hw4

Web搜索引擎 - 南开资源站

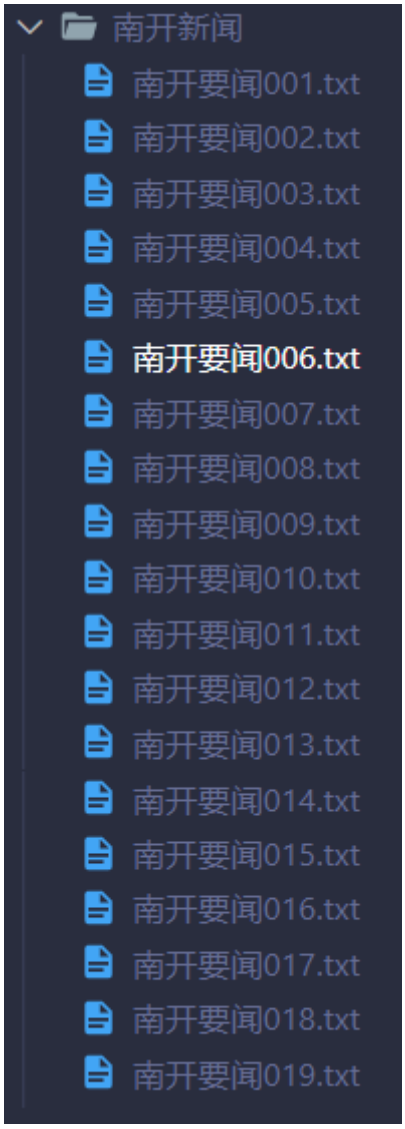
网页抓取

实验中我抓取了南开新闻网中的南开要闻(http://news.nankai.edu.cn/ywsd/index.shtml)板块，爬取其中的共533页网站，每页网站爬取的新闻条数都是30条。

主要使用了Python的requests包来爬取网站。

```
# 获取南开新闻网综合新闻的URL
html = requests.get('http://news.nankai.edu.cn/ywsd/index.shtml')
html_bytes = html.content
# 获取网页源代码
htmlcode = html_bytes.decode("utf-8")
```

