

NKU 风雨声

一、项目目录

```
WebSearchEngine/
├── data/          # 存放数据(包含爬虫脚本、解析脚本等)
│   ├── document_data # 文件数据
│   └── web_data      # 网页数据
│
├── src/
│   ├── create_indices.py # 建立索引脚本
│   ├── import_data.py    # 导入数据脚本
│   └── search.py         # 搜索脚本
│
├── webapp/
│   ├── static # 存放网页快照数据
│   ├── templates # 存放网页模板
│   ├── app.py  # 网站主程序
│   ├── elastic_utils.py # 与ElasticSearch交互的脚本
│   ├── search_logic.py  # 搜索功能的脚本
│   └── recommend_logic.py # 推荐功能的脚本
```

项目目录如上图所示，在提交的代码中也会留存项目的框架。

二、网页爬取

通过脚本，对南开新闻网的数据和学院的首页通知和推送进行爬取。注意礼貌爬取，不要给服务器造成太大压力。

1.spider.py

这类脚本用于爬取南开大学多个网站下的内容。主要使用了 requests 和 BeautifulSoup 库。获取网页内容，解析网页链接，并且使用 MD5 哈希 URL 生

成唯一文件名。文件数量展示如下：

```
126991 {"id": "f75accebead9d66f6041aca87e0f128b", "url": "https://zfxi.nankai.edu.cn/faculty/function
126992 {"id": "f4072a6f3babf9655adc484a3f1cb60", "url": "https://zfxi.nankai.edu.cn/system/_content/
126993 {"id": "f72dcf13b207165b2f86eeaa988e460a", "url": "https://zfxi.nankai.edu.cn/info/1220/8849.h
126994 {"id": "f9a0a04f27aed260dc6e2fefc5532e53", "url": "https://zfxi.nankai.edu.cn/faculty/public_a
126995 {"id": "f8dc6b46c8f9f7535031e097801d96ef", "url": "https://zfxi.nankai.edu.cn/info/1829/8638.h
126996 {"id": "f9c5af60a29a44c7817c18f760f5b7a8", "url": "https://zfxi.nankai.edu.cn/system/_content/
126997 {"id": "f7f38cab820923df12afdc4e58e2dd9f", "url": "https://zfxi.nankai.edu.cn/info/1156/5433.h
126998 {"id": "fc21658df8d06cd069e583419c9d922e", "url": "https://zfxi.nankai.edu.cn/info/1249/6787.h
126999 {"id": "fb0567a0e0bf70e5934d432985d8e384", "url": "https://zfxi.nankai.edu.cn/info/1139/1062.h
127000 {"id": "fbcbb29128de39a5fcf91004eb5cd9fc", "url": "https://zfxi.nankai.edu.cn/system/_content/
127001 {"id": "fcc122a05ec8f302ae8f4b4aaab631ab", "url": "https://zfxi.nankai.edu.cn/faculty/internat
127002 {"id": "fe8aacadf65037107548aa8a4e381b3e", "url": "https://zfxi.nankai.edu.cn/system/_content/
127003 {"id": "fc9c5dee2deec9b03e384d1272d01882", "url": "https://zfxi.nankai.edu.cn/faculty/politica
127004
```

网页的数量是 127003 条。

```
12043 {"id": "document_college_data/zfxi_data\\5b2a5d91a032c09a30d8a6be4147a9e8.doc", "url": "https:/
12044 {"id": "document_college_data/zfxi_data\\49543e541c1f4ee7b2686ee365786a7f.xls", "url": "https:/
12045 {"id": "document_college_data/zfxi_data\\12c57688fd39e77a7b5ae1645113a19a.xlsx", "url": "https:/
12046 {"id": "document_college_data/zfxi_data\\62545f736b047dd7defb8e581d405940.doc", "url": "https:/
12047 {"id": "document_college_data/zfxi_data\\c8604917f912a03b4c0df66898173d94.doc", "url": "https:/
12048 {"id": "document_college_data/zfxi_data\\7e1d611b27ad19cc5262e6a3b08589ba.pdf", "url": "https:/
12049
```

