电工导 C 第四次实验报告

姓名: 宋士祥　　学号:521030910013　　班级:F2103001

2022 年 10 月 19 日

1　实验概览

本次实验学习了使用`lucene` 进行一个简单的网页搜索引擎制作。

`lucene` 是一个基于 java 的全文搜索引擎，已经在多方面有所应用，较为成熟。对于实现网页搜索的方法，我们采用先对网页进行采样进入本地数据库，将其转化为`txt` 文本后进行搜索。

一般地，文档检索基于两个步骤：建立索引和搜索索引。

建立索引的步骤主要为分词、处理、得到索引组件。对于本次实验中采用的`CJKAnalyzer`，其对于仲浩文文本的处理方法为按照汉语的一般规律对每两个汉字进行排列，可以满足大部分的搜索需要。但是这种搜索方式存在弊端（后面会提及）。

搜索索引的步骤主要为输入、语言处理、搜索索引、对结果进行排序几个步骤。对于语言处理方面，一般我们会对文字进行拆解，语义分析等。

基于上述已有知识，本次实验将要进行一个简单的中文网页的搜索引擎。

我们需要首先需要爬取一定数量的中文网页（>5000 网页），然后将其转换为 txt 文档，最后建立相关索引，并实现一个简单的搜索引擎工具。

2　实验环境

本次实验采用所需的实验环境如下：

- Docker 中的`sjtumic/ee208` 镜像
- Python3（使用 VSCode 编译）
- `BeautifulSoup4` 扩展以及`lxml` 扩展。
- `java` 环境及`lucene` 扩展（在 SJTU EE208 中已经给出）。

3　问题重述与代码说明

本题要求我们爬取一定数量（>5k）的中文网页（可利用之前实验爬取的网页），修改`IndexFiles.py` 和`SearchFiles.py`，对这些中文网页建立索引并进行搜索，搜索时需要打印出检出文档的路径、网页标题、url。

本次实验共分为三个部分，`crawler_thread.py`、`IndexFiles.py` 和`SearchFiles.py`。

3.1　crawler_thread.py

该文件是基于 lab3 的多线程爬虫用于对网页进行采样的。本次实验主要对三个主流门户网站（新浪、腾讯、网易）进行采样。同时，由于本次实验需要对网页的标签进行采集，我们对代码进行了一定的改动。

3.1.1 针对添加网页标题和标签的改动

考虑到实验要求我们对网页的标签标题输出，我们需要提前提取出网页标签到index.txt 以便于后续的打印。同时，为了去除对应的标签，我们需要提前获取相关代码如下：

```
soup = BeautifulSoup(content, features="lxml")
title = soup.title.string
filename = valid_filename(page)  # 将网址变成合法的文件名
index = open(index_filename, 'a')
index.write(page.encode('ascii', 'ignore').decode() + "\t" + title + '\t' + filename + "
                                                    \n")
index.close()
index.close()
if not os.path.exists(folder):  # 如果文件夹不存在则新建
    os.mkdir(folder)
content = ''.join(soup.findAll(text = True))
```

3.1.2 针对网页存储的改动

考虑存储的网页仅为网页的网址，我们需要将其改为 txt 格式以便于运行，相应代码如下：

```
def valid_filename(s):
    valid_chars = "-_.() %s%s" % (string.ascii_letters, string.digits)
    s = ''.join(c for c in s if c in valid_chars)
    s += ".txt"
    return s
```

3.2 IndexFiles.py

对于IndexFiles.py 而言，我们需要提前建立索引以满足我们想要正常显示的效果。我们采取的方法为通过读取index.txt 的方式建立索引。我们一般采取逐行读取，相关代码如下：

```
file = open("index.txt", "r")
dictionary = dict()
while True:
    line = file.readline()
    if not line:
        break
    lst = line.split('\t')
    dictionary[lst[1] + '.txt'] = [lst[0]] + lst[2:]
for root, dirnames, filenames in os.walk(root):
    for filename in filenames:
        if not filename.endswith('.txt'):
            continue
        print("adding", filename)
        try:
            path = os.path.join(root, filename)
            file = open(path, encoding='utf-8')
            contents = file.read()
            file.close()
            doc = Document()
            doc.add(Field("name", filename, t1))
```

```
            doc.add(Field("path", path, t1))
            doc.add(Field("url", dictionary[filename][0], t1))
            doc.add(Field("imgsrc", dictionary[filename][1], t1))
            doc.add(Field("title", dictionary[filename][2], t1))
```