并将爬取下来的网站写到记事本中，截取部分网站如下：




```
link:http://news.nankai.edu.cn/ywsd/system/2010/05/31/000031203.shtml, title:南开大学思想政治教育先进经验获中央肯定
link:http://news.nankai.edu.cn/ywsd/system/2010/05/31/000031201.shtml, title:全国加强和改进大学生思想政治教育工作座谈
link:http://news.nankai.edu.cn/ywsd/system/2010/05/31/000031200.shtml, title:南开大学：以铸以陶 校园文化高品质育人
link:http://news.nankai.edu.cn/ywsd/system/2010/05/31/000031199.shtml, title:南开大学：创新育人载体加强思想政治教育工
link:http://news.nankai.edu.cn/ywsd/system/2010/05/30/000031177.shtml, title:南开-弗林德斯大学合作培养医院管理专业硕
link:http://news.nankai.edu.cn/ywsd/system/2010/05/30/000031178.shtml, title:新疆调研宣讲团访问南开大学
link:http://news.nankai.edu.cn/ywsd/system/2010/05/29/000031175.shtml, title:首届天津市大学生物理竞赛颁奖大会南开举行
link:http://news.nankai.edu.cn/ywsd/system/2010/05/29/000031171.shtml, title:德国慕尼黑工业大学副校长访问南开
link:http://news.nankai.edu.cn/ywsd/system/2010/05/29/000031170.shtml, title:西南交通大学代表团访问南开
link:http://news.nankai.edu.cn/ywsd/system/2010/05/28/000031128.shtml, title:教育系统将深入开展创先争优活动
link:http://news.nankai.edu.cn/ywsd/system/2010/05/28/000031126.shtml, title:南开大学新闻网进入改版试运行阶段
link:http://news.nankai.edu.cn/ywsd/system/2010/05/27/000031098.shtml, title:澳大利亚维多利亚州总督克雷茨访问南开
link:http://news.nankai.edu.cn/ywsd/system/2010/05/27/000031086.shtml, title:台北商业技术学院校长访问南开大学
link:http://news.nankai.edu.cn/zhxw/system/2010/05/25/000030891.shtml, title:促进产学研结合 区校携手共建“科技南开”
link:http://news.nankai.edu.cn/zhxw/system/2010/05/23/000030831.shtml, title:国内外专家南开研讨“蛋白质介导的膜塑形与
link:http://news.nankai.edu.cn/zhxw/system/2010/05/23/000030832.shtml, title:著名历史学家杨生茂先生追思会举行
link:http://news.nankai.edu.cn/zhxw/system/2010/05/22/000030822.shtml, title:南开动员学子积极参加“百人计划”
link:http://news.nankai.edu.cn/zhxw/system/2010/05/22/000030820.shtml, title:海南大学代表团访问南开
link:http://news.nankai.edu.cn/zhxw/system/2010/05/22/000030819.shtml, title:学校召开分党委党总支书记会部署近期党建工
link:http://news.nankai.edu.cn/zhxw/system/2010/05/20/000030796.shtml, title:校领导访问华南理工大学
link:http://news.nankai.edu.cn/zhxw/system/2010/05/20/000030785.shtml, title:南开大学举行旅游与服务学院教职工座谈会
link:http://news.nankai.edu.cn/zhxw/system/2010/05/20/000030781.shtml, title:南开大学召开共青团表彰大会
link:http://news.nankai.edu.cn/zhxw/system/2010/05/17/000030721.shtml, title:校领导应邀赴港深两地做主题报告
link:http://news.nankai.edu.cn/zhxw/system/2010/05/13/000030485.shtml, title:新加坡大学生代表团访问南开
link:http://news.nankai.edu.cn/zhxw/system/2010/05/08/000030379.shtml, title:学校召开旅游与服务学院成立情况通报会
link:http://news.nankai.edu.cn/zhxw/system/2010/05/08/000030376.shtml, title:校领导出席天津商业大学建校30周年庆典
```

爬取的过程中，可以发现南开要闻网址的规律，网址是呈线性增加的，因此可以通过循环，比较容易的爬取到数据，同时将数据写入文本中。

```
for i in range(532, 0, -1):
    # 网页的页数有规律
```

```
    i = str(i).zfill(3)
    html = requests.get('http://news.nankai.edu.cn/ywsd/system/count//0003000/000000000000/000/000/c000300000000000000000_000000' + i + '.shtml')
    html_bytes = html.content
    htmlcode = html_bytes.decode("utf-8")

    titlere = re.compile('<a href="(.*?)" target="_blank">(.*?)</a></div></td>')
    titles = titlere.findall(htmlcode)
    filehandle = open('./南开新闻/南开要闻' + i + '.txt', mode='w', encoding='utf-8')
    for title in titles:
        text = "link:" + str(title[0]) + ", title:" + str(title[1])
        filehandle.write(text + '\n')
        count += 1
    filehandle.close()
```

我使用了Python的re包，即正则表达式来提取网页，在网页中打开网页源代码，找到需要爬取的网页的位置，如下图：

```
llspacing="0">
t"><a href="http://news.nankai.edu.cn/ywsd/system/2021/12/13/030049427.shtml" target="_blank">第二届全国数量经济技术经济学博士后论坛南开举行</a></div></td>
div align="right">2021-12-13</div></td>
```

然后将需要爬取的网址和标题用(.*?)代替，即可匹配需要爬取的网站，然后再写入文本中即可。

```
    titlere = re.compile('<a href="(.*?)" target="_blank">(.*?)</a></div></td>')
    titles = titlere.findall(htmlcode)
```

