# 第四次实验（2）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：Windows 10
语言：Python 2
IDE：Pycharm 2019.2.3

### 2、 实验目的

2.1　了解组合查询
2.2　模拟实现搜索引擎的“site:”功能
2.3　实现一个图片索引

### 3、实验思路

3.1　重新修改 IndexFiles.py，使得创建索引时也会添加一个名为 site 的 field，同时修改一些 field 的属性

3.2　重新修改 SearchFiles.py，使得输出结果符合题目要求，并且运用组合查询并加以判断来模拟实现搜索引擎的“site:”功能

3.3 运用相关函数获得图片周围的文本，并以此新建索引来达到搜索图片的目的

## 二、实验过程

### 1、练习一 模拟实现搜索引擎的“site:”功能

#### 1.1 修改 IndexFiles.py

在练习一中我们依然使用上一次实验中爬取的相关网站的内容。但不同的是，我们需要对创建索引的部分的代码加以修改来满足对于 site 的搜索。

site 的获取需要借助 urlparse 库内的 urlparse 类。我们可以建立这个类的对象，并且由它的 netloc 属性获取服务器的地址，也就是我们在此需要的 site。

IndexFiles.py 中除了 field 添加这一部分，余下的代码与上一部分实验差异不大。

这一段修改的代码如下：

```
for line in index_file.readlines():
    url_and_name = line.split()
    url = url_and_name[0]
    filename = url_and_name[1]
    parsed = urlparse(url)
    site = parsed.netloc
    print "site:",site
    print "adding", filename
    try:
        path = os.path.join("html", filename)
        file = open(path)
        contents = file.read()
        soup = BeautifulSoup(contents, features="html.parser")
        title = soup.head.title.string
        contents = ''.join(soup.findAll(text=True))
        seg_list = jieba.cut(contents)
        contents = " ".join(seg_list)
        file.close()
        doc = Document()
        doc.add(Field("title", title,
                      Field.Store.YES,
                      Field.Index.NOT_ANALYZED))
        doc.add(Field("site", site,
                      Field.Store.YES,
                      Field.Index.NOT_ANALYZED))
        doc.add(Field("name", filename,
                      Field.Store.YES,
```

### 1.2 修改 SearchFiles.py

对于上机实验 5 中提供的 SearchFiles.py，已经为我们提供了组合搜索的大体框架。而对于本练习，我们需要的仅仅是将限定词改为 site，并且将输入的查询语句加入中文分词的功能，总体而言比较简单。

我们修改的代码部分如下：

```
allowed_opt = ['site']
command_dict = {}
opt = 'contents'
for i in command.split(' '):
    if ':' in i:
        opt, value = i.split(':')[:2]
        opt = opt.lower()
        if opt in allowed_opt and value != '':
            command_dict[opt] = command_dict.get(opt, '') + ' ' + value
    else:
        command = ' '.join(jieba.cut(command))
        command_dict[opt] = command_dict.get(opt, '') + ' ' + i
return command_dict
```

当然，我们还需要对于最终的输出格式加以修改，使其符合输出样例。

```
for scoreDoc in scoreDocs:
    doc = searcher.doc(scoreDoc.doc)
    ##             explanation = searcher.explain(query, scoreDoc.doc)
    print "------------------------"
    print 'path:', doc.get("path")
    print 'title:', doc.get('title')
    print 'url:', doc.get('url')
    print 'name:', doc.get("name")
```

### 1.3 代码运行结果

在本次实验中，我们选择进行“contents”与“site”的组合搜索，为了便于分辨搜索效果，我们以“民国”作为关键词，分别查看不限制 site 与限制 site 的不同情况下的结果。

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

Searching for: 民国
13 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
------------------------
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
------------------------
path: html/httpwww.sh.chinanews.com.cnshishang2017-12-1532903.shtml
title: “上海南市难民区纪念碑”在沪落成并揭幕-中新社上海
url: http://www.sh.chinanews.com.cn/shishang/2017-12-15/32903.shtml
name: httpwww.sh.chinanews.com.cnshishang2017-12-1532903.shtml
------------------------
path: html/httpwww.sh.chinanews.com.cngatq2018-08-0142827.shtml
title: 第六届台湾“单车天使”公益活动上海站启动-中新社上海
url: http://www.sh.chinanews.com.cn/gatq/2018-08-01/42827.shtml
name: httpwww.sh.chinanews.com.cngatq2018-08-0142827.shtml
------------------------
path: html/httpnews.sjtu.edu.cnztzl_lmfc2018041471086.html
title: 何友声：正身律物、轻利重德，学贵致用、以勤补拙_劳模风采_上海交通大学新闻
学术网
url: http://news.sjtu.edu.cn/ztzl_lmfc/20180414/71086.html
name: httpnews.sjtu.edu.cnztzl_lmfc2018041471086.html
```

