实验 5-Lucene2

姓名：王煌基　　学号：519030910100　　班级：F1903004

一、实验概览

基于对 Lucene 的初步认识，学习在多个域（Field）下的组合查询以及索引中的文档的更新来实现搜索带限定词的索引情况，同时学习图片搜索以完成对图片的文字标识进行索引

二、实验环境

1. 个人笔记本电脑
2. 操作系统：windows10 专业版
3. 使用软件：Visual Studio Code；Docker Desktop

三、实验过程

练习一：模拟实现搜索引擎的 “site:” 功能（对搜索的网站进行限制）

示例：搜索 sina.com.cn 和 baike.baidu.com 上包含“国家”的网页。提示：可以在原先的索引上添加一个可以索引的 domain 域，来对网址所在的域名进行索引。

【解】

基于在 lab4 中对 Lucene 的初步认识,我们已经能够简单地建立一定的索引并进行查找，只是这个任务中不同的地方在于限定了搜索的范围，而这显然更具有现实意义。

在实验 1 中，我们对 path、title、url、name、contents 都分别建立了个域，且前四项的域类型为 t1，最后的 contents 的域类型为 t2，t1 负责存储且 t2 负责分词：

```
t1 = FieldType()
t1.setStored(True)  #是否存储
t1.setTokenized(False)  #是否分词
t1.setIndexOptions(IndexOptions.NONE)  # Not Indexed
t2 = FieldType()
t2.setStored(False)
t2.setTokenized(True)
t2.setIndexOptions(IndexOptions.DOCS_AND_FREQS_AND_POSITIONS)
```

然后将它们分别写入文档中建立索引：

```
doc = Document()
doc.add(Field("path", path, t1))
doc.add(Field("title", title, t1))
doc.add(Field("url", dic[filename], t1))
doc.add(Field("name", filename, t1))
#对 contents 利用 jieba 进行分词操作
if len(contents) > 0:
    doc.add(Field("contents", contents, t2))
```

```
9.  else:
10.     print("warning: no content in %s" % filename)
11. writer.addDocument(doc)
```

通过对 lab5 的 IndexFiles.py 的对比可以发现，实际上 t2 是一个可以专门用来提供检索条目的域，这样我们只需要保证我们建立的 domain 索引是以 t2 为域即可，网址的域名获取通过函数 parse.urlparse(url).hostname[1]即可，site 和 contents 都需要建立域：

```
1.  site = parse.urlparse(dic[filename]).hostname
2.  doc.add(Field("site", site, t2))
3.  doc.add(Field("contents", contents, t2))
```

显然，site 并不需要分词，而 contents 的分词在 lab4 中已经完成，故我们的 IndexFiles.py 已经修改完毕，下面对 SearchFiles.py 进行修改。

SearchFiles.py 与 lab4 不同的地方在于检索的过程中存在了多个域的限制，故我们在具体的不同域的时候也要注意分离方式，如**长江流域 site:www.17yy.com title:无敌网站**，我们需要对**长江流域**分词为**长江 流域**，title 的**无敌网站**分词为**无敌 网站**，只是本次练习中只考虑 site 域，故可以不需要进行分词改写。

对于我们输入的搜索词进行分块的函数进行修改（原来用 title,author,language 进行限制，现在改成 site 域进行限制即可）：

```
1.  def parseCommand(command):
2.      #allowed_opt = ['title', 'author', 'language']
3.      allowed_opt = ['site']
```

同时，为了表现我们限制之后与限制之前的区别，我们特地修改了我们搜索结果的长度（调整至足够大来充分反映我们的结果量）并且只输出相关度高的前 20 个搜索结果：

```
1.  scoreDocs = searcher.search(querys.build(),10000).scoreDocs
2.  cnt=0
3.
4.  cnt += 1
5.  if(cnt >= 20):
6.      break
```

从而，我们的练习一得到了完成。

练习二：实现一个图片索引

新建一个索引，输入文本，输出相关的图片地址，图片所在网页的网址，图片所在网页的标题。提示：图片周围的文本可能会用到 parser 实验中的 parent, nextSibling, previousSibling 等函数。

做图片索引时最好选定某个你感兴趣的网站爬取。比如只对交大网站上的图片进行索引，这样可以对特定网站的结构进行分析，让搜索结果更精确。

[1] https://blog.csdn.net/fisherming/article/details/89450841 见评论区

【解】

此处我仍旧选择 17yy 的网站进行爬取（基于 17yy 的网站进行爬取），对其中的某个网页分析网页结构后我发现，该网站的 img 标签下 src 属性就是图片的地址，alt 属性就是图片的内容，因此可以简单地建立爬取模型来获取 imgurl（即 src）以及 data（即 alt）：


```
soup = BeautifulSoup(contents,features="html.parser")
for i in soup.findAll('img'):
    imgurl = i.get('src','')
    data= i.get('alt','')
```