文本索引

文本索引与上次的思路大致相当，在爬取数据的过程中，考虑索引的构建，因此在爬取的时候在网址前加关键字 `link` ，在标题前加关键字 `title` ，那么建索引可以依据这两个关键字来建立。

```
def create_index():
    mappings = {
        "mappings":{
            "properties":{
                "link":{
                    "type":"text",
                    "index":"true"
                },
                "title":{
                    "type":"text",
                    "index":"true"
                }
            }
```

```
        }
    }

    if es.indices.exists("nkindex") is not True:
        es.indices.create(index = 'nkindex', body = mappings)
    else:
        es.indices.delete(index = 'nkindex')
        es.indices.create(index = 'nkindex', body = mappings)
```

以上代码建立完索引后，往其中插入数据，遍历爬取数据的文件夹，同样使用正则表达式匹配之后，提取出每一行的链接和标题，存入一个列表之后，使用`bulk`函数将数据插入。

```
def loadnews():
    tempList = []
    for filepath, dirnames, filenames in os.walk('./南开新闻'):
        for filename in filenames:
            # 合成绝对路径
            ab_filepathos = os.path.join(filepath, filename)
            # 以r格式打开，即只读格式
            with open(ab_filepathos, 'r', encoding='utf-8') as f:
                # 将文件读取到lines中，按行读取
                lines = f.readlines()
            for line in lines:
                # 提取每一行中的链接和标题
                Re = re.compile('link:(.*?), title:')
                link = Re.findall(line)[0]
                title = line[line.index('title:') + 6:]
                body = {
                    'link':link,
                    'title':title,
                }
                # 使用json格式传输字典
                tempList.append(json.dumps(body))

    ACTIONS = [
        {
            '_index':'nkindex',
            '_type':'_doc',
            '_source': s
        }
        for s in tempList
    ]

    bulk(es, ACTIONS)
```

PageRank链接分析

可以发现在南开要闻网上，每一个新闻都会指向10个热点新闻：

南开新闻网讯（记者 吴军辉 通讯员 李伟）12月8日，南开大学服务庄浪县乡村振兴创新试验郑河乡生活垃圾就地处置示范站举行揭牌仪式。南开大学党委常委、副校长李靖，庄浪县委副书记、县长王敏共同为示范站揭牌。我校挂职干部、庄浪县副县长邵刚主持揭牌仪式。

该项目由南开大学环境科学与工程学院牵头捐赠实施，装置总价值100余万元。该工作站采用的生活垃圾就地处理装置运用高效热传导技术，采用先进的烟气净化工艺，大幅降低了收集、运输成本，实现了减量化、无害化、资源化处理，能够有效解决群众生活垃圾难于处理的问题，破解乡村治理中的生活垃圾处理难题，是解决农村生活垃圾的有效方式。

李靖表示，希望乡村两级把绿色发展理念与乡村振兴工作有效结合起来，在生活垃圾就地处理装置的运行过程中，帮助南开大学积累宝贵经验，多提改进意见，努力将该项目打造成为具有示范推广效应的创新帮扶项目，为农村人居环境整治、建设生态宜居美丽乡村探索经验。

王敏指出，南开大学为郑河乡卢洼村捐建生活垃圾就地处理装置，是贯彻落实习近平生态文明思想的具体实践，是南开大学帮扶庄浪县深入实施乡村振兴的实际行动，是庄浪县推进全域无垃圾农村人居环境整治的重要抓手。郑河乡要严格按照处置装置技术要求，规范垃圾分类，强化宣传引导，及时收集转运处理，真正把生活垃圾就地处理装置运行好，争取