3.3 SearchFiles.py

我们注意到原本的`StandradAnalyzer` 无法满足我们的需求，我们需要一个更好的分析工具。`CJKAnalyzer` 是一个较为有效的工具，可以用于解决这一问题。

对于这一情况，我们仅需要`from org.apache.lucene.analysis.cjk import CJKAnalyzer` 并将`analyzer` 替换为`CJKAnalyzer()` 即可。

4 运行结果

我们准备了几个 ppt 中的测试案例和最近的时事热点，分别如下：

```
Hit enter with no input to quit.
Query:战争 游戏

Searching for: 战争 游戏
50 total matching documents.
path: html/httpsappcenter.qzone.qq.comservicesgame_clickappid_via1101070761_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
name: httpsappcenter.qzone.qq.comservicesgame_clickappid_via1101070761_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
url: https://appcenter.qzone.qq.com/services/game_click?appid_via=1101070761_QZSTORE.V6.SIDE-CATEGORIZED-APPS
title: 空间游戏应用中心-网页游戏|热门游戏|新游推荐
------
path: html/httpsappcenter.qzone.qq.comservicesgame_clickappid_via363_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
name: httpsappcenter.qzone.qq.comservicesgame_clickappid_via363_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
url: https://appcenter.qzone.qq.com/services/game_click?appid_via=363_QZSTORE.V6.SIDE-CATEGORIZED-APPS
title: 空间游戏应用中心-网页游戏|热门游戏|新游推荐
------
path: html/httpsappcenter.qzone.qq.comservicesgame_clickappid_via600399_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
name: httpsappcenter.qzone.qq.comservicesgame_clickappid_via600399_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
url: https://appcenter.qzone.qq.com/services/game_click?appid_via=600399_QZSTORE.V6.SIDE-CATEGORIZED-APPS
title: 空间游戏应用中心-网页游戏|热门游戏|新游推荐
------
path: html/httpsappcenter.qzone.qq.comservicesgame_clickappid_via600464_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
name: httpsappcenter.qzone.qq.comservicesgame_clickappid_via600464_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
url: https://appcenter.qzone.qq.com/services/game_click?appid_via=600464_QZSTORE.V6.SIDE-CATEGORIZED-APPS
title: 空间游戏应用中心-网页游戏|热门游戏|新游推荐
------
path: html/httpsappcenter.qzone.qq.comservicesgame_clickappid_via600867_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
name: httpsappcenter.qzone.qq.comservicesgame_clickappid_via600867_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
url: https://appcenter.qzone.qq.com/services/game_click?appid_via=600867_QZSTORE.V6.SIDE-CATEGORIZED-APPS
title: 空间游戏应用中心-网页游戏|热门游戏|新游推荐
------
path: html/httpsappcenter.qzone.qq.comservicesgame_clickappid_via601249_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
name: httpsappcenter.qzone.qq.comservicesgame_clickappid_via601249_QZSTORE.V6.SIDE-CATEGORIZED-APPS.txt
```

```
Hit enter with no input to quit.
Query:战争 NOT 游戏

Searching for: 战争 NOT 游戏
50 total matching documents.
path: html/httpv.163.comspecialopencourseamericanrevolution.html.txt
name: httpv.163.comspecialopencourseamericanrevolution.html.txt
url: http://v.163.com/special/opencourse/americanrevolution.html
title: 耶鲁大学公开课：美国独立战争-网易公开课
------
path: html/httpsmil.sina.cnsd2020-04-23detail-iircuyvh9344850.d.htmlvt4cid65898node_id65898.txt
name: httpsmil.sina.cnsd2020-04-23detail-iircuyvh9344850.d.htmlvt4cid65898node_id65898.txt
url: https://mil.sina.cn/sd/2020-04-23/detail-iircuyvh9344850.d.html?vt=4&cid=65898&node_id=65898
title: 为何不进行全民动员_手机新浪网
------
path: html/httpsmil.sina.cnsd2020-04-13detail-iircuyvh7463302.d.htmlvt4cid65898node_id65898.txt
name: httpsmil.sina.cnsd2020-04-13detail-iircuyvh7463302.d.htmlvt4cid65898node_id65898.txt
