls.extend(list(self.urls))
6.     for url in urls:
7.         # 提交任务到线程池
8.         # print(url)
9.         f = executor.submit(Page.load_url, url, att_dir)
10.        fs.append(f)
11.    # 等待这些任务全部完成
12.    futures.wait(fs)
13.    # 获取任务的结果
14.    for f in fs:
15.        if f.result() is False:
16.            continue
17.        else:
18.            self.cnt += 1 # 成功爬取网页个数
19.            rr = f.result()
20.            self.att_cnt += rr[1] # 带有附件的网页个数
21.            self.pages.append(rr[0])
22.    end = time.time()
23.    print(f'多线程爬虫耗时:{end-start}s')
24.    print(f'共成功爬取{self.cnt}个网页数据, 其中带附件网页共计{self.att_cnt}个')
25.    return self.pages
```

标题处理:

观察网页源码可发现其标题存储格式如下:

```
ink rel="token-devel" href="http://today.hit.edu.cn/node/y4555/devel/token" />
ink rel="version-tree" href="http://today.hit.edu.cn/node/94555/tree" />
```

```
<title>关于评选第八届哈尔滨工业大学创新创业教育优秀指导教师奖的通知 | 今日哈工大: 哈尔滨工业大学校内综合信息网</title>
<link rel="stylesheet" media="all" href="/sites/default.prod1.dpweb1.hit.edu.cn/files/css/css_TQ6QC5zJacyOya0-Pi8NNRlw3BrzIe
ink rel="stylesheet" media="all" href="/themes/custom/hit_front/css/devel-tools/panels-ipe.css?r810im" />
ink rel="stylesheet" media="all" href="/sites/default.prod1.dpweb1.hit.edu.cn/files/css/css_ZCffksEfGLxS2xyDaMKYiZknwMSANWY2hf
ink rel="stylesheet" media="all" href="/themes/custom/hit_front/css/style.css?r810im" />
ink rel="stylesheet" media="all" href="/sites/default.prod1.dpweb1.hit.edu.cn/files/css/css_3c6vwXne105nYweRpkLvI3QLMGJb7LQr
ink rel="stylesheet" media="all" href="/sites/default.prod1.dpweb1.hit.edu.cn/files/css/css_Fmqpmmcjc6HKE-rE7ST9dU3ITMGsMwVKqi
```

```
<div class="article-title text-center">
<h3>关于评选第八届哈尔滨工业大学创新创业教育优秀指导教师奖的通知</h3>
</div>
<div class="top_misc">
<div class="left-attr first">2022-05-09 17:46</div>
```

由于爬虫过程中遇到问题，发现大部分标题以第二种形式存储，少部分仅以第一种形式存储，所以通过条件判断获取文章标题，核心代码如下：

```
1. title = soup.find_all('div', {"class": "article-title text-center"})
2. if len(title)!=0:
3.     title = title[0].get_text().strip()
4.     title = title.split(' ')[0].split('|')[-1].split('/')[0].strip()
5. else:
6.     title = str(soup.title.string)
7. # print(title)
8. title = title.split(' ')[0].split('|')[0].split('/')[0].strip()
```

内容处理：

```
8 <link rel="canonical" href="http://today.hit.edu.cn/article/2022/05/09/94555" />
9 <meta name="description" content="各学院、学部、校区：
10
11 为坚持立德树人根本任务，落实人才培养中心地位，激励广大教师积极投入创新创业教育教学
12
13 一、申报范围
14
15 忠诚于党和人民的教育事业，自觉贯彻党的教育方针，模范遵守职业道德规范，教书育人，为
16
17 二、申报条件" />
```

最初按上图所示选取 soup 中 deacription 中 content 的内容进行提取，发现信息不全，再次观察源码发现内容各段均存储于：