文件的数量是 12048 条。

条目的总数达到了 139051，接近十四万条，满足十万+的数目要求。

2.parser.py

这类脚本用于解析 HTML 文档并将其转换为结构化 JSONL 格式。解析文件内的

标题、内容、锚文本等等部分，为后面建立索引打下了基础。主要

使用 `extract_text_from_html(file_path)` 函数对内容进行解析；

使用 `process_file(filename, data_dir, url_map)` 函数对内容进行处理，返回

id,url,title,content……信息。

使用 `parse_all_html(data_dir, url_map_file, output_jsonl)` 函数对所有文件进行

处理。

使用 `import_data.py` 脚本将网页数据导入到索引中。在 `webapp` 文件夹下编写 `search_logic.py`，实现多种查询服务。

2. 查询服务

站内查询主要是通过 `filter` 进行实现，过滤掉其他的网页。

NKU风雨声

登录

☐ 短语查询

☐ 通配查询

搜索

搜索结果：10 条

“校长杯”排球赛落幕，商学院排球队发挥出色再创佳绩 网址: https://bs.nankai.edu.cn/2018/0628/c9007a103143/page.htm 查看快照(2025.5.21) 类型: webpage
关于举办2023年南开大学教职工气排球比赛的通知 网址: https://bs.nankai.edu.cn/2024/0516/c35466a542882/page.htm 查看快照(2025.5.21) 类型: webpage
商学院分会气排球嘉年华活动成功举办 网址: https://bs.nankai.edu.cn/2024/0516/c35467a542877/page.htm 查看快照(2025.5.21) 类型: webpage
欢迎各位老师加入乒乓球、气排球和羽毛球俱乐部 网址: https://bs.nankai.edu.cn/2024/0516/c35466a542883/page.htm 查看快照(2025.5.21) 类型: webpage

文档查询通过对搜索结果限定是网页还是文件，进行特定的输出。

```
# 爬取支持的文档类型
ALLOWED_FILE_EXTENSIONS = {
    '.pdf', '.doc', '.docx', '.xls', '.xlsx', '.ppt', '.pptx'
}
```

NKU风雨声

登录

☐ 短语查询

☐ 通配查询

搜索

搜索结果：10 条

理论学习摘编 网址: https://bs.nankai.edu.cn/_upload/article/files/a7/6c/19649a604466bfe2c9bd1d81b91b/2313297c-5c74-417d-bbf4-3a7ee08b1c3e.pdf 类型: document
校外学习交流课程认定程序 网址: https://medical.nankai.edu.cn/_upload/article/files/c3/d4/403c565f4f45a19022d7ead564e7/0718df8d-2c96-462b-8af0-05c31f92c808.docx 类型: document
爱知大学大学院中国研究科博士课程双重学位学习简介I.课程学习 网址: https://chem.nankai.edu.cn/_upload/article/files/d7/20/af330ccb46949822e5186d8ee78c/97840284-d7ba-4376-94c5-b862b368eaba.pdf 类型: document
鄂维南，北京大学国际机器学习研究中心教授，数学学院教授，中国科学院院士，美国数学学会、美国工业与应用数学学会Fellow，北京大数据研究院院长，北京科学智能研究院(AI for Science) In 网址: https://chem.nankai.edu.cn/_upload/article/files/22/18/82cd5df74cd48c6ffc649efd30ae/599becae-3d9b-4724-9b70-4c90b56bce08.docx 类型: document

短语查询使用使用 `match_phrase` 查询，要求词语按顺序出现，并且要求词语作为短语不可分割。

大学 排球

站内搜索

网页

☒ 短语查询

☐ 通配查询

搜索

搜索结果：10 条

人民网：第九届全国大运会 香港女排与南开大学激扬排球俱乐部进行交流-媒体南开-南开大学 网址: http://news.nankai.edu.cn/mtnk/system/2012/09/15/000087096.shtml  查看快照(2025.5.21) 类型: webpage
南开女排挺进中国大学生排球联赛全国总决赛-南开要闻-南开大学 网址: http://news.nankai.edu.cn/ywsd/system/2019/12/06/030036718.shtml  查看快照(2025.5.21) 类型: webpage
【关注大运会】香港女排与南开大学激扬排球俱乐部交流-南开要闻-南开大学 网址: http://news.nankai.edu.cn/ywsd/system/2012/09/14/000086654.shtml  查看快照(2025.5.21) 类型: webpage
喜讯：学院男排获得南开大学排球赛冠军 网址: https://mse.nankai.edu.cn/2016/0521/c9309a95946/page.htm  查看快照(2025.5.21) 类型: webpage

学生文化

站内搜索

网页

☒ 短语查询

☐ 通配查询

搜索

搜索结果：10 条

今晚报：大学生文化创意创新创业大赛颁奖-媒体南开-南开大学 网址: http://news.nankai.edu.cn/mtnk/system/2021/06/28/030047062.shtml  查看快照(2025.5.21) 类型: webpage
人民网：“津港大学生文化夏令营”在南开大学举行-媒体南开-南开大学 网址: http://news.nankai.edu.cn/mtnk/system/2009/07/26/000024819.shtml  查看快照(2025.5.21) 类型: webpage
北方网：首届天津市大学生文化创意作品竞赛南开获佳绩-媒体南开-南开大学 网址: http://news.nankai.edu.cn/mtnk/system/2008/11/15/000020099.shtml  查看快照(2025.5.21) 类型: webpage
天津日报：大学生文化创意创新创业大赛落幕-媒体南开-南开大学 网址: http://news.nankai.edu.cn/mtnk/system/2021/06/28/030047071.shtml  查看快照(2025.5.21) 类型: webpage