url: https://mil.sina.cn/sd/2020-04-13/detail-iircuyvh7463302.d.html?vt=4&cid=65898&node_id=65898
title: 航母不再重要？美军想靠这种黑科技打赢下一场战争_手机新浪网
------
path: html/httpbook.qq.combook-detail44964250.txt
name: httpbook.qq.combook-detail44964250.txt
url: http://book.qq.com/book-detail/44964250
title: 龙族（1-3合集）（修订版）(江南)小说_龙族（1-3合集）（修订版）全文在线阅读下载|无弹窗全文阅读-QQ阅读
------
path: html/httpv.163.comspecialopencoursetorture.html.txt
name: httpv.163.comspecialopencoursetorture.html.txt
url: http://v.163.com/special/opencourse/torture.html
title: 芝加哥大学公开课：酷刑、法律与战争-网易公开课
------
path: html/httpsnew.qq.comraina20221012A00HJC00.txt
name: httpsnew.qq.comraina20221012A00HJC00.txt
url: https://new.qq.com/rain/a/20221012A00HJC00
title: 【喜迎二十大·特稿】为世界永续和平发展贡献中国力量--“中国式现代化”深度探析_腾讯新闻
```

```
Hit enter with no input to quit.
Query:二十大

Searching for: 二十大
50 total matching documents.
path: html/httpswww.163.comgovarticleHJFINTOF002398HK.htmlclickfromw_gov.txt
name: httpswww.163.comgovarticleHJFINTOF002398HK.htmlclickfromw_gov.txt
url: https://www.163.com/gov/article/HJFINTOF002398HK.html?clickfrom=w_gov
title: 喜迎二十大|各地开展喜迎二十大主题文明实践活动|志愿服务|平度市_网易政务
------
path: html/httpsv.qq.comxpagek3359vvthda.html.txt
name: httpsv.qq.comxpagek3359vvthda.html.txt
url: https://v.qq.com/x/page/k3359vvthda.html
title: 新时代之声_高清1080P在线观看平台_腾讯视频
------
path: html/httpview.inews.qq.comaUTR2022100800156700.txt
name: httpview.inews.qq.comaUTR2022100800156700.txt
url: http://view.inews.qq.com/a/UTR2022100800156700
title: 腾讯新闻
------
path: html/httpswww.163.comgovarticleHJFJ13I2002398HK.htmlclickfromw_gov.txt
name: httpswww.163.comgovarticleHJFJ13I2002398HK.htmlclickfromw_gov.txt
url: https://www.163.com/gov/article/HJFJ13I2002398HK.html?clickfrom=w_gov
title: 喜迎二十大丨一线党员以奋斗书写时代荣光|党支部|蒋卫东|孙鹏_网易政务
------
path: html/httpwww.xinhuanet.com.txt
name: httpwww.xinhuanet.com.txt
url: http://www.xinhuanet.com/
title: 新华网_让新闻离你更近
------
path: html/httpsv.qq.comxpagew0044ekh44c.html.txt
name: httpsv.qq.comxpagew0044ekh44c.html.txt
```

```
Hit enter with no input to quit.
Query:蔡徐坤

Searching for: 蔡徐坤
50 total matching documents.
path: html/httpsv.qq.comxcovermzc00200nm5itm6.html.txt
name: httpsv.qq.comxcovermzc00200nm5itm6.html.txt
url: https://v.qq.com/x/cover/mzc00200nm5itm6.html
title: 第6季腾讯视频_综艺_高清1080P在线观看平台
------
path: html/httpsv.qq.comxcovermzc00200te7ifuq.html.txt
name: httpsv.qq.comxcovermzc00200te7ifuq.html.txt
url: https://v.qq.com/x/cover/mzc00200te7ifuq.html
title: 第6季腾讯视频_综艺_高清1080P在线观看平台
------
path: html/httpsv.qq.comxcovermzc00200gtc10mp.html.txt
name: httpsv.qq.comxcovermzc00200gtc10mp.html.txt
url: https://v.qq.com/x/cover/mzc00200gtc10mp.html
title: 沸腾校园腾讯视频_综艺_高清1080P在线观看平台
------
path: html/httpsv.qq.comxcovermzc00200gtc10mph0044igk9f9.html.txt
name: httpsv.qq.comxcovermzc00200gtc10mph0044igk9f9.html.txt
url: https://v.qq.com/x/cover/mzc00200gtc10mp/h0044igk9f9.html
title: 《沸腾校园》第7期下：四公来袭！INTO1献唱推广曲_综艺_高清1080P在线观看平台_腾讯视频
------
path: html/httpsv.qq.comxcovermzc00200gtc10mps0044oxat76.html.txt
name: httpsv.qq.comxcovermzc00200gtc10mps0044oxat76.html.txt
url: https://v.qq.com/x/cover/mzc00200gtc10mp/s0044oxat76.html
title: 社团12小时极限编舞_综艺_高清1080P在线观看平台_腾讯视频
------
path: html/httpsv.qq.comxcovermzc002003cloofr.html.txt
name: httpsv.qq.comxcovermzc002003cloofr.html.txt
url: https://v.qq.com/x/cover/mzc002003cloofr.html
```

