

# 第四次实验报告

学号：518030910308

姓名：刘文轩

## 一、实验准备

### 1、实验环境介绍

操作系统：ubuntu 14.04

语言：Python 2

IDE：Pycharm 2019.2.3

### 2、实验目的

2.1 学会创建中文索引，了解 lucene

2.2 学会搜索索引

2.3 运用第三方库实现中文分词

2.4 实现中文网页索引

### 3、实验思路

3.1 通过并行化爬虫爬取相对较大规模的网页

3.2 修改 IndexFiles.py，创建中文索引，从 index.txt 中获取 url 和对应的文件名，并且到 html 文件夹中找到相关文件，创建中文索引

3.3 修改 SearchFiles.py，搜索中文索引并返回结果

## 二、实验过程

### 1、爬取较大规模的网页集

#### 1.1 修改并行化爬虫代码

由于这一次实验中需要的网站量较大，因此直接使用串行爬虫速度较慢。因而我们选择对于并行化爬虫的代码加以修改。

我们这一次选择 8 线程并行，将获取页面的最大值设为 5100。同时，我们考虑到在显示器上输出爬取结果很耗时间，但是不显示可能又无法准确判断程序是否在正常运行以及预计还要多久才能结束，因此这次加入了一个判断语句，当新增爬取数量达到 50 的倍数时，我们将这个值输出一下。

我们选择的起始网站是 <http://news.sjtu.edu.cn>。

输出已爬取网页数量的部分代码如下：

```

def working():
    global count, max_page, crawled, q
    while count < max_page and q.qsize > 0:
        page = q.get()
        if page not in crawled:
            try:
                content = get_page(page)
                if content == '':
                    continue
                else:
                    add_page_to_folder(page, content)
                    crawled.append(page)
                    outlinks = get_all_links(content, page)
                    for link in outlinks:
                        q.put(link)
                    if count % 50 == 0:
                        print count
                    if varLock.acquire():
                        count += 1
                        varLock.release()
                    q.task_done()
            except:
                continue