我对这十个新闻提取出来，写入到一个文本中便于访问

```
link:http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049290.shtml, title:张伟平院士为青年学子讲述陈...
link:http://news.nankai.edu.cn/ywsd/system/2021/12/05/030049287.shtml, title:我校召开会议部署第十次党代...
link:http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049292.shtml, title:全国高校大学生讲思政课公开...
link:http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049289.shtml, title:南开大学举办2021年度研究生...
link:http://news.nankai.edu.cn/ywsd/system/2021/12/08/030049332.shtml, title:南开助你“创”青春
link:http://news.nankai.edu.cn/ywsd/system/2021/12/07/030049315.shtml, title:南开研究生支教团20周年纪念...
link:http://news.nankai.edu.cn/ywsd/system/2021/12/05/030049288.shtml, title:杨庆山会见校友企业家张文中一行
link:http://news.nankai.edu.cn/ywsd/system/2021/12/10/030049385.shtml, title:南开大学2021年民族团结进步...
link:http://news.nankai.edu.cn/ywsd/system/2021/12/11/030049412.shtml, title:杨庆山赴甘肃环县习仲勋红军...
link:http://news.nankai.edu.cn/ywsd/system/2021/12/07/030049302.shtml, title:践行南开特色研究性教学模式 ...
```

PageRank算法的含义是，为网页构建一个有向图，当一个网页指向另外一个网页，则该边加入有向图中。遍历爬取的资源，每一个网页都指向上述所说的十个网页，因此将指向这十个热点新闻的网页的边加入到有向图中：

```python
for filepath, dirnames, filenames in os.walk('./南开新闻'):
    for filename in filenames:
        ab_filepathos = os.path.join(filepath, filename)
        with open(ab_filepathos, 'r', encoding='utf-8') as f:
            # 将文件读取到lines中，按行读取
            lines = f.readlines()
        for line in lines:
            Re = re.compile('link:(.*?), title:')
            # 找到该行的链接
            link = Re.findall(line)[0]
            for i in range(10):
                G.add_edge(link, hit[i][0])
        f.close()
```

使用Python的networkx包中的Pagerank函数，即可方便地计算出网页的PageRank值。

```
pr = nx.pagerank(G)
for node, value in pr.items():
    print(node, value)
for i in range(10):
    print(hit[i][0], pr[hit[i][0]])
```

在终端中打印一部分信息如下：

根据上面的分析可知，只有这十个热点新闻的PageRank值较高，其余的值都一样，如下：

```
http://news.nankai.edu.cn/ywsd/system/2021/12/10/030049379.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/09/030049378.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/09/030049377.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/09/030049375.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/09/030049365.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/09/030049363.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/09/030049349.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/08/030049337.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/07/030049330.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/07/030049317.shtml 3.409158217542268e-05
http://news.nankai.edu.cn/ywsd/system/2021/12/05/030049287.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049290.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049292.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049289.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/08/030049332.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/07/030049315.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/05/030049288.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/10/030049385.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/11/030049412.shtml 0.04562051727201998
http://news.nankai.edu.cn/ywsd/system/2021/12/10/030049392.shtml 0.04562051727201998
```

最后十行便是热点最高的十个新闻。

查询服务

站内查询

站内查询是指定一个查询网址，在该网址中进行查询，查询体如下：

```
query_json = {
        'query':{
            'bool':{
                'must':[
                    {'match':{'title':key}},
                    {'match_phrase':{'link':web}}
                ]
            }
        },
        'size': 10
    }
```