5 问题探究与拓展思考

5.1 StandradAnaylzer 为何不适用

在中文分词的语境中，StandradAnalyzer 很难发挥出其应有的作用。例如，当笔者使用 StandradAnalyzer 搜索“上海交大”时，呈现出的第一个义项包含了“上海”、“交”、“大”这四个关键词，但是很显然这并不是我们要的答案。这也就是将中文单字分割存在的问题。但如果我们采用 CJK 这种双字节的问题，再次搜索时就避免了这一问题。如下：

```
Hit enter with no input to quit.
Query:上海交大

Searching for: 上海交大
50 total matching documents.
path: html/httpswww.163.comeduarticleHAHLEB3M00297VGM.html.txt
name: httpswww.163.comeduarticleHAHLEB3M00297VGM.html.txt
url: https://www.163.com/edu/article/HAHLEB3M00297VGM.html
title: C9、E9、华五、中九、五院四系、二龙四虎……这些黑话你懂吗？|志愿填报|大学_网易教育
------
path: html/httpbiz.finance.sina.com.cnfuturesask.txt
name: httpbiz.finance.sina.com.cnfuturesask.txt
url: http://biz.finance.sina.com.cn/futures/ask/
title: 期货专家坐堂_首页_财经频道_新浪网
------
path: html/httpswww.163.comeduarticleH9DILVH200297VGM.html.txt
name: httpswww.163.comeduarticleH9DILVH200297VGM.html.txt
url: https://www.163.com/edu/article/H9DILVH200297VGM.html
title: 清华调整强基计划方案：初试分省计算机考试，复试分区域|清华大学_网易教育
------
path: html/httpedu.sina.com.cn.txt
name: httpedu.sina.com.cn.txt
url: http://edu.sina.com.cn/
title: 教育频道_新浪教育_新浪网
------
path: html/httpzhiyuan.edu.sina.cnvt4pos108.txt
name: httpzhiyuan.edu.sina.cnvt4pos108.txt
url: http://zhiyuan.edu.sina.cn/?vt=4&pos=108
title: 教育频道_新浪教育_新浪网
------
path: html/httpsv.qq.comxpagep004418tr7x.html.txt
name: httpsv.qq.comxpagep004418tr7x.html.txt
```

```
Hit enter with no input to quit.
Query:上海交大 NOT 复旦

Searching for: 上海交大 NOT 复旦
50 total matching documents.
path: html/httpbiz.finance.sina.com.cnfuturesask.txt
name: httpbiz.finance.sina.com.cnfuturesask.txt
url: http://biz.finance.sina.com.cn/futures/ask/
title: 期货专家坐堂_首页_财经频道_新浪网
------
path: html/httpswww.163.comeduarticleH9DILVH200297VGM.html.txt
name: httpswww.163.comeduarticleH9DILVH200297VGM.html.txt
url: https://www.163.com/edu/article/H9DILVH200297VGM.html
title: 清华调整强基计划方案：初试分省计算机考试，复试分区域|清华大学_网易教育
------
path: html/httpsv.qq.comxpagep004418tr7x.html.txt
name: httpsv.qq.comxpagep004418tr7x.html.txt
url: https://v.qq.com/x/page/p004418tr7x.html
title: 上海交响乐团2022-23乐季开幕！埃尔加刮起英伦风_高清1080P在线观看平台_腾讯视频
------
path: html/httpswww.163.comnewsarticleHASE7HSN00018AOQ.html.txt
name: httpswww.163.comnewsarticleHASE7HSN00018AOQ.html.txt
url: https://www.163.com/news/article/HASE7HSN00018AOQ.html
title: 谁是钱七虎？“消失”16年“修长城”，还在珠海搞出“天下第一爆”！|院士|教授|力学|中国工程院_网易新闻
------
path: html/httpswww.163.comdyarticleH89OK1H20530WJIN.html.txt
name: httpswww.163.comdyarticleH89OK1H20530WJIN.html.txt
url: https://www.163.com/dy/article/H89OK1H20530WJIN.html
title: 山东第一医科大学2022年普通高等教育招生章程|本科|入学|学校_网易订阅
------
```

