

电工导实验报告 3

F1403023 5140309534 韩坤言

一、实验目的

1. 了解索引的创建和搜索，中文分词
2. 了解组合查询，更新索引中的文档，简单的图片搜索

二、实验内容

1. 了解 Lucene，索引创建和搜索索引，实现中文的全文检索
2. 组合查询来对搜寻的结果进行限制，更改目录中的文档以及搜索网页中出现的图片。

三、实验环境

1. Firefox + Firebug 插件或 Chrome
2. Python 2.7 + easy_install + BeautifulSoup
3. JCC + PyLucene

四、实验步骤

Lucene 是一个高效的，基于 Java 的中文检索库。全文索引主要分为两个部分，索引的创建和利用索引来进行搜索。原文先被分词，得到的词元传给语言处理组件，最后将一个个 term 传给索引组件，得到倒排索引。而用户查询时的关键词先被分词，创建语法树，进而搜索。对中文而言，先用中文分词软件分好词再用 WhitespaceAnalyzer 或者 SimpleAnalyzer 进行分析是比较方便而且可行的。

例如 Google 等搜索引擎，都能在搜索时加上限定，比如 site:某某网站，更好符合用户的需求。通过使用 BooleanQuery 可以将不同的查询组合成复杂的查询方式。BooleanQuery 本身是一个布尔子句的容器，这个子句可以是逻辑“或”、“与”或“非”。可以用 BooleanQuery 的 add 方法将一个查询子句添加到某个 BooleanQuery 对象中，这样能够优化搜寻的结果，得到更令人满意的搜索结果。图片搜索要依据图片周围的文字来获取关于图片的信息，用这些信息来建立索引，最后依据这进行查找。

五、问题及其解决

1. 实现一个中文网页索引与搜索程序

爬取一定数量(>5k)的中文网页，修改 IndexFiles.py 和 SearchFiles.py，对这些中文网页建立索引。

doc 的 Field 中需要有 name(文件名)，path(文件路径)，title(网页标题)，url(网页地址)，contents(索引的文件内容)

搜索时显示出相关信息

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

```
Searching for: 战争 游戏  
10 total matching documents.  
path: C:/Users/Alpha/Desktop/py/baidu_sim/html/httpwf.qq.com title: 战争前线-WarF  
ace-官方网站-腾讯游戏-孤岛危机系列射击巨作 url: http://wf.qq.com/ name: httpwf.qq.com  
path: C:/Users/Alpha/Desktop/py/baidu_sim/html/httpwww.pcgames.com.cnkzztpcgameG  
OW title: 战争机器PC_战争机器_太平洋游戏网战争机器专题 url: http://www.pcgames.com.cn  
/kzztpcgame/GOW/ name: httpwww.pcgames.com.cnkzztpcgameGOW  
path: C:/Users/Alpha/Desktop/py/baidu_sim/html/httpwww.7k7k.comflash_fl491_1.htm  
title: 战争小游戏_战争小游戏大全_战争小游戏全集_7k7k战争小游戏 - 7k7k小游戏 url: http:  
//www.7k7k.com/flash_fl/491_1.htm name: httpwww.7k7k.comflash_fl491_1.htm
```