```

## 1.2 爬取结果

我们在此部分展示我们爬取得到的 index.txt 和 html 文件夹。

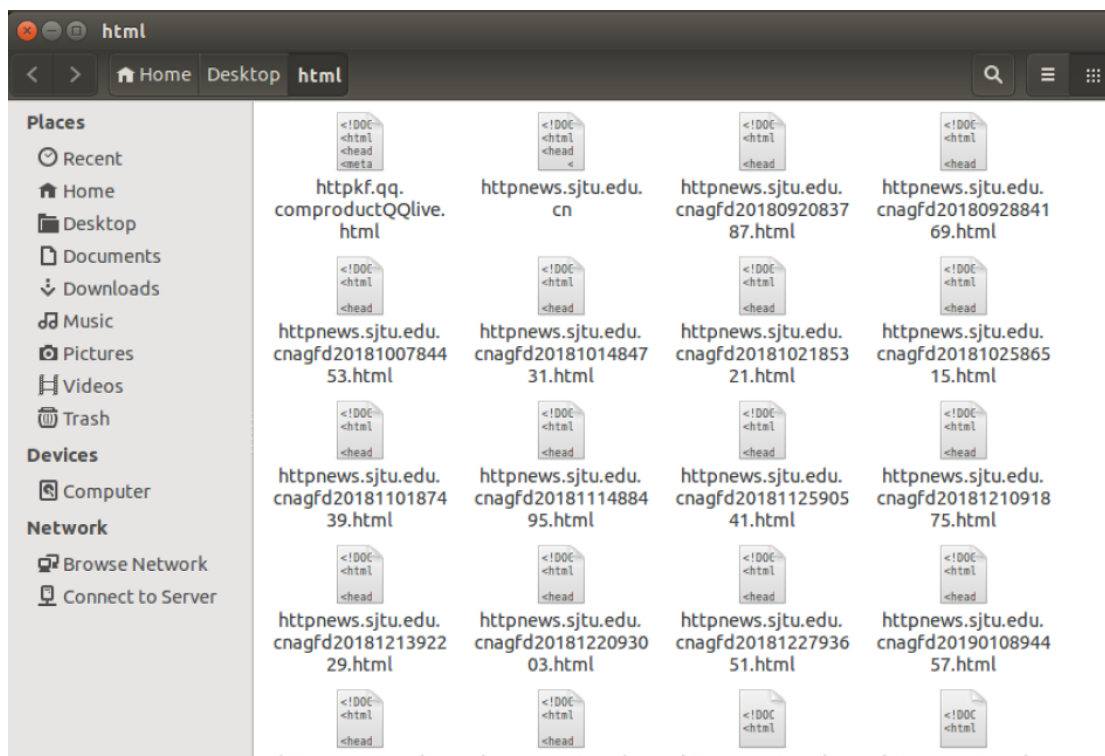
index.txt x

```

http://news.sjtu.edu.cn httpnews.sjtu.edu.cn
http://news.sjtu.edu.cn/jdyw/20191017/112851.html httpnews.sjtu.edu.cnjdyw20191017112851.html
http://news.sjtu.edu.cn/jdyw/20191018/112921.html httpnews.sjtu.edu.cnjdyw20191018112921.html
http://news.sjtu.edu.cn/jdyw/20191018/112899.html httpnews.sjtu.edu.cnjdyw20191018112899.html
http://news.sjtu.edu.cn/jdyw/20191014/112535.html httpnews.sjtu.edu.cnjdyw20191014112535.html
http://news.sjtu.edu.cn/jdyw/20191014/112627.html httpnews.sjtu.edu.cnjdyw20191014112627.html
http://news.sjtu.edu.cn/jdyw/20191016/112793.html httpnews.sjtu.edu.cnjdyw20191016112793.html
http://news.sjtu.edu.cn/jdyw/index.html httpnews.sjtu.edu.cnjdywindex.html
http://news.sjtu.edu.cn/jdyw/index.html httpnews.sjtu.edu.cnjdywindex.html
http://news.sjtu.edu.cn/jdyw/20191018/113101.html httpnews.sjtu.edu.cnjdyw20191018113101.html
http://news.sjtu.edu.cn/jdyw/20191018/112893.html httpnews.sjtu.edu.cnjdyw20191018112893.html
http://news.sjtu.edu.cn/jdyw/20191018/112919.html httpnews.sjtu.edu.cnjdyw20191018112919.html
http://news.sjtu.edu.cn/jdyw/20191018/112925.html httpnews.sjtu.edu.cnjdyw20191018112925.html
http://news.sjtu.edu.cn/jdyw/20191017/112849.html httpnews.sjtu.edu.cnjdyw20191017112849.html
http://news.sjtu.edu.cn/ntjj/20191014/112617.html httpnews.sjtu.edu.cnntjj20191014112617.html
http://news.sjtu.edu.cn/jdyw/20191018/112923.html httpnews.sjtu.edu.cnjdyw20191018112923.html
http://news.sjtu.edu.cn/ntjj/index.html httpnews.sjtu.edu.cnntjjindex.html
http://news.sjtu.edu.cn/ntjj/20191016/112763.html httpnews.sjtu.edu.cnntjj20191016112763.html
http://news.sjtu.edu.cn/ntjj/20191015/112739.html httpnews.sjtu.edu.cnntjj20191015112739.html
http://news.sjtu.edu.cn/ntjj/20191015/112735.html httpnews.sjtu.edu.cnntjj20191015112735.html
http://news.sjtu.edu.cn/ntjj/20191015/112743.html httpnews.sjtu.edu.cnntjj20191015112743.html
http://news.sjtu.edu.cn/jdzh/index.html httpnews.sjtu.edu.cnjdzhindex.html
http://news.sjtu.edu.cn/ntjj/20191015/112727.html httpnews.sjtu.edu.cnntjj20191015112727.html
http://news.sjtu.edu.cn/ntjj/20191015/112723.html httpnews.sjtu.edu.cnntjj20191015112723.html
http://news.sjtu.edu.cn/tsfx/index.html httpnews.sjtu.edu.cnstfxindex.html
http://news.sjtu.edu.cn/jdzh/20191018/113069.html httpnews.sjtu.edu.cnjdzh20191018113069.html
http://news.sjtu.edu.cn/jdzh/20191018/113071.html httpnews.sjtu.edu.cnjdzh20191018113071.html
http://news.sjtu.edu.cn/jdzh/20191012/112457.html httpnews.sjtu.edu.cnjdzh20191012112457.html
http://news.sjtu.edu.cn/jdzh/20191010/112203.html httpnews.sjtu.edu.cnjdzh20191010112203.html
http://news.sjtu.edu.cn/xsjz/index.html httpnews.sjtu.edu.cnxsjzindex.html
http://news.sjtu.edu.cn/ztzl/index.html httpnews.sjtu.edu.cnztzlindex.html
http://news.sjtu.edu.cn/ztzl_jdms/20191010/112263.html httpnews.sjtu.edu.cnztzl_jdms20191010112263.html
http://news.sjtu.edu.cn/ztzl_jdms/20191016/112789.html httpnews.sjtu.edu.cnztzl_jdms20191016112789.html
http://news.sjtu.edu.cn/ztzl_jdms/20190923/110843.html httpnews.sjtu.edu.cnztzl_jdms20190923110843.html
http://news.sjtu.edu.cn/ztzl_jdms/20190930/111735.html httpnews.sjtu.edu.cnztzl_jdms20190930111735.html
http://news.sjtu.edu.cn/ztzl_xyfc/20191016/112767.html httpnews.sjtu.edu.cnztzl_xyfc20191016112767.html
http://news.sjtu.edu.cn/ztzl_jdms/20190916/110343.html httpnews.sjtu.edu.cnztzl_jdms20190916110343.html
http://news.sjtu.edu.cn/ztjy/20191010/112261.html httpnews.sjtu.edu.cnztjy20191010112261.html