其中 `match_phrase` 为模糊匹配，可以满足站内查询的条件，因为该网址内的资源，都会匹配包含它的网站。运行如下：

```
input the web you want to query in:http://news.nankai.edu.cn/ywsd/system/2021/12/06/

input the key word you want to query:南开大学
C:\Users\86183\Desktop\南开大学\信息检索\1911501_宣恩允_hw4\search.py:29: DeprecationWarning:
The 'body' parameter is deprecated for the 'search' API and will be removed in a future
version. Instead use API parameters directly. See https://github.com/elastic/elasticsearch-
py/issues/1698 for more information
  res = es.search(index='nkindex', body=query_json)
C:\ProgramData\Anaconda3\lib\site-packages\elasticsearch\connection\base.py:209:
ElasticsearchWarning: Elasticsearch built-in security features are not enabled. Without
authentication, your cluster could be accessible to anyone. See https://www.elastic.co/
guide/en/elasticsearch/reference/7.15/security-minimal-setup.html to enable security.
  warnings.warn(message, category=ElasticsearchWarning)
2021-12-14 23:10:44.875 GET http://localhost:9200/ [status:200 request:0.022s]
2021-12-14 23:10:44.890 POST http://localhost:9200/nkindex/_search [status:200 request:
0.013s]
results are as follow
Got 11 Hits:
{'link': 'http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049289.shtml', 'title': '南开
大学举办2021年度研究生...\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049289.shtml', 'title': '南开
大学举办2021年度研究生指导教师培训活动\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049289.shtml', 'title': '南开
大学举办2021年度研究生指导教师培训活动\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049301.shtml', 'title': '南开
大学新能源转化与存储交叉科学中心联合eScience编辑部召...\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049301.shtml', 'title': '南开
大学新能源转化与存储交叉科学中心联合eScience编辑部召...\n'}
```

以上查询目的是查询出来2021年12月6日发布的并且匹配关键字“南开大学”的新闻，可以看到结果符合预期。

文档查询

该查询需要查询出来elasticsearch中的一个文档，由于每个文档的都有如下相同的结构：

```
ACTIONS = {
        '_id': i
    }
```

可以根据'_id'的值来查询文档，用它作为查询体的一部分。

```
query_json = {
        'query':{
            'bool':{
                'must': [
                    {'match':ACTIONS}
                ]
            }
        },
        'size': 10
    }
```

若想查询到如下的文档：

运行如下：

```
input the id of document you want to query:WlAEuX0BszfXlfpMBfFx
C:\Users\86183\Desktop\南开大学\信息检索\1911501_宣恩允_hw4\search.py:61: DeprecationWarning:
The 'body' parameter is deprecated for the 'search' API and will be removed in a future
version. Instead use API parameters directly. See https://github.com/elastic/elasticsearch-
py/issues/1698 for more information
  res = es.search(index='nkindex', body=query_json)
2021-12-14 23:18:39.966 GET http://localhost:9200/ [status:200 request:0.005s]
2021-12-14 23:18:39.980 POST http://localhost:9200/nkindex/_search [status:200 request:
0.013s]
results are as follow
Got 1 Hits:
{'link': 'http://news.nankai.edu.cn/zhxw/system/2010/03/25/000029345.shtml', 'title': '美国
洛克菲勒集团总裁受聘南开大学客座教授\n'}
```

输入该id的值，即可查询到准确的文档。

短语查询

该部分在elasticsearch中已经实现，只需要编写查询体，使用`match_phrase`匹配关键词即可。

```
query_json = {
        'query':{
            'bool':{
                'must': [
                    {'match_phrase':{'title':key}}
                ]
            }
        },
        'size': 10
    }
```

运行如下：

```
input the phrase you want to query:张伟平
C:\Users\86183\Desktop\南开大学\信息检索\1911501_宣恩允_hw4\search.py:87: DeprecationWarning:
The 'body' parameter is deprecated for the 'search' API and will be removed in a future
version. Instead use API parameters directly. See https://github.com/elastic/elasticsearch-
py/issues/1698 for more information
  res = es.search(index='nkindex', body=query_json)
2021-12-14 23:27:52.620 GET http://localhost:9200/ [status:200 request:0.004s]
2021-12-14 23:27:52.639 POST http://localhost:9200/nkindex/_search [status:200 request:
0.017s]
results are as follow
Got 12 Hits:
{'link': 'http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049290.shtml', 'title': '张伟
平院士为青年学子讲述陈...\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2013/02/02/000110254.shtml', 'title': '校领
导春节前夕看望张伟平院士\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2015/02/06/000220855.shtml', 'title': '薛进
文春节前看望慰问张伟平院士\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2015/03/14/000225296.shtml', 'title': '薛进
文春节前看望慰问张伟平院士\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2021/01/18/030044248.shtml', 'title': '天津
市财政局领导慰问张伟平院士\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2014/03/30/000171874.shtml', 'title': '光明
日报：张伟平：心中藏着“数学强国梦”\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2016/12/05/000309174.shtml', 'title': '纪念
陈省身先生诞辰 张伟平院士做学术报告\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2017/02/20/000318616.shtml', 'title': '张伟
平院士在国际顶尖数学刊物《Annals of Mathematics》发表...\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2014/03/25/000171294.shtml', 'title': '张伟
平院士学术论文在国际顶级数学杂志《Acta Mathematica》发表\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2021/12/06/030049290.shtml', 'title': '张伟
平院士为青年学子讲述陈省身先生的学术成就\n'}
```