```
d--name-body field--type-text-with-summary field--label-hidden field--item"><p align="left" style="text-align:left">各学院、学部、校区：</p>
t:28.0pt">为坚持立德树人根本任务，落实人才培养中心地位，激励广大教师积极投入创新创业教育教学工作，根据《哈尔滨工业大学创新创业教育优秀指导教师评选：
t:28.1pt">一、申报范围</p>
t:28.0pt">忠诚于党和人民的教育事业，自觉贯彻党的教育方针，模范遵守职业道德规范，教书育人，为人师表，积极投身本科生、研究生（以下统称学生）创新创业教
t:28.1pt">二、申报条件</p>
t:28.0pt">多年坚持组织、指导本科生大一年度项目、大学生创新创业训练计划项目、学科竞赛、创业孵化、本科生科研项目等学生创新创业教育教学工作，取得成果或
t:28.0pt">1. 开设高水平创新创业课程。</p>
2. 指导本科生大一年度项目和大学生创新创业训练计划项目合计5项以上，并在学校优秀项目评选中获得校一等奖奖励。</p>
3. 指导学生参加“互联网+”、“挑战杯”比赛获省级金奖（或一等奖）或入选国赛。</p>
4. 指导学生参加高水平学科竞赛，并获国家级二等奖及以上奖励；指导的学生项目入选大学生创新创业训练计划项目年会。</p>
5. 指导学生在SCI、SSCI、EI、ISTP、ISSHP等检索的期刊或会议上发表论文1篇及以上，或其他核心期刊上发表学术论文2篇及以上。学生为论文第一作者，指导教师为通
6. 指导学生申请发明专利（研究生申请与竞赛直接相关的专利），并被受理或已授权。学生为第一或第二申请人，指导教师为申请人之一。</p>
7. 指导学生创业实践并取得标志性成绩。</p>
8. 创新创业教育成果产生了示范和辐射性影响，扩大了学校创新创业教育工作在全国的影响力。</p>
```

最终采用提取 html 文件解析中所有标签 p 下的句子进行合并

核心代码如下：

```
1. # 提取正文，没有则用标题充当正文内容
2. content = soup.find_all('p')
3. if content is None:
4.     paragraphs = title
5. else:
6.     paragraphs = ''
7. for i in range(len(content)):
```

```
8. paragraphs = paragraphs + content[i].get_text()
```

附件处理:

```
<div class="field field-name-attachments field-type-file field-label-above">
  <div class="field-label">附件</div>
  <div class="field-items">
    <div class="field-item"><span class="file file--mime-application-vnd-openxmlformats-officedocument-wordprocessingml-document file--x-office-document"><a href="http://today.hit.edu.cn/sites/todayl.prod1.qweb1.hit.edu.
  </div>
</div>
</div>
```

附件存在于:

标签下, 可通过查找 span 标签下 class 为 file--x-office-document 的内容并解析获取附件名称, 判断其后缀是否为 (doc, docx, xlxs, txt), 若为, 则将其名称加入 file_name 列表中, 解析 href 标签获得下载连接, 加入 file_url 中; 其中某一网页可能对应多个附件, 各网页附件对应存储路径为 attachment/title/附件 1...附件 x

核心代码如下:

```
1. # 提取附件名称并下载附件
2. all_href = soup.find_all('span', {'class': 'file--x-office-document'})
3. file_name = []
4. download_url = [] # 记录附件下载地址
5. for h in all_href:
6.     h = h.select('a')
7.     cur_name = h[0].get_text() # 提取附件名称
8.     if cur_name.endswith(attachment_type): # 只提取指定类型的附件
9.         cur_url = h[0].get('href')
10.        download_url.append(cur_url)
11.        file_name.append(cur_name) # 添加附件名称至当前 cur_url 对应的附件名称列表
12.
13. if len(file_name) == 0:
14.     file_path = None
15. else: # 附件列表不为空, 下载附件
16.     att_cnt += 1
17.     if not (os.path.exists(file_path)):
18.         os.mkdir(file_path)
19.     for i in range(len(file_name)):
20.         urlretrieve(download_url[i], f'{file_path}/{file_name[i]}')
```

分词及去停用词处理:

停用词字典树构建:

在 utils.py 中实现字典树 TrieTree 结构, 提高停用词查找速率:

```

16 '''字典树结构, 构建停用词词典'''
17 class Trie:
18     def __init__(self):
19         self.root = {} # 用字典存储
20         self.end_of_word = '#' # 用#标志一个单词的结束
21
22     def insert(self, word: str):
23         node = self.root
24         for char in word:
25             node = node.setdefault(char, {})
26         node[self.end_of_word] = self.end_of_word
27
28     # 查找一个单词是否完整的存在于字典树里, 判断node是否等于#
29     def search(self, word: str):
30         node = self.root
31         for char in word:
32             if char not in node:
33                 return False
34             node = node[char]
35         return self.end_of_word in node
36
37 # In[27]:
38
39
40
41 '''基于路径文件构建停用词字典树'''
42 def get_stop_dic(file_path):
43     stop_dic = Trie()
44     with open(file_path, 'r', encoding='utf-8') as f:
45         for line in f.readlines():
46             stop_dic.insert(line.strip())
47     return stop_dic

```

进一步利用 ltp 进行分词处理, 查找停用词字典树去除停用词, 返回处理后的分词列表, 关键代码如下:

```

1. ht = HarvestText()
2. ltp = LTP()
3. def tokenize(content:str):
4.
5.     # 数据清洗部分
6.     content = ht.clean_text(content)
7.
8.     # 分词
9.     words = []
10.    segs,_ = ltp.seg([content])
11.    for w in segs[0]:
12.        if stop_dic.search(w) or w.strip()=='':
13.            continue
14.        else:
15.            words.append(w.strip())
16.
17.    return words # 返回分词结果列表

```

segment.py 分词去停用词处理:

使用 utils 中的 tokenize 函数对标题及内容进行处理，并覆盖原有数据

```
if __name__ == '__main__':
    # 读入按行存储的json格式的pages数据
    pages_dic = read_pages('../data/result/craw.json')

    # 进行分词及去停用词处理，并用新数据覆盖原始数据
    for i in range(len(pages_dic)):
        pages_dic[i]['segmented_title'] = tokenize(pages_dic[i].pop('title'))
        pages_dic[i]['segmented_paragraphs'] = tokenize(pages_dic[i].pop('paragraphs'))
        pages_dic[i]['file_name'] = pages_dic[i].pop('file_name') # 不改变数据格式顺序

    # 将处理好的数据前10行存储至指定目录下
    pages_to_json(pages_dic[:10], '../data/result/preprocessed.json')
```

craw.py 多线程爬虫:

利用 Craw 类先对 url 进行 bfs 获得网页树，再利用 multi_craw 进行多线程爬虫

```
if __name__ == '__main__':
    att_dir = '../data/attachment/' # 附件存储路径
    save_dir = '../data/result/' # 结果存储路径

    # bfs 根url获取多个url (线程池默认最大工作数目为5)
    uus = [f'http://today.hit.edu.cn/category/10?page={i}' for i in range(50)]
    mm = Craw()
    urls = mm.multi_geturls(uus)

    # 多线程爬取urls的网页数据，存储在指定路径文件下 (线程池默认最大工作数目为10)
    pages = mm.multi_craw(att_dir)
    pages_dic = mm.show()
    pages_to_json(pages_dic, f'{save_dir}craw.json')
```

实验中的问题:

数据爬取过程中可能遇到页面跳转的情况，查找本地浏览器的 User-Agent 及 cookie 等参数，构建请求头，在 url 中加入请求头在进行请求处理：