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

Searching for: 战争 NOT 游戏
10 total matching documents.
path: C:/Users/Alpha/Desktop/py/baidu_sim/html/httpbaike.baidu.comview14949.htm
title: 越南战争_百度百科 url: http://baike.baidu.com/view/14949.htm name: httpbaik
e.baidu.comview14949.htm
path: C:/Users/Alpha/Desktop/py/baidu_sim/html/httpbaike.baidu.comview67404.htm
title: 普法战争_百度百科 url: http://baike.baidu.com/view/67404.htm name: httpbaik
e.baidu.comview67404.htm
```

原以为这并不是什么难事，但是问题接踵而至。

首先，我发现我之前的 `crawl_thread` 爬如此大量的网站时经常会爬着爬着就突然停了下来，既没有终止程序，也没有任何的报错信息。有时候甚至到 499X 的时候停下来，让人心烦不已。我不得不进一步修改我的程序。考虑到很多文件比如 pdf, doc 这些都不是需要的，而且这些文件一般爬的很慢，我于是将把得到的 url 后缀是这些的都过滤掉了，爬取网页的速度提升了不少。我还加了一些异常捕获，尽量减少发生错误的可能性。

其次是汉字的编码问题，虽然我查阅了大量关于编码转换和机理，但是遇到有的网站还是无能为力，毫无办法，乱码还是乱码，但是大部分的网站都能正常处理，这个问题有待之后解决。

随着网页越爬越多，最后速度会越来越慢，我原以为是 `crawled` 里的元素多了以后，判断在不在 `crawled` 中是占用时间的元凶，但是就算我加上了 hash 来优化，依旧没有好转，到最后 100 至 200 左右速度会大幅下降，反而用了哈希以后，平均速度反而降低了，我还是保留了原来的方法。

爬 5000 个网页大约用时如下：

```
1582.36638843
```

大约 26 分钟，可以接受。

修改 index 的创建程序不是很难，无非改一改路径，我运用了第三方的库 `chardet` 来识别网页的编码方式，把其解码成 `unicode` 的格式，因为 `unicode` 可以跨平台运用，实用最方便，也不容易出错。bs 自带的识别编码方式的函数不太理想，安装了 `chardet` 库以后，遇到不确定的情况还会调用 `chardet` 模块，所以我直接用了 `chardet` 模块。但又是仍然有无法处理的情况，毕竟少数，加个异常捕获跳过就行了。读取爬网页时的 `index.txt`，找到对应 url 的文件名，然后再打开，读取，分词，把 html 的 tag 去除。开始我打算实用 `nltk` 库，但是高版本貌似不支持这个函数了，我于是选择了使用 bs 自带的 `get_text()` 函数。虽然效果不是非常好，但是减少了查询时所不必要的时间。Search 的程序也是相应改动，改动不大。5000 个做成 index 还是花了不少时间。查询时效果不错。

```
|optimizing index... done
|0:38:56.929000
```

用时挺长，或许不把 title print 出来时间会短一些。

```
Python 2.7.10 (default, May 23 2015, 09:40:32) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
lucene 3.6.2

Hit enter with no input to quit.
Query:国家

Searching for: 国家
50 total matching documents.
-----
path: html\httpwww.seiee.sjtu.edu.cn
title: 上海交通大学-电子信息与电气工程学院-电子信息与电气工程学院
url: http://www.seiee.sjtu.edu.cn/
name: httpwww.seiee.sjtu.edu.cn
-----
path: html\http120.sjtu.edu.cnWebShoww35p7f1769
title: 上海交通大学建校120周年
url: http://120.sjtu.edu.cn/Web/Show?w=35&p=7&f=1769
name: http120.sjtu.edu.cnWebShoww35p7f1769
-----
path: html\httpwww.china-language.gov.cn322015_9_71_32_6090_0_1441609177858.html
title: 中国语言文字网
url: http://www.china-language.gov.cn/32/2015_9_7/1_32_6090_0_1441609177858.html
name: httpwww.china-language.gov.cn322015_9_71_32_6090_0_1441609177858.html
-----
path: html\httpwww.shjbzx.cnjbptnode132904ulai594.html
title: 上海市互联网举报平台
url: http://www.shjbzx.cn/jbpt/node132904/ulai594.html
name: httpwww.shjbzx.cnjbptnode132904ulai594.html
-----
path: html\httpwww.china-language.gov.cn7index.htm
title: 中国语言文字网
url: http://www.china-language.gov.cn/7/index.htm
name: httpwww.china-language.gov.cn7index.htm
-----
path: html\httpwww.shjbzx.cnjbptnode131566ulai687.html
title: 上海市互联网举报平台
url: http://www.shjbzx.cn/jbpt/node131566/ulai687.html
name: httpwww.shjbzx.cnjbptnode131566ulai687.html
```

测试图片

2. 模拟实现搜索引擎的“site:”功能（对搜索的网站进行限制）

