USTC 分布式校园文档搜索引擎 - 实验报告

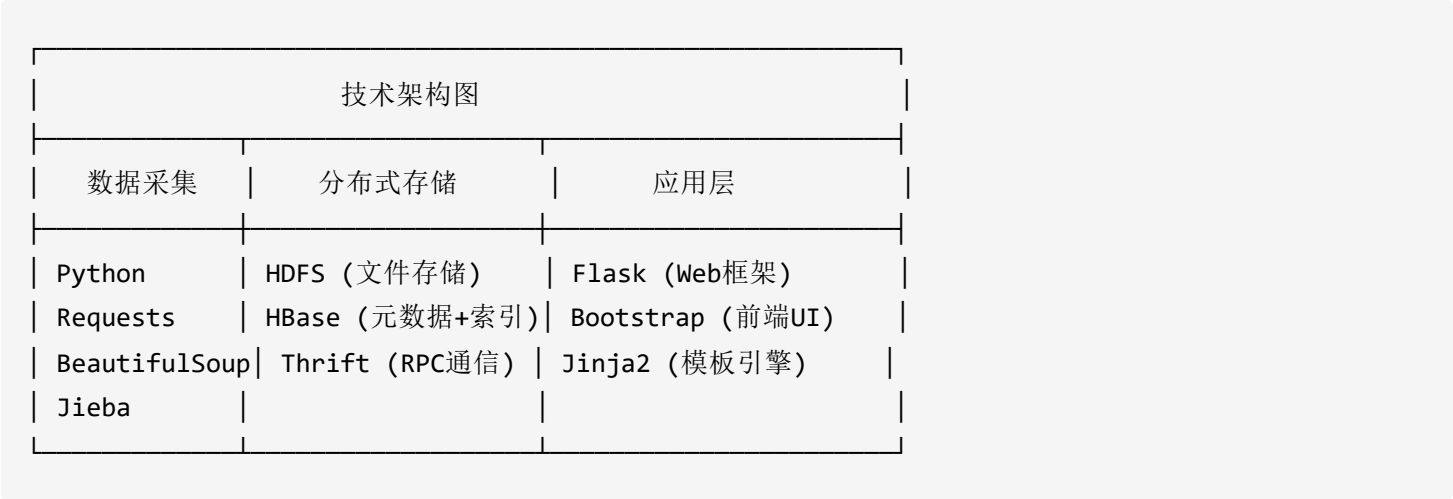
课程名称： 大数据系统综合实验
项目名称： USTC分布式校园文档搜索引擎
完成日期： 2026年1月

一、小组成员与分工

姓名	学号	分工内容
图力格尔	PB22051044	独立完成全部工作：需求分析、系统设计、爬虫开发、 存储架构搭建、Web应用开发、测试与调试

二、技术路线

2.1 项目技术栈概览



2.2 核心技术介绍

2.2.1 网络爬虫技术 (与课程相关)

Requests库：Python HTTP客户端库，用于发送网络请求

- 支持Session会话保持，复用TCP连接提高效率
- 配置自动重试策略(Retry)，应对网络波动和服务器5xx错误
- 流式下载(stream=True)，防止大文件撑爆内存

BeautifulSoup库：HTML/XML解析库

- 使用`html.parser`解析器提取页面中的超链接
- 实现BFS广度优先搜索策略进行链接发现

反爬对抗策略：

- 随机User-Agent轮换(模拟多浏览器访问)
- 随机延时3-6秒(模拟人类浏览行为)
- 异常退避机制(检测到封禁后休眠60秒)

2.2.2 分布式存储技术

HDFS (Hadoop Distributed File System)：

- 用于存储爬取的原始文件(PDF/DOC/DOCX等)
- 基于MD5哈希命名实现文件去重
- 通过WSL调用HDFS命令行完成跨系统文件传输

HBase (分布式列族数据库)：

- 存储文档元数据(URL、标题、日期)
- 存储文档内容摘要(用于全文检索)
- 存储关键词索引(Jieba提取的TF-IDF关键词)
- 使用happybase库通过Thrift协议连接

表结构设计：

```
表名: ustc_search_engine
行键: URL的MD5哈希值(32位定长,高效索引)
列族:
  - meta: 元数据 (url, title, type, date)
  - data: 内容数据 (hdfs_path, content)
  - index: 检索索引 (keywords)
```