```
@staticmethod
def load_url(url, file_dir, tokenizer = False, attachment_type=('txt', 'doc', 'docx', 'xlsx')):
    att_cnt = 0
    # 爬虫时会遇到页面跳转的问题，携带请求头部进行数据爬取
    headers = {
        'Accept': 'application/json, text/javascript, */*; q=0.01',
        'Cookie': 'SESS0b78b3575298f2ed94ea5549d866ad3c=gCKCtOU7h4zyZGkmGp2Debo-I22w6dzCk16UfS-y8vU; Drupal.visitor.DRUPAL_UID=36477',
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/101.0.4951.54 Safari/537.36'
    }
    request = Request(url=url, headers=headers) # 加入请求头
```

BFS 广搜的网页中部分网页可能无访问权限，多线程爬虫过程中，可能由于方位次数过多过快 IP 被禁止，于是在网页请求过程中加入异常处理：

实验结果:

共爬取了 1081 个网页数据, 带附件网页共计 161 个; 多线程爬虫共耗时 104.52s

对 1081 个网页数据进行分词及去停用词处理, 用时 34s

部分数据展示如下：

此电脑 > 新加卷 (G:) > jupyter-notebook > 信息检索 > lab1-wyd > data > attachment

名称	修改日期	类型	大小
“4.23世界读书日”系列活动通知	2022-05-10 16:14	文件夹	
“风雨同舟·青春记忆”图书馆摄影比赛通知	2022-05-10 16:15	文件夹	
【不忘初心、牢记使命】数学学院关于2...	2022-05-10 16:15	文件夹	
【哈工大之春】关于举办首届哈尔滨工业...	2022-05-10 16:14	文件夹	
【回望百年，青春向团】土木学院团委...	2022-05-10 16:15	文件夹	
【心系国防	2022-05-10 16:15	文件夹	
【知理远航】物理学院关于开展“班级夺...	2022-05-10 16:15	文件夹	
2021-2022年度苏州育才奖学金评选通知	2022-05-10 16:15	文件夹	
2022	2022-05-10 16:14	文件夹	
2022年第六届全国大学生集成电路创新...	2022-05-10 16:14	文件夹	
2022年度国家自然科学基金申请有关事...	2022-05-10 16:14	文件夹	

« data > attachment > 【不忘初心、牢记使命】数学学院关于2019届毕业生党员德育答辩的通知

名称	修改日期	类型	大小
附件1 数学学院毕业生党员德育论文撰写...	2022-05-10 16:15	Microsoft Word ...	16 KB
附件2 数学学院毕业生党员德育论文范文...	2022-05-10 16:15	Microsoft Word ...	18 KB

附件

此电脑 > 新加卷 (G:) > jupyter-notebook > 信息检索 > lab1-wyd > data > result

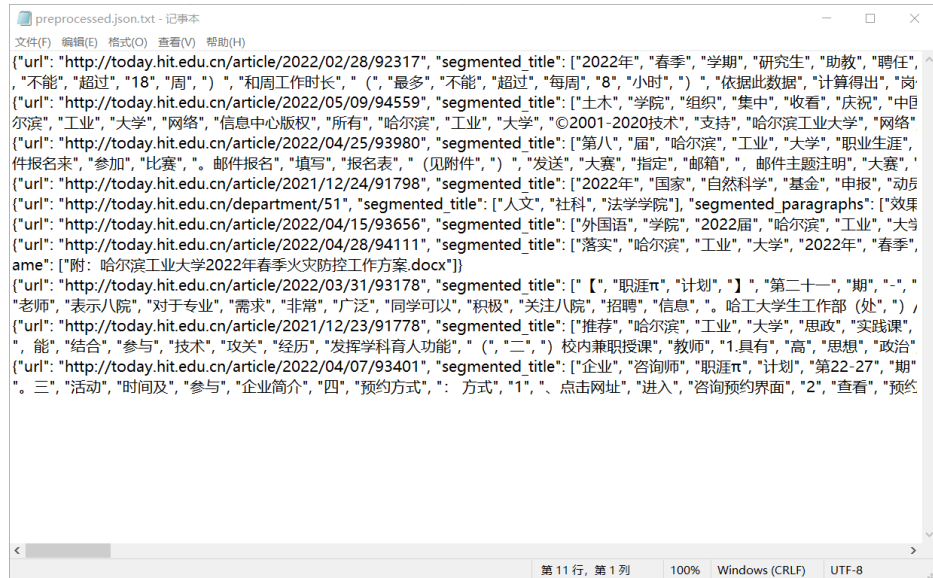
名称	修改日期	类型	大小
craw.json	2022-05-10 16:15	JSON File	2,518 KB
preprocessed.json	2022-05-10 16:16	JSON File	25 KB

craw - 副本.json.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

["url": "http://today.hit.edu.cn/article/2022/02/28/92317", "title": "关于2022年春季学期研究生助教聘任工作的通知", "paragraphs":
予以计算酬金。工作时长以周志中填写的总额（不得超过教师申报岗位时预设的总时长）为准。3.各学院（部）管理员统计酬金发放表时，
["url": "http://today.hit.edu.cn/article/2022/05/09/94559", "title": "土木学院关于组织集中收看庆祝中国共产党百年成立100周年