```
Hit enter with no input to quit.      Hit enter with no input to quit.
Query:国家 site:sina.com.cn           Query:国家 site:baike.baidu.com

Searching for: 国家 site:sina.com.cn   Searching for: 国家 site:baike.baid
10 total matching documents.           10 total matching documents.
-----
path: html\httpmatch.2012.sina.com.cn  path: html\httpbaike.baidu.comview
untry_t00001                           title: 美国_百度百科
title: 法国奖牌榜_2012伦敦奥运会_新浪网 url: http://baike.baidu.com/view/2
url: http://match.2012.sina.com.cn/m_   name: httpbaike.baidu.comview2398.1
_country_t00001                        -----
name: httpmatch.2012.sina.com.cnmeda_   path: html\httpbaike.baidu.comview
-----                                title: 英国_百度百科
path: html\httpmatch.2012.sina.com.cn  url: http://baike.baidu.com/view/3
untry_t00001                           name: httpbaike.baidu.comview3565.1
title: 韩国奖牌榜_2012伦敦奥运会_新浪网 -----
url: http://match.2012.sina.com.cn/m_   path: html\httpbaike.baidu.comview
_country_t00001                        -----
name: httpmatch.2012.sina.com.cnmeda_
```

开始我是用 tld 的第三方模块来得到一个 url 的域名,把域名加到 index 里之后进行匹配。但是这样得到的域名只是顶级的域名,遇过搜索的时候用下级的域名变无法搜索到。后来我修改 search 的程序

```
querys = BooleanQuery()
for k,v in command_dict.iteritems():
    if (k=='site'):
        t = Term('url', '*' + v + '*')
        query = WildcardQuery(t)
        querys.add(query, BooleanClause.Occur.MUST)
    else:
        query = QueryParser(Version.LUCENE_CURRENT, k, analyzer).parse(v)
        querys.add(query, BooleanClause.Occur.MUST)
scoreDocs = searcher.search(querys, 50).scoreDocs
print "%s total matching documents." % len(scoreDocs)
```

通过 WildcardQuery 可以部分来进行匹配,顺利解决了分级域名无法匹配的问题。

我的测试:

```
Python 2.7.10 (default, May 23 2015, 09:40:32) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
lucene 3.6.2

Hit enter with no input to quit.
Query:餐餐

Searching for: 餐餐
12 total matching documents.
-----
path: html\httpyouth.sjtu.edu.cnplusview.phpaid6937
title: 团旗飘飘
url: http://youth.sjtu.edu.cn/plus/view.php?aid=6937
name: httpyouth.sjtu.edu.cnplusview.phpaid6937
-----
path: html\http120.sjtu.edu.cnWebShoww29p9f1724
title: 上海交通大学建校120周年
url: http://120.sjtu.edu.cn/Web/Show?w=29&p=9&f=1724
name: http120.sjtu.edu.cnWebShoww29p9f1724
-----
path: html\http120.sjtu.edu.cnWebShoww42p8f1517
title: 上海交通大学建校120周年
url: http://120.sjtu.edu.cn/Web/Show?w=42&p=8&f=1517
name: http120.sjtu.edu.cnWebShoww42p8f1517
-----
path: html\http120.sjtu.edu.cnWebShoww30p9f1459
title: 上海交通大学建校120周年
url: http://120.sjtu.edu.cn/Web/Show?w=30&p=9&f=1459
name: http120.sjtu.edu.cnWebShoww30p9f1459
-----
path: html\httpwww.sjtu.edu.cnjdwmmwcontent.jspurltypetree.TreeTempUrlwbtreetid1453
title: IV-20-4-上海交通大学
url: http://www.sjtu.edu.cn/jdwmmw/content.jsp?urltype=tree.TreeTempUrl&wbtreetid=1453
name: httpwww.sjtu.edu.cnjdwmmwcontent.jspurltypetree.TreeTempUrlwbtreetid1453
-----
path: html\httpmy.xinhuanet.com
title: 马来西亚频道-新华网
url: http://my.xinhuanet.com/
name: httpmy.xinhuanet.com
```

选了一个比较少见的词，全网站搜索有 12 个符合的