通配符查询

通配符类似于正则表达式，python中有两个主要的通配符

- `?`: which matches any single character
- `*`: which can match zero or more characters, including an empty one

`?` 匹配任意单个字符，`*` 匹配任意字符。

同时在查询体中，使用`wildcard`字段，可以使用通配符进行查询：

```python
query_json = {
        'query':{
            'wildcard':{
                'title':{
                    'value':key
                }
            }
        },
        'size': 10
    }
```

运行结果如下：

```
input the key words you want to query:*张*
C:\Users\86183\Desktop\南开大学\信息检索\1911501_宣恩允_hw4\search.py:114:
DeprecationWarning: The 'body' parameter is deprecated for the 'search' API and will be
removed in a future version. Instead use API parameters directly. See https://github.com/
elastic/elasticsearch-py/issues/1698 for more information
  res = es.search(index='nkindex', body=query_json)
C:\ProgramData\Anaconda3\lib\site-packages\elasticsearch\connection\base.py:209:
ElasticsearchWarning: Elasticsearch built-in security features are not enabled. Without
authentication, your cluster could be accessible to anyone. See https://www.elastic.co/
guide/en/elasticsearch/reference/7.15/security-minimal-setup.html to enable security.
  warnings.warn(message, category=ElasticsearchWarning)
2021-12-14 23:38:54.039 GET http://localhost:9200/ [status:200 request:0.006s]
2021-12-14 23:38:54.050 POST http://localhost:9200/nkindex/_search [status:200 request:
0.010s]
results are as follow
Got 171 Hits:
{'link': 'http://news.nankai.edu.cn/ywsd/system/2010/07/01/000032291.shtml', 'title': '解放
军总医院副政委张保全一行访问南开\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2010/10/18/000034226.shtml', 'title': '南开
大学九旬院士讲述张伯苓教育风采\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2010/10/17/000034189.shtml', 'title': '南开
大学周末影坛校庆特别奉献电视剧《张伯苓》精编版 \n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2010/11/13/000035276.shtml', 'title': '张力
解读《国家中长期教育改革和发展规划纲要》\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2010/12/03/000035984.shtml', 'title': '张再
旺同志逝世\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2010/12/01/000035939.shtml', 'title': '南开
大学颁发2010年张舜尧暨香港友人奖助学金\n'}
{'link': 'http://news.nankai.edu.cn/ywsd/system/2010/12/07/000036099.shtml', 'title': '【缅
怀】深切怀念张再旺同志\n'}
```

查询日志

在上述查询函数中将查询条件、查询时间、查询类型写入文档中(以短语查询为例)