2.2.3 中文自然语言处理

Jieba分词库：

- 使用TF-IDF算法提取文档关键词(Top 5)
- 支持中文分词，适配校园文档场景
- 关键词用于构建离线索引，加速检索

文档解析：

- pdfplumber：解析PDF文档提取文本
- python-docx：解析Word文档提取段落

2.2.4 Web应用技术

Flask框架：轻量级Python Web框架

- 路由映射：首页、搜索、下载三个核心接口
- 模板渲染：使用Jinja2引擎生成动态页面

搜索算法：三维加权相关度计分

```
Score = 标题命中×100 + 关键词命中×50 + 正文词频×1(上限20)
```

三、实现功能介绍与效果展示

3.1 功能一：分布式网络爬虫

功能描述：
自动爬取USTC校内各部门网站的文档资源，支持PDF、DOC、DOCX、XLS、XLSX等格式。

技术特点：

- BFS广度优先搜索策略
- 域名过滤(仅爬取ustc.edu.cn)
- URL去重(MD5哈希)
- 文件优先队列(文档链接优先处理)
- 流式下载(15MB上限)

爬取来源：

- 教务处 (teach.ustc.edu.cn)
- 研究生院 (gradschool.ustc.edu.cn)

- 财务处 (finance.ustc.edu.cn)
- 学工处 (stuhome.ustc.edu.cn)
- 计算机学院、信息学院等院系

效果展示：

```
[运行日志示例]
14:23:15 [INFO] Crawler started. Seeds: 26
14:23:18 [INFO] Status: 0 scanned | 0 files saved. Current: https://www.teach.ustc.edu.cn/downlo
14:23:25 [INFO] [SAVED] 本科生学籍管理规定.pdf
14:23:32 [INFO] [SAVED] 2024-2025学年校历.docx
14:24:01 [INFO] Status: 10 scanned | 5 files saved. Current: https://gradschool.ustc.edu.cn/...
...
16:45:33 [INFO] Task finished. Total files: 1247
```

3.2 功能二：分布式存储系统

功能描述：
将爬取的文件存储到HDFS，元数据和索引存储到HBase。

存储架构：

```
HDFS: /search_engine/raw_data/
  ├── a1b2c3d4e5f6...abc.pdf    (文件以MD5命名)
  ├── 1234abcd5678...xyz.docx
  └── ...

HBase: ustc_search_engine
  ┌──────────────────┬──────────────────────────────────────┐
  │ Row Key          │ Column Families                      │
  ├──────────────────┼──────────────────────────────────────┤
  │ md5(url)         │ meta:url, meta:title, meta:date      │
  │                  │ data:hdfs_path, data:content         │
  │                  │ index:keywords                       │
  └──────────────────┴──────────────────────────────────────┘
```