```

Hit enter with no input to quit.
Query:餐餐 site:sjtu.edu.cn

Searching for: 餐餐 site:sjtu.edu.cn
5 total matching documents.
-----
path: html\httpyouth.sjtu.edu.cnplusview.phpaid6937
title: 团旗飘飘
url: http://youth.sjtu.edu.cn/plus/view.php?aid=6937
name: httpyouth.sjtu.edu.cnplusview.phpaid6937
-----
path: html\http120.sjtu.edu.cnWebShoww29p9f1724
title: 上海交通大学建校120周年
url: http://120.sjtu.edu.cn/Web/Show?w=29&p=9&f=1724
name: http120.sjtu.edu.cnWebShoww29p9f1724
-----
path: html\http120.sjtu.edu.cnWebShoww42p8f1517
title: 上海交通大学建校120周年
url: http://120.sjtu.edu.cn/Web/Show?w=42&p=8&f=1517
name: http120.sjtu.edu.cnWebShoww42p8f1517
-----
path: html\http120.sjtu.edu.cnWebShoww30p9f1459
title: 上海交通大学建校120周年
url: http://120.sjtu.edu.cn/Web/Show?w=30&p=9&f=1459
name: http120.sjtu.edu.cnWebShoww30p9f1459
-----
path: html\httpwww.sjtu.edu.cnjdwmmwcontent.jspurltypetree.TreeTempUrlwbtreeid1453
title: IV-20-4-上海交通大学
url: http://www.sjtu.edu.cn/jdwmmw/content.jsp?urltype=tree.TreeTempUrl&wbtreeid=1453
name: httpwww.sjtu.edu.cnjdwmmwcontent.jspurltypetree.TreeTempUrlwbtreeid1453

```

限定了 sjtu 后就只有 5 个了，site 的限制颇有成效。

3. 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。

示例：

```

Hit enter with no input to quit.
Query:男装
Searching for: 男装
10 total matching documents.
imgurl: http://img01.taobaocdn.com/
url: http://list.taobao.com/market/
d=all&atype=b&style=grid&ppath=1404
urlltitle:
薄款-夹克-男装-淘宝网

-----
imgurl: http://img01.taobaocdn.com/
url: http://list.taobao.com/market/
ype=0&random=false&viewIndex=1&yp4g
urlltitle:
中老年专区-男装-淘宝网

```

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网站爬取。比如只对淘宝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。

我是选择京东，因为我发现爬京东比淘宝快多了。首先要修改 crawler，因为我需要爬取那些商品的界面，并不需要别的没有用的网页。

我添加了正则匹配，因为商品的 url 都有共同的部分