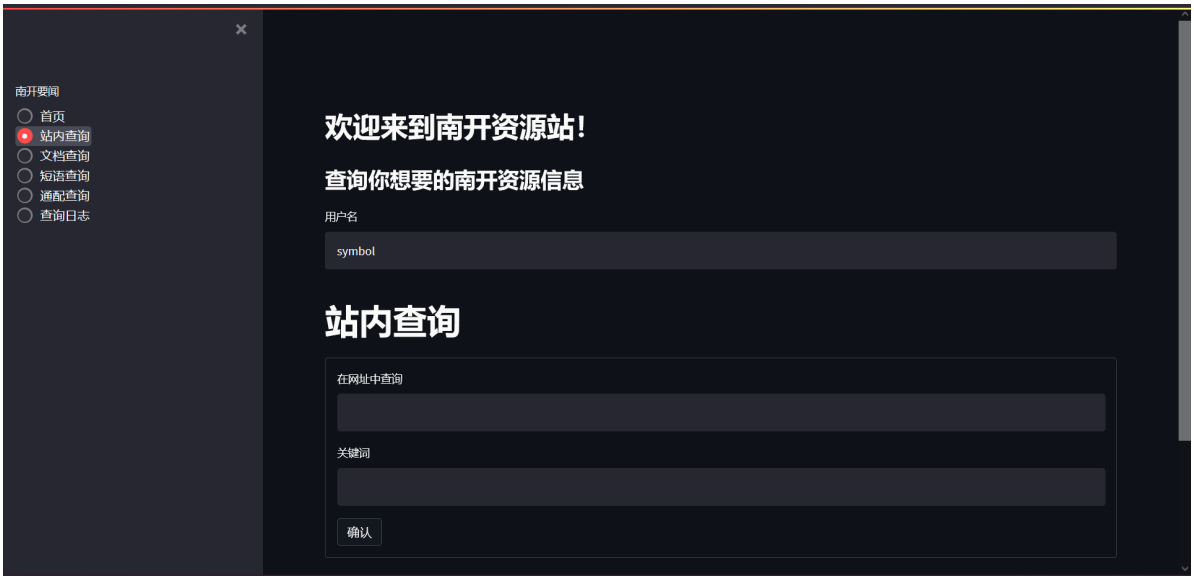
```python
with open('./查询日志.txt', mode='a', encoding='utf-8') as f:
        time_str = datetime.datetime.now().strftime('%Y-%m-%d %H:%M:%S')
        f.write('phraseSearch' + '\n')
        f.write(time_str + '\n')
        f.write('key:' + key + '\n')
        f.close()
```

```
phraseSearch
2021-12-14 09:47:02
key:张伟平
inWebSearch
2021-12-14 23:07:51
web:http://news.nankai.edu.cn/ywsd/index.shtml
key:中国社会科学
inWebSearch
2021-12-14 23:08:27
web:http://news.nankai.edu.cn/ywsd/system/count//0003000/000
key:建设美丽乡村
inWebSearch
2021-12-14 23:08:57
web:http://news.nankai.edu.cn/ywsd/system/
key:建设美丽乡村
inWebSearch
2021-12-14 23:09:55
web:http://news.nankai.edu.cn/ywsd/system/2021/12/06/
key:南开大啊学
inWebSearch
2021-12-14 23:10:44
web:http://news.nankai.edu.cn/ywsd/system/2021/12/06/
key:南开大学
docSearch
2021-12-14 23:18:39
docid:WlAEuX0BszfXlfpMBfFx
phraseSearch
2021-12-14 23:27:52
key:张伟平
```