通配查询需要下载 ik 插件进行分词，还需要添加 title.wildcard 字段(专门为通配查询优化的字段)。

温*

站内搜索





网页

☐ 短语查询

☒ 通配查询

搜索

搜索结果：10 条

温馨提示 网址: https://bs.nankai.edu.cn/2019/0102/c10228a117719/page.htm  查看快照(2025.5.21) 类型: webpage
温延龙 网址: https://cc.nankai.edu.cn/2021/0323/c13619a551346/page.htm  查看快照(2025.5.21) 类型: webpage
温馨提示 网址: https://chem.nankai.edu.cn/2014/1201/c24120a369171/page.htm  查看快照(2025.5.21) 类型: webpage
温志慧 网址: https://chem.nankai.edu.cn/2019/0906/c24405a379817/page.htm  查看快照(2025.5.21) 类型: webpage

站内搜索

网页

☐ 短语查询

☒ 通配查询

搜索

搜索结果: 2 条

- 中国日报网: 发挥排球育人功能 践行南开体育精神-媒体南开-南开大学

网址:<http://news.nankai.edu.cn/mtnk/system/2021/03/31/030045185.shtml>

查看快照(2025.5.21) 类型: webpage
- 发挥排球育人功能 践行南开体育精神-南开要闻-南开大学

网址:<http://news.nankai.edu.cn/ywzd/system/2021/03/30/030045155.shtml>

查看快照(2025.5.21) 类型: webpage

搜索日志的实现在实现用户的注册和登录基础上，为每个用户记录搜索历史，并进而进行个性化查询和个性化推荐。

站内搜索

网页

☐ 短语查询

☐ 通配查询

搜索

机器学习

属性

结构

蛋白质

植物

[5255/page.htm](#)

结构性质

[i033/page.htm](#)

网页快照的实现基于之前的网页抓取。在爬取网页内容后，网页被存储在本地永久保存，用户可以选择从网页进入查询结果，也可以通过快照(特定的时间的本地文件)进入。

news.nankai.edu.cn/ywzd/system/2021/03/30/030045155.shtml

签 网址导航 JD 京东 论文查重 爱淘宝 万能小in AI助手 热门游戏 百度 创客贴 https://eamis.nan... NKU-OJ AlchatOS 注销页 基本资料|南开大