3.3 功能三：Web搜索界面

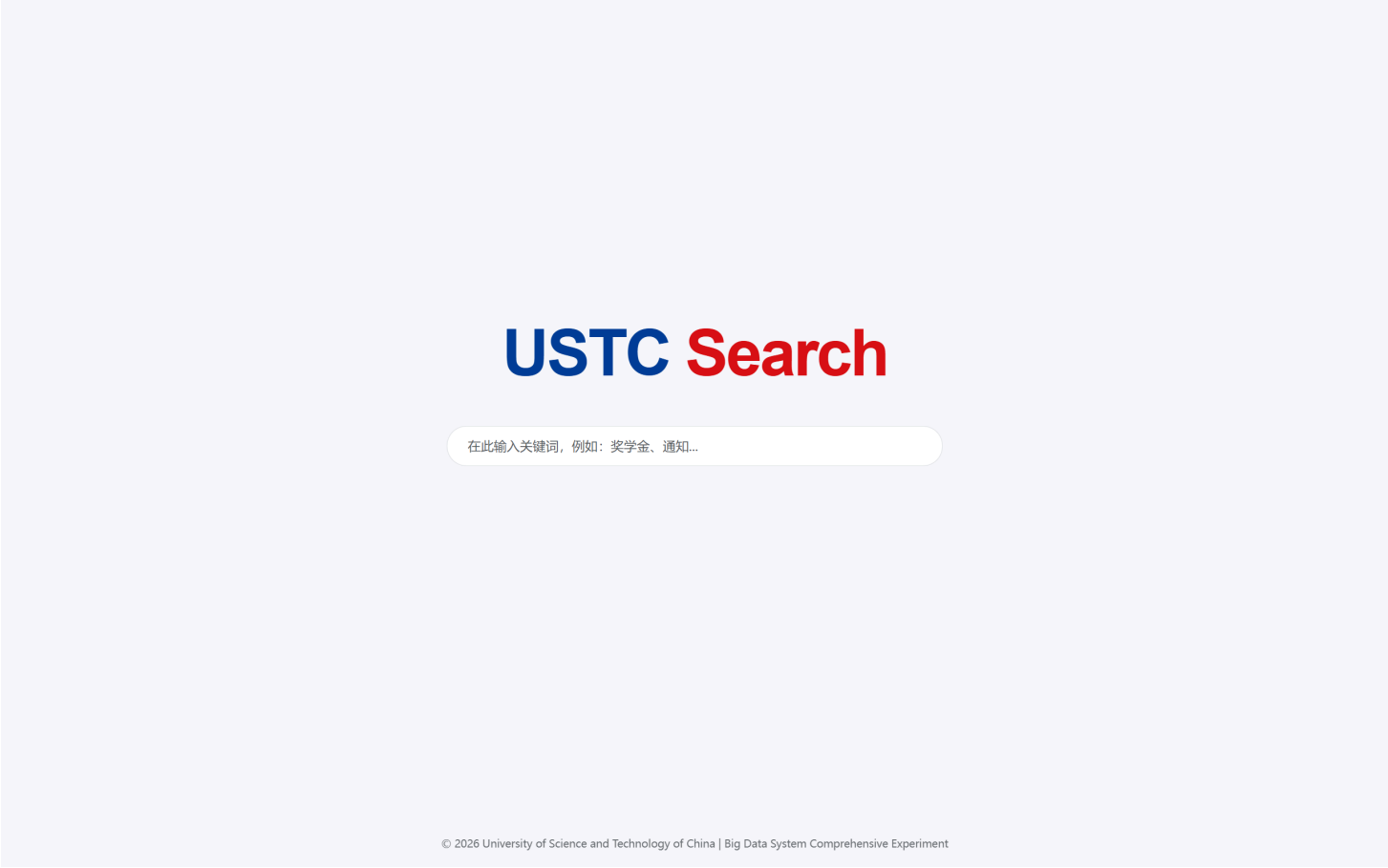
功能描述：
提供类似Google风格的搜索界面，支持关键词检索和文件下载。

主要功能：

1. **关键词搜索**：输入关键词返回相关文档列表
2. **相关度排序**：按三维计分模型排序结果
3. **摘要高亮**：关键词在摘要中红色高亮显示
4. **分页展示**：每页10条结果，支持翻页
5. **文件下载**：点击可直接从HDFS下载原文件

界面效果：

首页界面：



搜索结果页：

USTC Search 计算机 搜索

找到约 127 条结果 （用时 0.94 秒）

https://stuhome.ustc.edu.cn/_upload/article/files/2023/11... · 2025-12-24

我院刘利刚教授获2024年中国计算机图形学杰出奖.txt

10 月 11 日上午，在安徽黄山召开的第十五届中国**计算机**图形学大会（ Chinagraph 2024 ）开幕式上，两年一度的"中国**计算机**图形学贡献奖"和"中国**计算机**图形学杰出奖"评选结果揭晓。经广泛征集提名、奖励委员会推荐、 Chinagraph 2...

图形学 计算机 中国 学会 Chinagraph

https://www.teach.ustc.edu.cn/download/files/2024/我院... · 2025-12-24

我院成功举办第三届计算机图形学前沿暑期课程.txt

21日至7月25日，我院图形与几何计算实验室成功举办了第三届**计算机**图形学前沿暑期课程。该课程由我院刘利刚教授牵头组织，以介绍**计算机**图形学领域的最新的研究成果及进展为主要目的，同时为了兼顾本科生也介绍该领域的一些基本问题和研究方向，多视角、多层面地深入到各个热点...