```
def get_all_links(content, page):
    link = []
    soup = BeautifulSoup(content)
    for i in soup.findAll('a', {'href': re.compile('.*//item.jd.com/.html$') }):
        url = i.get('href', '')
        link.append(urlparse.urljoin(page, url))
    #print len(link)
    return link
```

图片好找，但是我还需要关于图片的描述用来匹配搜索。我剖析了京东商品页面的 html 树



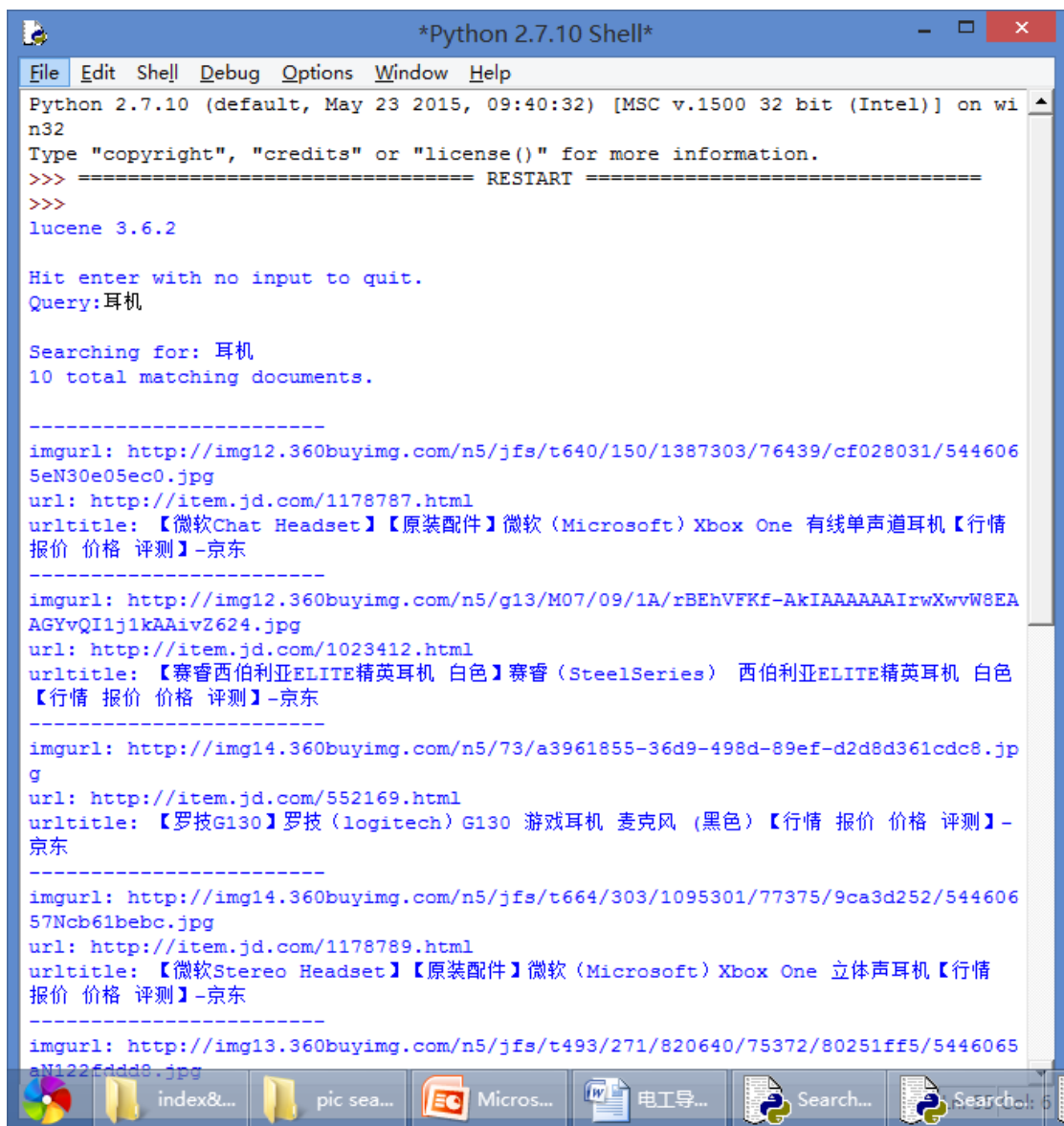
主要相关的图片和描述都在这个叫 p-box 的 tag 里，虽然还有很多商品的图片，那些图片反正爬取的时候会爬到是它们的界面，所以就不处理了。下面商品介绍虽然有很多图片，但没有相应的描述，而且图片相关性参差不齐，所以就算了，我主要把左上角一副大图和下面数量不等的小图存下来，这些都是相关性极高的图片，也有详尽的描述。

这部分代码我是这样处理的