个性化查询&Web界面

实验中将这两个部分合成了一个部分，使用了python中的streamlit包来搭建web界面。

搭建的界面如下图：



在这部分中，定义了一个选择框：

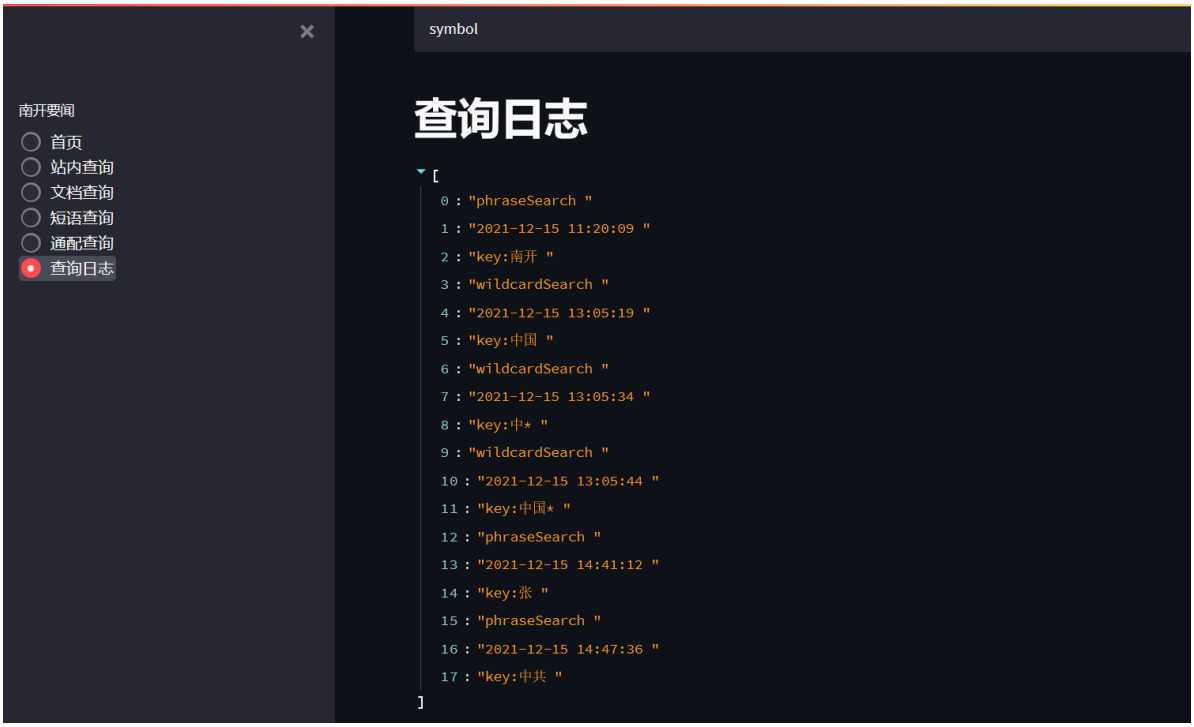
```
# 设置左侧导航栏
sidebar = st.sidebar.radio(
    "南开要闻",
    ("首页", "站内查询", "文档查询", "短语查询", "通配查询", "查询日志")
)

```

当点击不同的选择框出现不同的查询结果，以站内查询为例子：

```python
if sidebar == '站内查询':
    st.title('站内查询')
    # 创建一个表单
    with st.form('站内查询'):
        webtext = st.text_input('在网址中查询')
        keytext = st.text_input('关键词')
        confirm = st.form_submit_button('确认')
        # 点击按钮
    if confirm:
        res = search.inWebSearch(name, webtext, keytext)
        st.success('查询到了' + str(res['hits']['total']['value']) + '条结果')
        for hit in res['hits']['hits']:
            st.write(hit['_source'])
```

在子界面中，调用了站内查询函数，然后进行相应的输出即可：

同时点击查询日志，可以看到查询的记录：


南开要闻
首页
站内查询
文档查询
短语查询
通配查询
查询日志
symbol
查询日志
[
0 : "phraseSearch "
1 : "2021-12-15 11:20:09 "
2 : "key:南开 "
3 : "wildcardSearch "
4 : "2021-12-15 13:05:19 "
5 : "key:中国 "
6 : "wildcardSearch "
7 : "2021-12-15 13:05:34 "
8 : "key:中* "
9 : "wildcardSearch "
10 : "2021-12-15 13:05:44 "
11 : "key:中国* "
12 : "phraseSearch "
13 : "2021-12-15 14:41:12 "
14 : "key:张 "
15 : "phraseSearch "
16 : "2021-12-15 14:47:36 "
17 : "key:中共 "
]

以此来个性化查询。