5.2 实验有何不足

一方面的不足之处在于样本量较少。由于时间原因我们爬取的网页数目并不是很全面。这导致如果我们想要寻找的是一些小众的内容，会很难找到正确的解答，比如我们搜索一个小众番剧时：

```
Hit enter with no input to quit.
Query:缘之空

Searching for: 缘之空
4 total matching documents.
path: html/httpsv.qq.comxcovericklrc9ls2gruri.html.txt
name: httpsv.qq.comxcovericklrc9ls2gruri.html.txt
url: https://v.qq.com/x/cover/icklrc9ls2gruri.html
title: 寒战_电影_高清1080P在线观看平台
------
path: html/httpsv.qq.comxcoverd0womh06pzk3h9k.html.txt
name: httpsv.qq.comxcoverd0womh06pzk3h9k.html.txt
url: https://v.qq.com/x/cover/d0womh06pzk3h9k.html
title: 西游记女儿国_电影_高清1080P在线观看平台
------
path: html/httpbook.qq.combook-detail44215435.txt
name: httpbook.qq.combook-detail44215435.txt
url: http://book.qq.com/book-detail/44215435
title: 野枪（全集）(何楚舞)小说_野枪（全集）新人全文免费阅读|全文在线阅读下载-QQ阅读
------
path: html/httpbook.qq.combook-detail521449.txt
name: httpbook.qq.combook-detail521449.txt
url: http://book.qq.com/book-detail/521449
title: 九幽天帝(给力)小说_九幽天帝新人全文免费阅读|全文在线阅读下载-QQ阅读
------
```

而且，由于现在有很多网页使用动态网页的技术。这样爬取出的网页会出现许多 javascript 语言。一些程序员会在程序里添加一些注释，加入一些彩蛋等，会使结果有所误差。

另一方面，CJKAnalyzer 的特性也使得查询结果相对不准确。比如我们搜索“二十大”时，结果可能会被理解为“二”、“十大”，这样就会出现“第二...”“十大”这样相对不准确的答案。

一个更加好的方法为使用 SmartChineseAnalyzer 或 IKAnalyzer，前者为中科院 ICTCLAS 开发，二者均有不错的效果（详情见 ref）。

二者都存在一个很大的问题在于 SJTUEE208 的 docker 均未安装这两个库，而且这两个库都是基于 java 编写，这就意味着我们需要重新安装 lucene 以适应。

一个折中的办法为采用 jpype 的方式导入这一 analyzer。然而这一做法的问题在于 python 程序不支持同时开启两个 java 虚拟机（在 lucene 已经采用了一个虚拟机，不能再申请一个 JVM(Java Virtual Machine) 来解决）。

因此最好的解决方法为重装。

Reference

开源中文分词框架分词效果对比 smartcn 与 IKanalyzer_ 阿里云：
https://developer.aliyun.com/article/30900