```
collection = [] #存放imgurl和对应的discription
dic = {}
p_box = soup.find(id='p-box') #处理一开始左上角一大图和下面几张小图
#print p_box.get('id', '')
sub_p_box = p_box.div.nextSibling.nextSibling.nextSibling.nextSibling
#print sub_p_box.get('class', '')
#print sub_p_box
big_pic = sub_p_box.div.div.div.img
dic['imgurl'] = urlparse.urljoin(url, big_pic.get('src', ''))
dic['discription'] = big_pic.get('alt', '')
#print dic
collection.append(dic)

small_pic_group = big_pic.parent.nextSibling.nextSibling.div.ul
for i in small_pic_group.findAll('li'):
    small_pic = i.img
    dic['imgurl'] = urlparse.urljoin(url, small_pic.get('src', ''))
    dic['discription'] = " ".join(jieba.cut(small_pic.get('alt', '')))
    #print dic
    collection.append(dic)
```


由于京东商品的页面结构是固定的，这样能大幅减少出错的概率，尝试很多次基本没有问题。做成 index 时描述是要分析的，别的信息只要存储即可。



```
*Python 2.7.10 Shell*
File Edit Shell Debug Options Window Help
Python 2.7.10 (default, May 23 2015, 09:40:32) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
lucene 3.6.2

Hit enter with no input to quit.
Query:耳机

Searching for: 耳机
10 total matching documents.

-----
imgurl: http://img12.360buyimg.com/n5/jfs/t640/150/1387303/76439/cf028031/5446065eN30e05ec0.jpg
url: http://item.jd.com/1178787.html
urltitle: 【微软Chat Headset】【原装配件】微软 (Microsoft) Xbox One 有线单声道耳机【行情 报价 价格 评测】-京东
-----
imgurl: http://img12.360buyimg.com/n5/g13/M07/09/1A/rBEhVFKf-AkIAAAAAAIrwXwvW8EAGYvQI1j1kAAivZ624.jpg
url: http://item.jd.com/1023412.html
urltitle: 【赛睿西伯利亚ELITE精英耳机 白色】赛睿 (SteelSeries) 西伯利亚ELITE精英耳机 白色【行情 报价 价格 评测】-京东
-----
imgurl: http://img14.360buyimg.com/n5/73/a3961855-36d9-498d-89ef-d2d8d361cdc8.jpg
url: http://item.jd.com/552169.html
urltitle: 【罗技G130】罗技 (logitech) G130 游戏耳机 麦克风 (黑色)【行情 报价 价格 评测】-京东
-----
imgurl: http://img14.360buyimg.com/n5/jfs/t664/303/1095301/77375/9ca3d252/54460657Ncb61bebc.jpg
url: http://item.jd.com/1178789.html
urltitle: 【微软Stereo Headset】【原装配件】微软 (Microsoft) Xbox One 立体声耳机【行情 报价 价格 评测】-京东
-----
imgurl: http://img13.360buyimg.com/n5/jfs/t493/271/820640/75372/80251ff5/5446065aN122fddd8.jpg
```

小样本进行测试，结果令人满意，如果样本再大一些，产品更加多样的话，效果会更好（这些网页都和数码产品相关，所以大部分都是这些商品）

六、实验总结

这次在上次作业的基础上，难度提升不少，但是实用性也越来越有所提升。index 的建立和关键词的搜索大大提升了搜索的效率，也能对大量爬到的网站进行处理，为我所用（否则爬了一堆网站什么都做不了并没有什么用）。之后的 site 的限制来是搜索结果更能让人满意，以及对图片的搜索，都有一种一步步在做搜索引擎的感觉，总而言之，通过这两周的学习让我受益匪浅，对此也有了兴趣。