```

Plain Text ▾ Tab Width: 8 ▾ Ln 1, Col 1 INS



## 2、修改 SearchFils.py，创建中文索引

### 2.1 修改文件获取方式与路径

首先，我们在上一实验中创建了 index.txt，保存了我们爬取的网站的 url 与我们保存的文件名，并将对应文件保存在了 html 这一文件夹中，因此我们首先要修改文件的获取方式以及路径。

考虑到 index.txt 中每一行以空格间隔，第一段为 url，第二段为文件名，我们将相关代码修改如下：

```
if not root.endswith('.txt'):
    print "Please give the index file end with .txt !"
    return

index_file = open(root)
for line in index_file.readlines():
    url_and_name = line.split()
    url = url_and_name[0]
    filename = url_and_name[1]
```

### 2.2 设置中文分词与创建索引

在这里我们使用 jieba 库来进行中文分词，由于 jieba 库划分好的结果是以空格间隔开来的，因此我们修改原来的 StandardAnalyzer 为 WhitespaceAnalyzer，WhitespaceAnalyzer 使用空格作为间隔符的词汇分割分词器。处理词汇单元的时候，以空格字符作为分割符号。分词器不做词汇过滤，也不进行小写字母转换。

```

import sys, os, lucene, threading, time, re
import jieba
from datetime import datetime
from bs4 import BeautifulSoup

from java.io import File
from org.apache.lucene.analysis.miscellaneous import LimitTokenCountAnalyzer
from org.apache.lucene.analysis.core import WhitespaceAnalyzer
from org.apache.lucene.document import Document, Field, FieldType
from org.apache.lucene.index import FieldInfo, IndexWriter, IndexWriterConfig
from org.apache.lucene.store import SimpleFSDirectory
from org.apache.lucene.util import Version

analyzer = WhitespaceAnalyzer(Version.LUCENE_CURRENT)
IndexFiles('index.txt', "index", analyzer)

```

由于我们最终搜索时仅仅依据网页的标题来返回搜索结果，因此我们在此仅将读取到的网页的 title 进行分词。

```

path = os.path.join("html", filename)
file = open(path)
contents = file.read()
soup = BeautifulSoup(contents, features="html.parser")
if soup.head.title:
    title = soup.head.title.string
    title = jieba.cut(title)
    if title:
        title = ' '.join(title)
    else:
        title = ' '
else:
    title = ' '

```

接下来我们设置 field 属性，我们不采取原代码中分别设置 t1 和 t2 的方式，而转而选择一种更加直观的方法，那就是在添加 field 时，直接指出对应的属性，在此，只有 title 需要被设置为要被索引与分析。

```

doc = Document()
doc.add(Field("title", title,
              Field.Store.YES,
              Field.Index.ANALYZED))
doc.add(Field("name", filename,
              Field.Store.YES,
              Field.Index.NOT_ANALYZED))
doc.add(Field("path", path,
              Field.Store.YES,
              Field.Index.NOT_ANALYZED))
doc.add(Field("url", url,
              Field.Store.YES,
              Field.Index.NOT_ANALYZED))
if len(contents) > 0:
    doc.add(Field("contents", contents,
                  Field.Store.NO,
                  Field.Index.NOT_ANALYZED))
else:
    print "warning: no content in %s" % filename
writer.addDocument(doc)