```
wenxuanliu@ubuntu: ~/Desktop
Hit enter with no input to quit.
Query:民国 site:www.sh.chinanews.com.cn

Searching for: 民国 site:www.sh.chinanews.com.cn
11 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
-----------------------
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
-----------------------
path: html/httpwww.sh.chinanews.com.cnshishang2017-12-1532903.shtml
title: "上海南市难民区纪念碑"在沪落成并揭幕-中新社上海
url: http://www.sh.chinanews.com.cn/shishang/2017-12-15/32903.shtml
name: httpwww.sh.chinanews.com.cnshishang2017-12-1532903.shtml
-----------------------
path: html/httpwww.sh.chinanews.com.cngatq2018-08-0142827.shtml
title: 第六届台湾"单车天使"公益活动上海站启动-中新社上海
url: http://www.sh.chinanews.com.cn/gatq/2018-08-01/42827.shtml
name: httpwww.sh.chinanews.com.cngatq2018-08-0142827.shtml
-----------------------
path: html/httpwww.sh.chinanews.com.cnshishang2019-05-2457155.shtml
title: 90后法国美女来华创业，交中国男友、跳钢管舞、爱老上海-中新社上海
url: http://www.sh.chinanews.com.cn/shishang/2019-05-24/57155.shtml
name: httpwww.sh.chinanews.com.cnshishang2019-05-2457155.shtml
-----------------------
```

我们可以看到，限制 site 后搜索得到的结果变少，并且它们的 site 都与我们指定的 site 相同，而不相同的那些网站则不呈现在搜索结果中，可见实验实现效果较好。

## 2、练习二 实现图片索引

### 2.1 设计 IndexFilesImg.py

图片索引的本质依然是对于图片周边文字的索引，故而 IndexFilesImg.py 和 IndexFiles.py 的主体基本差别不大。

我们首先设计获取图片相关文本的代码，由于图片周边文本的出现方式是数不胜数的，我们在此选择一些比较主流的情况：

```
def getinfo(img):
    try:
        contents = img['alt']
    except:
        try:
            contents = img.parent['title']
        except:
            try:
                contents = img.parent.nextSibling.h4.string
            except:
                return None
    return contents
```

由于我们要向 doc 中加入 imgurl，所以在添加 field 部分的代码我们进行如下修改，利

用 html 的“src”标签来获取图片的链接。而且很多网站的图片地址以相对路径保存，我们记得将其改写为绝对路径。

```
try:
    path = os.path.join("html", filename)
    file = open(path)
    contents = file.read()
    soup = BeautifulSoup(contents, features="html.parser")
    imgs = soup.find_all('img')
    title = soup.head.title.string
    file.close()
    for img in imgs:
        imgurl = img.get('src')
        imgurl = urlparse.urljoin(url,imgurl)
        contents = getinfo(img)
        if imgurl not in crawled_imgurls:
            crawled_imgurls.append(imgurl)
            if contents:
                contents_list = jieba.cut(contents)
                contents = ' '.join(contents_list)
            doc = Document()
            doc.add(Field("imgurl", imgurl,
                          Field.Store.YES,
                          Field.Index.NOT_ANALYZED))
            doc.add(Field("title", title,
                          Field.Store.YES,
                          Field.Index.NOT_ANALYZED))
            doc.add(Field("url", url,
                          Field.Store.YES,
                          Field.Index.NOT_ANALYZED))
```

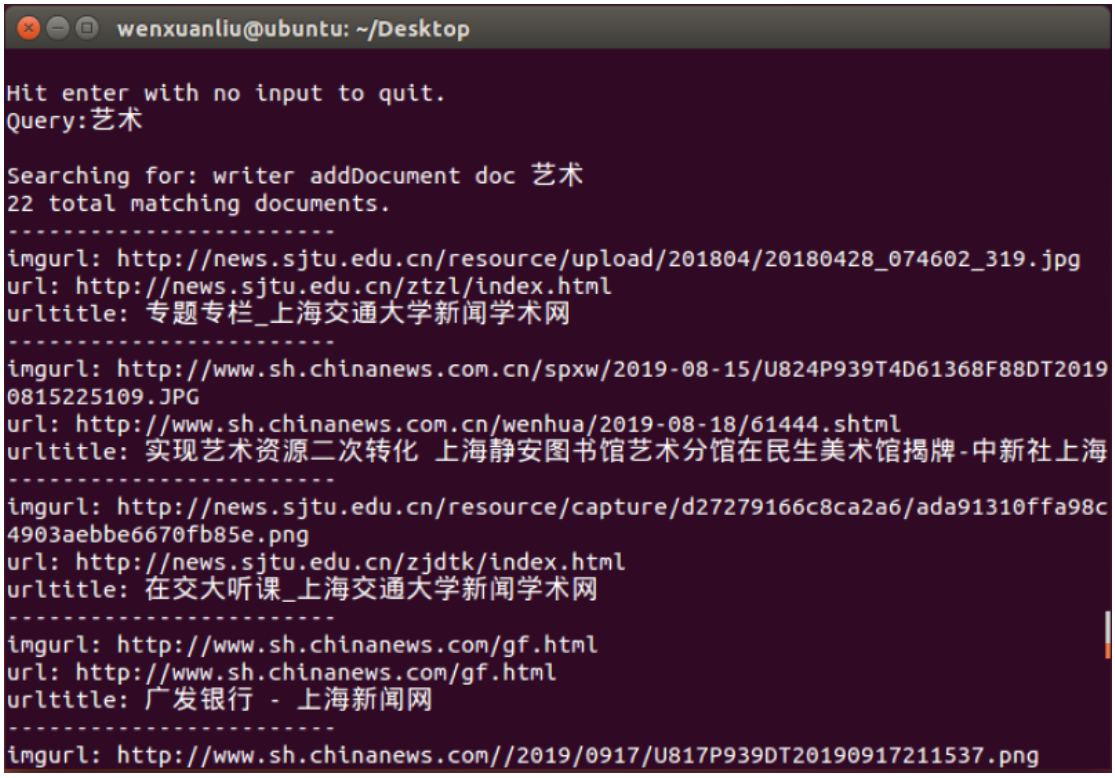
### 2.2 设计 SearchFilesImg.py

SearchFilesImg.py 的代码更是和 SearchFiles.py 相差无几，注意修改一下 main 函数里的索引文件夹为 indeximg，并且修改一下输出格式就好。