授课 课程 图形学 计算机 暑期

https://finance.ustc.edu.cn/_upload/article/files/2023/计算... · 2025-12-24

计算机学院期末考试考场记录表.txt

Title: **计算机**学院期末考试考场记录表 Date: 2023-07-03 期末考试记录表(1).doc 仅供**计算机**学院开设的研究生课程使用...

记录表 期末考试 计算机 学院 Title

https://www.teach.ustc.edu.cn/download/files/2024/软件... · 2025-12-24

软件学院受邀在第六届中国计算机教育大会专题论坛分享建设成效.txt

12月7日至8日，第六届中国**计算机**教育大会在福建厦门举行。我校软件学院院长陈华平教授在大会特色化软件人才培养专题论坛作题为《嵌入式智能计算系统的研究、应用及人才培养》的主题报告，分享近几年来中国科学技术大学在特色化示范性软件学院...

软件 人才培养 特色化 计算机教育 嵌入式

https://www.teach.ustc.edu.cn/download/files/2024/信息... · 2025-12-24

信息与计算机实验教学中心召开教学指导委员会第四次会议.txt

2021年1月7日下午，信息与**计算机**实验教学中心教学指 导委员会第四次会议在线上召开。委员会专家清华大学杨士强教授、北京大学郝永胜教授、北京交通大学陈后金教授、 中国科学技术大学校长助理吴枫教授、 中国科学技术大学许胤龙教授出席会...

实验教学 实验 中心 教授 建设

https://gradschool.ustc.edu.cn/static/upload/article/file/2... · 2025-12-24

我院成功举办第六届计算机图形学前沿暑期课程.txt

国科学技术大学数学科学学院图形与几何计算实验室举办了第六届《**计算机**图形学前沿》暑期课程。本年度课程的主题为" 3D 几何感知与建模、虚拟现实、机器人与机器学习"，由刘利刚老师及多位国内外学者共同授课，介绍当前**计算机**图形学当今时代的研究热点和最新成果，同时兼顾本...

USTC Search 奖学金 搜索

找到约 48 条结果 （用时 0.141 秒）

https://stuhome.ustc.edu.cn/_upload/article/files/2024/02... · 2025-12-24

研究生学业奖学金实施细则.pdf

研字 (20 11号 19） 中国科学技术大学研究生学业**奖学金** 实施细则 根据财政部、教育部、人力资源社会保障部、退役军人 li ` ¢ 部、中央军委国防动员部印发的 《研究生学业**奖学金**实施细 则》,结 合我校实际,特制定 《中国科学技术大学研究生学 业奖...

研究生 学业 奖学金 参评 评定

https://www.teach.ustc.edu.cn/download/files/2024/关于2... · 2025-12-24

关于2024年部分专项奖学金和优秀学生奖学金候选人的公示.txt

Title: 关于2024年部分专项**奖学金**和优秀学生**奖学金**候选人的公示 Publisher: 发布者：少年班学院 Publish Date: 发布时间：2024/11/12 根据《关于评选2024年优秀学生**奖学金**的通知》（学字〔202...

奖学金 2024 公示 优秀学生 候选人

https://finance.ustc.edu.cn/_upload/article/files/2023/202... · 2025-12-24

2022年度求是奖学金评选通知.txt

进一步加强校企之间的联系， 求是科技基金会在我校捐赠设立求是**奖学金**， 用以 激励学子不断进取、追求卓越，使其为未来报效祖国、服务社会做好充足的知识与能力储备。 即日启动 2022 年度求是**奖学金**评选工作，现将有关事项通知如下： 一、 申报条件 1. 具有中国科...

奖学金 求是 电子版 2022 附件

https://cs.ustc.edu.cn/_upload/tpl/00/00/1/template1/styl... · 2025-12-24

关于2024年部分专项奖学金和奋进奖学金候选人的公示.txt

Title: 关于2024年部分专项**奖学金**和奋进**奖学金**候选人的公示 Publisher: 发布者：少年班学院 Publish Date: 发布时间：2024/10/09 根据《关于评选2024年奋进**奖学金**的通知》（学字〔2024〕48...