同时，为了建立合理的索引，需要将 data 进行分词后再存入 doc 中，并且，为了避免网页之间出现大量重复图片，我们利用一个 url_picture 列表来避免重复，最后将所有内容放入 doc 中即可：

```
if(imgurl in url_picture):
    continue
url_picture.append(imgurl)
#分词
data = jieba.cut(data, cut_all=False)
data = ' '.join(data)
#将内容放入 doc 中，其中 data 需要建立搜索，故需要以 t2 的域
if len(data) > 0:
    doc.add(Field("data", data, t2))
    doc.add(Field("imgurl", imgurl, t1))
    doc.add(Field("url", dic[filename], t1))
```

从而，我们的 Indeximg.py 便实现完毕了，下面进行 Searchimg.py 的实现，本质上与 SearchFiles.py 是类似的，当然，由于我们这次并没有限定具体的域名等等，我们可以直接采用 lab4 的 SearchFiles.py 来实现。

需要注意的是，为了能够完全体现我们的搜索结果，这里将搜索结果扩大至最多 100000 条以贴近现实，并且只输出前 20 个相关度最高的内容。

```
query = QueryParser("data", analyzer).parse(command)
scoreDocs = searcher.search(query, 100000).scoreDocs
print("%s total matching documents." % len(scoreDocs))
cnt = 0
for scoreDoc in scoreDocs:
    doc = searcher.doc(scoreDoc.doc)
    print('imgurl:', doc.get("imgurl"))
    print('url:', doc.get("url"))
    print(doc.get('data').replace(' ',''))
    print("\n------------------------")
    cnt += 1
    if(cnt >= 20):
        break
```

从而我们的练习二得以解决。

四、实验问题及解决

问题：在练习二中我只对 17yy 上的图片进行索引，而该网站的结构简单，如果遇到不同的网站，结构不同的时候应该怎么办比较好呢？

【解】 我尝试着去查看如交大网站、知乎日报、百度百科等网站，并分析它们的图片结构：

上海交通大学主页(https://www.sjtu.edu.cn)：

```
<div class="owl-item cloned" style="width: 940px;">
  <div class="item">
    <a href="https://news.sjtu.edu.cn/ztzl_syh/index.html" title="思源湖·邀你共倘徉" target="_blank"
    class="flex-box block-sm">
      <div class="col-sm-12 no-padding">
        <img src="/resource/upload/202009/20200928_072857_590.jpg" alt="思源湖·邀你共倘徉">
        <h4 class>…</h4>
      </div>
    </a>
  </div>
</div>
<div class="owl-item" style="width: 940px;">
  <div class="item">
    <a href="https://news.sjtu.edu.cn/jdyw/20201019/132601.html" title="刘学新在上海交大调研" target=
    "_blank" class="flex-box block-sm">
      <div class="col-sm-12 no-padding">
        <img src="/resource/upload/202010/20201019_132340_400.jpg" alt="刘学新在上海交大调研">
        <h4 class>…</h4>
      </div>
    </a>
  </div>
</div>
<div class="owl-item" style="width: 940px;">
  <div class="item">
    <a href="https://ddh11.sjtu.edu.cn/jdzs.html" title="迎上海交通大学第十一次党代会" target="_blank"
    class="flex-box block-sm">
      <div class="col-sm-12 no-padding">
        <img src="/resource/upload/202010/20201016_023857_807.png" alt="迎上海交通大学第十一次党代会">
        <h4 class>…</h4>
      </div>
    </a>
  </div>
</div>
```

知乎日报(https://daily.zhihu.com/)：

```
<div class="box">
  <a href="/story/9729106" class="link-button">
    <img src="https://pic2.zhimg.com/v2-fbd0235….jpg?source=8673f162" class="preview-image">
    <span class="title">大误 · 万能食材刘看山</span>
  </a>
</div>
<div class="wrap">
  <div class="box">
    <a href="/story/9729016" class="link-button">
      <img src="https://pic1.zhimg.com/v2-7ef909e….jpg?source=8673f162" class="preview-image">
      <span class="title">瞎扯 · 如何正确地吐槽</span>
    </a>
  </div>
```