南开大学 1919

新闻网 Nankai University

微信 微博 抖音 快手 哔哩哔哩 知乎 学习 头条

首页 南开要闻 媒体南开 南开校史 光影南开 南开故事 南开大学报 视频 广播

您当前的位置：南开大学 >> 南开要闻

发挥排球育人功能 践行南开体育精神

来源：南开大学新闻网 发稿时间：2021-03-30 14:28

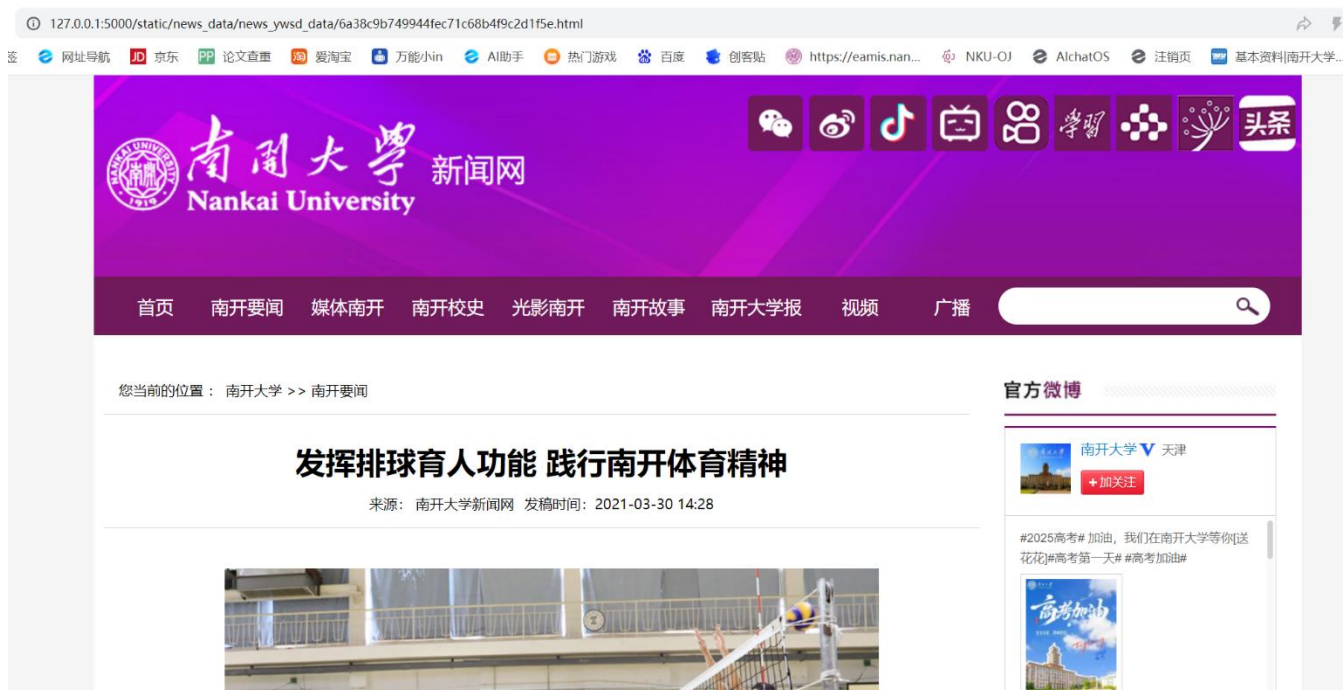


官方微博

南开大学 天津

#2025高考# 加油，我们在南开大学等你送花花|#高考第一天# #高考加油#





四、个性化查询/推荐

记录用户的搜索历史，根据不同用户的偏好，对查询结果进行“再排序”，并且主动推荐一些用户可能喜欢的内容。个性化查询根据用户的常用查询词，若相关文章的标题、内容出现了常用查询词，则根据出现的位置和次数，进行不同的加分策略。个性化推荐是根据常用查询词进行查询，然后将综合结果返回给用户。

NKU风雨声

欢迎, test

机器学习

站内查询

网页

☐ 短语查询

☐ 通配查询

搜索结果: 10 条

基于机器学习的蛋白质结构与属性预测

网址: <https://math.nankai.edu.cn/2021/1110/c34410a525255/page.htm>

类型: webpage

【“开悟”物理前沿讲座（第110期）】机器学习预测原子核结构性质

网址: <https://physics.nankai.edu.cn/2024/1111/c573a556033/page.htm>

类型: webpage

3290数据挖掘和机器学习停课通知

网址: <https://stat.nankai.edu.cn/2023/1218/c12312a531902/page.htm>

类型: webpage

若用户多次查询“蛋白质”、“属性”相关的词，那么搜索“机器学习”就会将上图的第一个结果返回给用户。

切换账号，拥有以下的搜索历史，就有相应的相关推荐。

NKU风雨声

欢迎, gyx 搜索历史 退出

输入关键词

站内查询

网页

☐ 短语查询

☐ 通配查询

搜索

运动	
南开	
篮协	092/page.htm
体育	
篮球	4/page.htm

为你推荐

<div>南开大学与中国篮协签约合作《中国篮球通志》编纂全面启动</div> <div>https://history.nankai.edu.cn/2023/0410/c16078a508786/page.htm</div> <div> 查看快照(2025.5.21) 类型: webpage</div>
<div>南开大学与中国篮协签约合作《中国篮球通志》编纂全面启动-南开要闻-南开大学</div> <div>http://news.nankai.edu.cn/ywsd/system/2023/04/10/030055324.shtml</div> <div> 查看快照(2025.5.21) 类型: webpage</div>
<div>天津日报：重现天津篮球运动历史荣光-媒体南开-南开大学</div> <div>http://news.nankai.edu.cn/mtnk/system/2023/08/04/030057342.shtml</div> <div> 查看快照(2025.5.21) 类型: webpage</div>
<div>北方网：姚明现身南开大学 中国篮球博物馆将落地天津-媒体南开-南开大学</div> <div>http://news.nankai.edu.cn/mtnk/system/2019/07/27/030034648.shtml</div> <div> 查看快照(2025.5.21) 类型: webpage</div>
<div>赵晶教授做客“中国篮球史志与文化系列讲座”第二讲</div> <div>https://history.nankai.edu.cn/2023/0605/c16078a514395/page.htm</div> <div> 查看快照(2025.5.21) 类型: webpage</div>

五、总结

以上就是本搜索引擎 NKU 风雨声的说明。在满足 10w+数据的前提下，实现了六种查询服务，并且设计了直观的 WEB 界面，并且实现了个性化查询和个性化推荐的功能。