奖学金 2024 奋进 公示 蔷薇

https://www.teach.ustc.edu.cn/content/wp-content/uploa... · 2025-12-24

2018年度"华罗庚奖学金"颁奖典礼在我院举行.txt

2019年1月17日，2018年度"华罗庚**奖学金**"颁奖典礼在我校东区管理科研楼二楼报告厅举行。中国科学院大学教授、华罗庚先生之子华光、华光教授夫人江泳、中国科学院数学与系统科学研究院副院长巩馥洲、研究生部主任邵欣、副主任刘霞等嘉宾出席了颁奖礼...

华班 同学 华罗庚 奖学金 院长

https://www.ustc.edu.cn/files/download/2023/关于2024年... · 2025-12-24

关于2024年部分专项奖学金候选人的公示.txt

Title: 关于2024年部分专项**奖学金**候选人的公示 Publisher: 发布者：少年班学院 Publish Date: 发布时间：2024/11/08 根据《关

四、核心代码块

4.1 BFS爬虫主循环

```
def run(self):
    """BFS主循环: 队列处理直到达到MAX_PAGES或队列为空"""
    logger.info(f"Crawler started. Seeds: {len(self.queue)}")
    count = 0

    while self.queue and count < Config.MAX_PAGES:
        url = self.queue.popleft()
        # 用MD5哈希作为去重键(固定长度、高效)
        url_hash = hashlib.md5(url.encode()).hexdigest()

        if url_hash in self.visited: continue

        try:
            # 流式下载(避免大文件内存溢出)
            resp = self.session.get(
                url,
                headers=self._get_headers(),
                timeout=30,
                stream=True
            )

            if resp.status_code == 200:
                self._process_response(resp, url, url_hash)
        except Exception as e:
            # 连接被中断时长时间休息(避免IP封禁)
            if "10053" in str(e) or "Connection aborted" in str(e):
                logger.warning("Connection aborted. Sleeping 60s...")
                time.sleep(60)

        self.visited.add(url_hash)
        count += 1
        # 随机延时防封(模拟人类浏览行为)
        time.sleep(random.uniform(*Config.DELAY_RANGE))
```

4.2 文件存储到HDFS

```
def save_file_to_hdfs(self, content, ext):
    """保存文件到HDFS，使用MD5命名实现去重"""
    file_hash = hashlib.md5(content).hexdigest()
    filename = f"{file_hash}{ext}"
    hdfs_path = f"{Config.HDFS_ROOT}/{filename}"

    # 通过WSL管道传输到HDFS(避免临时文件)
    cmd_str = f'wsl {Config.HDFS_BIN} dfs -put -f - "{hdfs_path}"'

    process = subprocess.Popen(
        cmd_str,
        shell=True,
        stdin=subprocess.PIPE,
        stdout=subprocess.PIPE,
        stderr=subprocess.PIPE
    )
    stdout, stderr = process.communicate(input=content)

    if process.returncode == 0:
        return hdfs_path
    return None
```

4.3 HBase元数据存储

```
def _handle_file(self, resp, url, url_hash, ext):
    """文件处理: 下载 -> 解析 -> 存储元数据"""
    # ... 文件下载代码省略 ...

    hdfs_path = self.storage.save_file_to_hdfs(content, ext)

    if hdfs_path:
        # 提取文本内容和关键词
        text = ContentParser.parse_text(content, ext)
        keywords = ContentParser.extract_keywords(text)

        # 构造HBase行数据
        data = {
            b'meta:url': url.encode(),
            b'meta:title': fname.encode('utf-8', 'ignore'),
            b'meta:type': b'file',
            b'meta:date': time.strftime("%Y-%m-%d").encode(),
            b'data:hdfs_path': hdfs_path.encode(),
            b'data:content': text[:5000].encode('utf-8', 'ignore'),
            b'index:keywords': keywords.encode('utf-8', 'ignore')
        }
        self.storage.save_metadata(url_hash, data)
```