["url": "http://today.hit.edu.cn/article/2022/04/25/93980", "title": "第八届哈尔滨工业大学职业生涯规划大赛培训通知（第一期）", "
["url": "http://today.hit.edu.cn/article/2021/12/24/91798", "title": "2022年度国家自然科学基金申报动员会通知", "paragraphs": "版
["url": "http://today.hit.edu.cn/department/51", "title": "人文社科与法学院", "paragraphs": "效果演示效果演示哈尔滨工业大学 ©
["url": "http://today.hit.edu.cn/article/2022/04/15/93656", "title": "外国语学院2022届哈尔滨工业大学优秀毕业生", "paragraphs":
["url": "http://today.hit.edu.cn/article/2022/04/28/94111", "title": "关于落实《哈尔滨工业大学2022年春季火灾防控工作方案》的通知
["url": "http://today.hit.edu.cn/article/2022/03/31/93178", "title": "【生涯π计划】第二十一期-中国航天科技八院专场", "paragraph
["url": "http://today.hit.edu.cn/article/2021/12/23/91778", "title": "关于推荐哈尔滨工业大学思政实践课兼职授课教师的通知", "para
1和2），纸质版文件签字盖章后扫描成为PDF文件。各学院推荐材料（WORD文件和PDF文件）由学院思政实践课程组负责人（学院学
["url": "http://today.hit.edu.cn/article/2022/04/07/93401", "title": "企业咨询师来了！“生涯π”计划第22-27期等你来预约", "para
["url": "http://today.hit.edu.cn/article/2022/05/01/94259", "title": "讣告", "paragraphs": "哈尔滨工业大学机电工程学院退休教授张
["url": "http://today.hit.edu.cn/article/2022/04/12/93549", "title": "【电气大讲堂】电气学院关于认定秋季学期学生讲师带辅工作的
["url": "http://today.hit.edu.cn/category-ranking", "title": "类别信息数排行", "paragraphs": "版权所有：哈尔滨工业大学 ©2001-20
["url": "http://today.hit.edu.cn/article/2022/03/28/93052", "title": "图书馆图书借阅权限相关通知", "paragraphs": "根据学校工作提
["url": "http://today.hit.edu.cn/article/2022/01/16/92030", "title": "2022年度国家自然科学基金艺术学重大项目申报通知", "parag
→办事大厅→全国艺术科学规划项目申报管理系统，以下简称“系统”）https://yskx.mct.gov.cn/index进行在线申报，请申请人（首席
["url": "http://today.hit.edu.cn/article/2022/04/02/93243", "title": "【学理启航】物理学院“学知讲师团”之学知讲堂第六讲", "para
["url": "http://today.hit.edu.cn/article/2022/05/09/94555", "title": "关于评选第八届哈尔滨工业大学创新创业教育优秀指导教师奖的
奖申报佐证材料》（附件3，以下简称佐证材料），并将相关佐证材料原件一并提交学院（含学部 and 校区，下同）。佐证材料原件应包括以
["url": "http://today.hit.edu.cn/article/2022/04/26/93998", "title": "【洪晶讲堂—实验室宣讲】先进光子学材料与器件实验室", "para
—442-341-436欢迎同学们积极参加，共同探索神秘的物理学新世界！版权所有：哈尔滨工业大学 ©2001-2020技术支持：哈尔滨工业

< 第1行，第1列 100% Windows (CRLF) UTF-8