百度百科（https://baike.baidu.com/）：

```
<ul style="position: absolute; width: 1500px; height: 316px; left: 0px;">
  <li class="wgt_carousel_aniUnit" style="float: left;">
    <div class="card" alog-alias="publicizing-slide-0-card-0">…</div>
    <div class="card" alog-alias="publicizing-slide-0-card-1">…</div>
    <div class="card" alog-alias="publicizing-slide-0-card-2">
      <img src="https://bkimg.cdn.bcebos.com/pic/908fa0e…?x-bce-process=image/
      resize,m_lfit,w_255,h_340,limit_0">
      <div class="card_shd">
        <a href="https://baike.baidu.com/item/%E6%B2%B3%E9%A9%AC/596547#hotspotmining" target="_blank"></a>
      </div>
      <div class="card_cnt">
        <div class="card_cnt_tit">
          <a href="https://baike.baidu.com/item/%E6%B2%B3%E9%A9%AC/596547#hotspotmining" target="_blank">河
          马打哈欠是因为困？</a>
        </div>
        <div class="card_cnt_cnt">河马的外形看起来并不像马，倒略似一只特别大的猪。事实上，河马在进化过程中的确与猪
        类的亲缘关系更为接近。</div>
      </div>
    </div>
    ::after
  </li>
</ul>
```

淘宝网热卖（网址太长直接超链接附送）：

```
<div class="item">
  <a target="_blank" href="https://click.simba.taobao.com/cc_im?spm=a2e15.8261149.07626516002.4&p=…
  D&union_lens=lensId%3An%401603190561%400b0b2afa_0ec2_175459a9b50_556d%4001" data-spm-anchor-id=
  "a2e15.8261149.07626516002.4" data-spm-act-id="a2e15.8261149.07626516002.4">
    <div class="imgContainer">
      <span class="imgLink">
        <img src="//img.alicdn.com/imgextra/i1/17971277/O1CN01ks0KHq1LIuMQRiqFR_!!0-
        saturn_solar.jpg_220x220.jpg_.webp" data-spm-anchor-id="a2e15.8261149.07626516002.i0.7e3229b4C15vYv">
      </span>
    </div>
    <div class="info">
      <p class="price">…</p>
      <span class="title" title="Vero Moda2020春夏新款工装简约铆钉口袋连">Vero Moda2020春夏新款工装简约铆钉口袋连
      </span>
      <p class="shopNum">…</p>
      <div class="moreInfo">…</div>
    </div>
  </a>
</div>
```

分析以上几个网页之后，可以发现，实际上不同的网页它们对于图片的存储方式未必是相同的，因而如果我们想要设计出一个大一统的算法来获取所有图片的相关文字是相当不容易的一个操作过程，但是通过对上面几个比较大型的网站的存储方式进行分析后，我有一个初步的想法，就是对于一张具体的图片，大概率它的相关文字是网页中从它开始往下读取的第一串中文文字串，并且可以考虑利用这种查找函数去查找到第一个中文文字串，直接将其视为图片的相关介绍（如果实验基数较大，这个过程中出现的错误数量应该可以忽略不计），这样就可以对每一个图片进行文字的匹配（如果没有，则利用 try 语句防止程序陷入崩溃）。

但是这样的想法我并没有去尝试，很大的原因在于实现起来似乎比较麻烦，并且这一块内容在教学大纲中也有提出，故目前暂时略去，等到之后“**第 12 周：图像索引创建　实验 6：图像索引创建及查找**”再进行具体尝试，可能会更全面。

然而，我这个想法应该只能适用于中文网页，对于英文网页，很难判断哪一块的内容是用来存放具体内容的，因而我认为我这个方法并不是一个很高效的方法，也不具有太大的普适性，希望在接下来的学习中能够逐步接触并解决这个问题。

五、实验结果

【练习一：模拟实现搜索引擎的 “site:” 功能（对搜索的网站进行限制）】

运行 IndexFiles.py 建立索引，运行结果：

```
adding httpwww.mumayi.com.html
adding httpwww.muzhiwan.com.html
adding httpwww.niba.com.html
adding httpwww.veryhuo.com.html
adding httpwww.wangye.cn.html
adding httpwww.wanyx.com.html
adding httpwww.yxbao.com.html
adding httpwww.yzz.cn.html
adding httpxiaoyouxi.2345.com.html
commit index
...done
0:03:31.237434
```

运行 SearchFiles.py 进行索引的检索：

检索“小游戏”，看待是否存在域名限制的区别（第一图无限制，第二图有限制）：

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

Searching for: 小游戏
{'contents': '小游戏'}
5186 total matching documents.
-----------------------
path: html/httpwww.17yy.comyouxigonglueyizhi.html
title: 小游戏攻略_小游戏秘籍 - 17yy经典小游戏
url: http://www.17yy.com/youxigonglue/yizhi
name: httpwww.17yy.comyouxigonglueyizhi.html
-----------------------
path: html/httpwww.17yy.comyouxigonglue.html
title: 小游戏攻略_小游戏秘籍 - 17yy经典小游戏
url: http://www.17yy.com/youxigonglue/
name: httpwww.17yy.comyouxigonglue.html
-----------------------
```