```

### 3、修改 SearchFiles.py，实现中文搜索索引

#### 3.1 修改 Analyzer 与分词器

需要注意的是，我们应当确保索引创建与搜索时使用的 analyzer 相同，因此我们同样需要修改 SearchFiles.py 中的 analyzer 为 WhitespaceAnalyzer，并使用 jieba 库进行中文分词：

```

import sys, os, lucene
import jieba

from java.io import File
from org.apache.lucene.analysis.core import WhitespaceAnalyzer
from org.apache.lucene.index import DirectoryReader
from org.apache.lucene.queryparser.classic import QueryParser
from org.apache.lucene.store import SimpleFSDirectory
from org.apache.lucene.search import IndexSearcher
from org.apache.lucene.util import Version

```

```

command = raw_input("Query:")
command = unicode(command, 'utf-8')
command = jieba.cut(command)
command = ' '.join(command)
if command == '':
    return

```

#### 3.2 设置索引输出

我们在输出题目要求的数据时，由于存入的 title 是已经用空格分词后的结果，因此我们将其再次回复为原来连贯的形态。



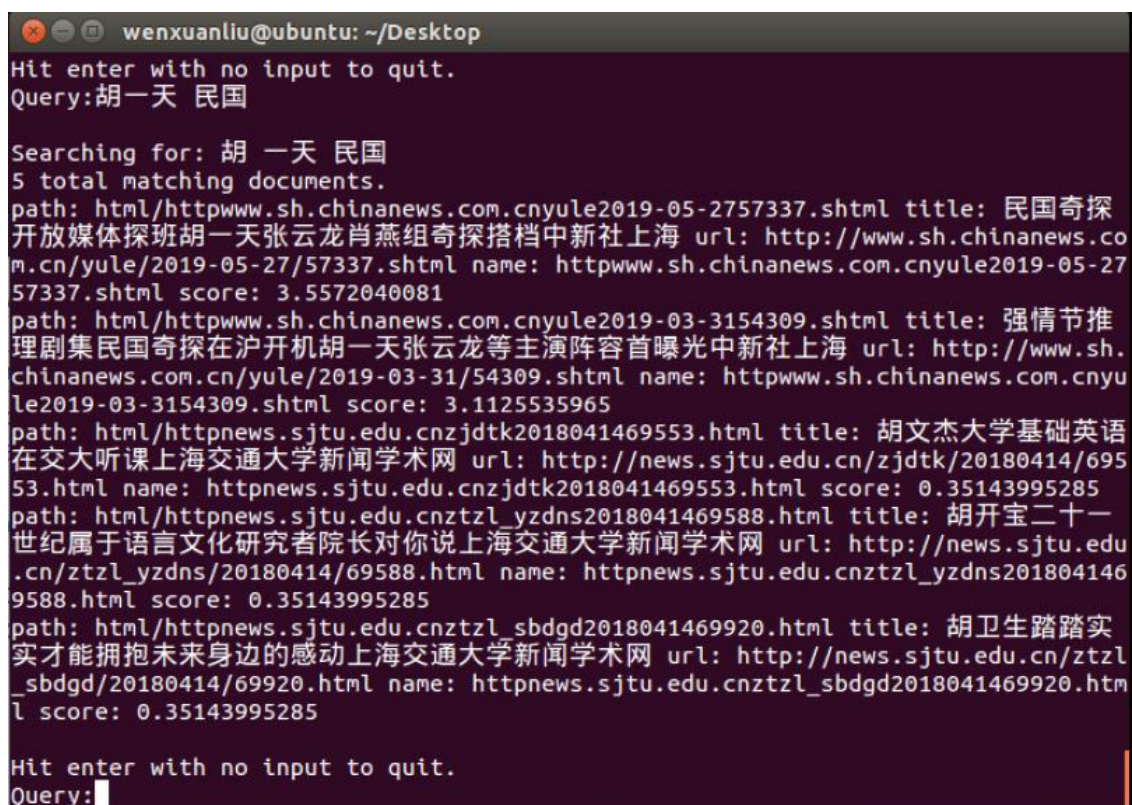
```

for i, scoreDoc in enumerate(scoreDocs):
    doc = searcher.doc(scoreDoc.doc)
    print 'path:', doc.get("path"), \
          'title:', doc.get("title").encode("utf-8").replace(' ',''), \
          'url:', doc.get("url"), \
          'name:', doc.get("name"), 'score:', scoreDoc.score