```
{
  "url": "http://today.hit.edu.cn/article/2022/02/28/92317",
  "segmented_title": ["2022年", "春季", "学期", "研究生", "助教", "聘任",
    , "不能", "超过", "18", "周", ")", "和周工作时长", "(", "最多", "不能", "超过", "每周", "8", "小时", ")", "依据此数据", "计算得出", "岗
    ({
      "url": "http://today.hit.edu.cn/article/2022/05/09/94559",
      "segmented_title": ["土木", "学院", "组织", "集中", "收看", "庆祝", "中巨
        尔滨", "工业", "大学", "网络", "信息中心版权", "所有", "哈尔滨", "工业", "大学", "©2001-2020技术", "支持", "哈尔滨工业大学", "网络"
      ({
        "url": "http://today.hit.edu.cn/article/2022/04/25/93980",
        "segmented_title": ["第八", "届", "哈尔滨", "工业", "大学", "职业生涯",
          件报名表", "参加", "比赛", "。 邮件报名", "填写", "报名表", " (见附件", ")", "发送", "大赛", "指定", "邮箱", "。 邮件主题注明", "大赛", "
        ({
          "url": "http://today.hit.edu.cn/article/2021/12/24/91798",
          "segmented_title": ["2022年", "国家", "自然科学", "基金", "申报", "动员
            ({
              "url": "http://today.hit.edu.cn/department/51",
              "segmented_title": ["人文", "社科", "法学院"],
              "segmented_paragraphs": ["效果
                ({
                  "url": "http://today.hit.edu.cn/article/2022/04/15/93656",
                  "segmented_title": ["外国语", "学院", "2022届", "哈尔滨", "工业", "大学",
                    ({
                      "url": "http://today.hit.edu.cn/article/2022/04/28/94111",
                      "segmented_title": ["落实", "哈尔滨", "工业", "大学", "2022年", "春季",
                        ame": ["附：哈尔滨工业大学2022年春季火灾防控工作方案.docx"]
                      ({
                        "url": "http://today.hit.edu.cn/article/2022/03/31/93178",
                        "segmented_title": ["【", "职涯π", "计划", "】", "第二十一", "期", "-", "
                          "老师", "表示八院", "对于专业", "需求", "非常", "广泛", "同学可以", "积极", "关注八院", "招聘", "信息", "。 哈工大学生工作部 (处", ")",
                            ({
                              "url": "http://today.hit.edu.cn/article/2021/12/23/91778",
                              "segmented_title": ["推荐", "哈尔滨", "工业", "大学", "思政", "实践课",
                                , "能", "结合", "参与", "技术", "攻关", "经历", "发挥学科育人功能", " (", "二", ") 校内兼职授课", "教师", "1.具有", "高", "思想", "政治"
                                  ({
                                    "url": "http://today.hit.edu.cn/article/2022/04/07/93401",
                                    "segmented_title": ["企业", "咨询师", "职涯π", "计划", "第22-27", "期"
                                      "。 三", "活动", "时间及", "参与", "企业简介", "四", "预约方式", "： 方式", "1", "、 点击网址", "进入", "咨询预约界面", "2", "查看", "预约
```

实验结果文件

四、实验心得

- 学会使用 BeautifulSoup 处理网页 html
- 学会使用 urlopen 请求网页数据下载文件
- 学习了 json 文件进行处理存储
- 学会了多线程操作，减少大量数据处理时间
- 学会了处理网页免登录携带请求头的申请方法
- 对异常有了更深刻的认知与学习异常处理