```
Hit enter with no input to quit.
Query:小游戏 site:www.17yy.com

Searching for: 小游戏 site:www.17yy.com
{'contents': '小游戏', 'site': 'www.17yy.com'}
5161 total matching documents.
-----------------------
path: html/httpwww.17yy.comyouxigonglueyizhi.html
title: 小游戏攻略_小游戏秘籍 - 17yy经典小游戏
url: http://www.17yy.com/youxigonglue/yizhi
name: httpwww.17yy.comyouxigonglueyizhi.html
```

【练习二：实现一个图片索引】

运行 Indeximg.py 建立图片的索引：

需要注意的是，在这个过程中由于许多网站是纯文字类型的或者有一定的超文本类型的（没有图片），因此在刚开始实验爬取的网站出现大量的找不到内容的网页：

```
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
warning: no content in httpswww.11773.com.html
```

但是，当到了后面有着大量图片的网页的时候就显得相对比较正常：

```
adding httpwww.17yy.comf234343.html.html
adding httpwww.17yy.comf234357.html.html
adding httpwww.17yy.comf234360.html.html
adding httpwww.17yy.comf234362.html.html
adding httpwww.17yy.comf234364.html.html
adding httpwww.17yy.comf234369.html.html
adding httpwww.17yy.comf234373.html.html
adding httpwww.17yy.comf234374.html.html
adding httpwww.17yy.comf234375.html.html
adding httpwww.17yy.comf234376.html.html
adding httpwww.17yy.comf234394.html.html
adding httpwww.17yy.comf234451.html.html
adding httpwww.17yy.comf234452.html.html
adding httpwww.17yy.comf234453.html.html
adding httpwww.17yy.comf234457.html.html
```

并且由于有去重操作，所以程序的运行速度也是比较迅速的：

```
warning: no content in httpwww.yzz.cn.html
warning: no content in httpwww.yzz.cn.html
adding httpxiaoyouxi.2345.com.html
warning: no content in httpxiaoyouxi.2345.com.html
commit index
.done
0:03:22.984238
```

运行 Searchimg.py 进行索引的检索：

检索“闪客快打”出现 77 条项目

```
Hit enter with no input to quit.
Query:闪客快打
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 0.602 seconds.
Prefix dict has been built successfully.

Searching for: 闪客 快 打
77 total matching documents.
imgurl: http://pic1.17yy.com/swf/dongzuo/2010-02-25/6732d8863109c7917888ccaef4c43c35.gif
url: http://www.17yy.com/f/19805.html
闪客快打4

------------------------
imgurl: http://pic1.17yy.com/swf/dongzuo/2014-01-13/aa6b58052432229ce5509984524b1b45.jpg
url: http://www.17yy.com/f/108678.html
闪客快打之无路可退

------------------------
```

检索“金庸群侠传”出现 49 条项目

```
Hit enter with no input to quit.
Query:金庸群侠传

Searching for: 金庸 群侠传
49 total matching documents.
imgurl: http://pic3.17yy.com/swf/zonghe/2014-01-14/6238858c269f2fb2eda80679cd204dec.jpg
url: http://www.17yy.com/f/108747.html
金庸群侠传x0.4

-----------------------
imgurl: http://pic1.17yy.com/swf/xiuxian/2010-01-31/9f2ce7fa703f71c73caed335ce75f087.gif
url: http://www.17yy.com/f/12423.html
金庸群侠传连连看

-----------------------
imgurl: http://pic1.17yy.com/swf/maoxian/2015-06-11/9f0a6968ee0af4608c96b1e3532bc848.jpg
```

六、实验体会

这是本课程的第四次实验作业，可以认为是第三次实验作业的提升版，故基于对第三次实验作业的了解，处理起来并不算难，而且容易上手，极大地提升了我对电工导实验的自信心。

在本次实验中，我再一次加深了对 Field 的应用以及对 doc 的存储方式的理解，从而能够很好地建立域，并且正确选择域的类型，不会像刚开始入手的时候不清楚 t1、t2 的区别而茫然无措。

当然，本次实验也存在不足之处，主要是由于选取的网站图片的网页结构过于简单，因而需要对第四部分的实验问题进行更加深入的思考，思考是否有更一般化的解法。