```

### 3.3 代码运行结果

至此，两个相关程序的修改就已经完成，我们在已经运行了 IndexFiles.py 的情况下，在命令行中进行搜索尝试：



```

wenxuanliu@ubuntu: ~/Desktop
Hit enter with no input to quit.
Query:胡一天 民国

Searching for: 胡 一天 民国
5 total matching documents.
path: html/httpwww.sh.chinanews.com.cnyule2019-05-2757337.shtml title: 民国奇探
开放媒体探班胡一天张云龙肖燕组奇探搭档中新社上海 url: http://www.sh.chinanews.com.cn/yule/2019-05-27/57337.shtml name: httpwww.sh.chinanews.com.cnyule2019-05-2757337.shtml score: 3.5572040081
path: html/httpwww.sh.chinanews.com.cnyule2019-03-3154309.shtml title: 强情节推理剧集民国奇探在沪开机胡一天张云龙等主演阵容首曝光中新社上海 url: http://www.sh.chinanews.com.cn/yule/2019-03-31/54309.shtml name: httpwww.sh.chinanews.com.cnyule2019-03-3154309.shtml score: 3.1125535965
path: html/httpnews.sjtu.edu.cnzjdk2018041469553.html title: 胡文杰大学基础英语在交大听课上海交通大学新闻学术网 url: http://news.sjtu.edu.cn/zjdk/20180414/69553.html name: httpnews.sjtu.edu.cnzjdk2018041469553.html score: 0.35143995285
path: html/httpnews.sjtu.edu.cnztzl_yzdns2018041469588.html title: 胡开宝二十一世纪属于语言文化研究者院长对你说上海交通大学新闻学术网 url: http://news.sjtu.edu.cn/ztzl_yzdns/20180414/69588.html name: httpnews.sjtu.edu.cnztzl_yzdns2018041469588.html score: 0.35143995285
path: html/httpnews.sjtu.edu.cnztzl_sbdgd2018041469920.html title: 胡卫生踏踏实实才能拥抱未来身边的感动上海交通大学新闻学术网 url: http://news.sjtu.edu.cn/ztzl_sbdgd/20180414/69920.html name: httpnews.sjtu.edu.cnztzl_sbdgd2018041469920.html score: 0.35143995285

Hit enter with no input to quit.
Query:

```

由图可见，搜索结果还是比较恰当的。得分高的数据也就是我们需要的结果。

## 三、实验总结

### 1、实验概述

本次实验的主要任务，可以总结为创建中文索引并实现中文索引搜索，内容涉及中文分词，总体相当重要。

### 2、感想总结

在这次的实验中，学会了的东西有很多，其中最重要的就是提高了自己处理问题、收集相关资料、解决问题的能力，这在我们将来的学习和工作生活中都是很重要的。而具体细化开来，在本次实验中：

- 2.1 了解了 lucene 的有关内容
- 2.2 学会了运用 jieba 库实现中文分词
- 2.3 学会了创建中文索引
- 2.4 学会了中文索引搜索

### 3、创新点

首先是为了更好承接上一次实验的内容, 我将索引文件获取的过程拆分成从 index.txt 中获取 url 与文件名, 再从 html 文件夹中找到文件。

```
if not root.endswith('.txt'):
    print "Please give the index file end with .txt !"
    return

index_file = open(root)
for line in index_file.readlines():
    url_and_name = line.split()
    url = url_and_name[0]
    filename = url_and_name[1]
```

同时我使用了一种相对而言更加直接的设置 field 的方式, 使得代码更加直观明了:

```
doc = Document()
doc.add(Field("title", title,
              Field.Store.YES,
              Field.Index.ANALYZED))
doc.add(Field("name", filename,
              Field.Store.YES,
              Field.Index.NOT_ANALYZED))
doc.add(Field("path", path,
              Field.Store.YES,
              Field.Index.NOT_ANALYZED))
doc.add(Field("url", url,
              Field.Store.YES,
              Field.Index.NOT_ANALYZED))
if len(contents) > 0:
    doc.add(Field("contents", contents,
                  Field.Store.NO,
                  Field.Index.NOT_ANALYZED))
else:
    print "warning: no content in %s" % filename
writer.addDocument(doc)
```

### 4、遇到的问题

由于中文分词需要我们自己完善, 因此一开始由于 try-except 部分的代码, 我无法观察到完整的报错信息, 在去掉后, 将代码调试的更加合格后, 再加上 try-except 来加强程序的适应性。

同时我们在爬取一开始的文件时, 如果不加限制, 会得到很多我并不想要的 exe, pdf 等文件, 因此我在后期的代码中限制了爬取的文件为 html 文件, 使得后期处理更加方便。