```
for scoreDoc in scoreDocs:
    doc = searcher.doc(scoreDoc.doc)
    ##            explanation = searcher.explain(query, scoreDoc.doc)
    print "------------------------"
    print 'imgurl:', doc.get('imgurl')
    print 'url:', doc.get('url')
    print 'urltitle:', doc.get('title')
```

### 2.3 代码运行结果

在此我们依然运用上一次实验爬取到的文件，在建立好了图片索引后，我们搜索“艺术”，得到的结果如图所示，还是很恰当的。




## 三、实验总结

### 1、实验概述

本次实验的主要任务，可以总结为模拟实现搜索引擎的“site:”功能,实现图片索引两个实验，两道练习分别从不同的角度，让我们熟悉了相关的代码设计。

### 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 学会了组合查询的相关方法

2.2 学会了更新以及修改索引中的文档

2.3 懂得了如何实现图片索引

### 3、创新点

首先是对于网页的 site 信息的获取，没有使用蛮力，而是用了 urlparse(url).netloc 来处理，代码简洁而准确。

其次是在图片索引建立中，由于同一张图片可能会出现许多次，我额外加入了判断来避免大量重复，同时我还将相对路径改写为绝对路径来避免误会：

```
for img in imgs:
    imgurl = img.get('src')
    imgurl = urlparse.urljoin(url,imgurl)
    contents = getinfo(img)
    if imgurl not in crawled_imgurls:
        crawled_imgurls.append(imgurl)
        if contents:
            contents_list = jieba.cut(contents)
            contents = ' '.join(contents_list)
        doc = Document()
        doc.add(Field("imgurl", imgurl,
                      Field.Store.YES,
                      Field.Index.NOT_ANALYZED))
        doc.add(Field("title", title,
                      Field.Store.YES,
                      Field.Index.NOT_ANALYZED))
        doc.add(Field("url", url,
                      Field.Store.YES,
                      Field.Index.NOT_ANALYZED))
```

同时通过对 urlparse.urljoin()函数和 re.compile().match()的使用，我掌握了一种更加完整和更加简洁的将各种相对路径改写为绝对路径，去除无关数据的干扰的方式。

```
for i in soup.findAll('a'):
    temp = urlparse.urljoin(page, i.get('href'))
    if re.compile('^http').match(temp):
        links.append(temp)
```

同时在获取链接时，考虑到了很多不合规的情形，并且通过正则表达式，re.match()这个方法将其一一过滤或者修改，具体的情形已经在上面进行了详细的分析。

**4、遇到的问题**

在练习一中，原代码对输入的 command 调用 unicode 时，属性为“GBK”，会出现乱码，我们将属性改为“utf-8”就可以解决这一问题。

```
Hit enter with no input to quit.
Query:民国

Searching for: 姘戝浗
```