4.4 三维相关度计分搜索

```
@app.route('/search')
def search():
    query = request.args.get('q', '').strip()
    scored_results = []

    for key, data in table.scan():
        title = data.get(b'meta:title', b'').decode('utf-8', 'ignore')
        content = data.get(b'data:content', b'').decode('utf-8', 'ignore')
        keywords_list = data.get(b'index:keywords', b'').decode().split(',')

        score = 0

        # 基础过滤
        if query not in title and query not in content:
            continue

        # 维度一：标题命中(权重100)
        if query in title:
            score += 100

        # 维度二：关键词命中(权重50)
        if query in keywords_list:
            score += 50

        # 维度三：正文词频(权重1,上限20)
        term_count = content.count(query)
        score += min(term_count, 20)

        scored_results.append({'score': score, ...})

    # 按分数排序
    scored_results.sort(key=lambda x: x['score'], reverse=True)
```

4.5 HDFS文件下载接口

```
@app.route('/download/<path:row_key>')
def download(row_key):
    """从HDFS流式下载文件"""
    # 查询HBase获取HDFS路径
    data = table.row(row_key.encode())
    hdfs_path = data.get(b'data:hdfs_path', b'').decode()
    file_name = data.get(b'meta:title', b'download.file').decode()

    # 通过WSL调用HDFS读取文件
    cmd = f'wsl {HDFS_BIN} dfs -cat "{hdfs_path}"'
    process = subprocess.Popen(cmd, shell=True,
                               stdout=subprocess.PIPE,
                               stderr=subprocess.PIPE)
    file_data, stderr = process.communicate()

    # 流式传输给浏览器
    return send_file(
        io.BytesIO(file_data),
        as_attachment=True,
        download_name=file_name
    )
```

五、个人总结与心得

5.1 踩坑记录

坑1：HBase连接超时

问题：频繁出现`TTransportException: Connection timed out`错误
原因：HBase Thrift服务默认超时较短，网络延迟导致连接断开
解决：增加happybase连接超时参数`timeout=30000`，并在异常时重连

坑2：大文件内存溢出

问题：爬取大型PDF文件时Python进程被系统杀死
原因：`requests.get()`默认将整个响应加载到内存
解决：使用`stream=True`流式下载，分块读取并设置15MB上限

坑3：WSL与Windows文件传输

问题：无法直接使用Python操作WSL中的HDFS
原因：happybase只能连接HBase，无法操作HDFS
解决：通过`subprocess`调用WSL命令，使用管道(`stdin`)传输二进制数据，避免临时文件

坑4：爬虫被封禁

问题：爬取一段时间后出现大量连接失败
原因：请求频率过高触发服务器反爬机制
解决：

- 随机User-Agent轮换
- 随机延时3-6秒
- 检测到封禁后休眠60秒

坑5：HBase中文乱码

问题：存储的中文文件名显示乱码
原因：未指定UTF-8编码
解决：所有字符串使用`.encode('utf-8', 'ignore')`编码，读取时使用`.decode('utf-8', 'ignore')`

5.2 错误总结

1. **忽视异常处理**：初期代码缺少try-catch，单个URL失败导致整个爬虫崩溃
2. **资源未释放**：HBase连接未及时关闭，导致连接池耗尽
3. **硬编码配置**：IP地址写死在代码中，迁移环境时需要修改多处

5.3 实验收获

1. **分布式系统实践**：深入理解了HDFS和HBase的设计理念——HDFS适合大文件顺序读写，HBase适合海量小数据随机访问
2. **爬虫工程化**：学会了如何构建一个健壮的爬虫系统，包括去重、限速、容错、反爬对抗等关键技术
3. **全栈开发能力**：从数据采集、存储设计到Web应用开发，完整经历了一个搜索引擎的全生命周期
4. **调试技巧**：学会了在分布式环境下定位问题，如检查HBase Shell、HDFS命令行、查看日志等
5. **性能优化意识**：理解了为什么需要倒排索引（当前全表扫描在数据量大时会很慢），为后续优化指明方向

5.4 未来改进方向

1. **引入Elasticsearch**：替换全表扫描，实现真正的倒排索引检索

2. **增加分布式爬虫**：使用Scrapy-Redis实现多节点并行爬取
3. **支持增量更新**：定期检测文档变化，更新索引
4. **用户功能完善**：增加用户登录、收藏、搜索历史等功能

六、参考资料

1. Apache HBase官方文档: https://hbase.apache.org/
2. Apache Hadoop HDFS文档: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/
3. Flask官方文档: https://flask.palletsprojects.com/
4. Jieba分词库: https://github.com/fxsjy/jieba
5. Beautiful Soup文档: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
