第五周实验报告

1. 实验概览

本实验旨在进一步理解索引中“域”的概念，实现浏览器中的组合搜索（“site:”）；理解 Luence 索引的更新和删除；理解图片搜索的过程，基于 Luence 库和 jieba 中文分词库，实现组合搜索和图片索引。

2. 实验环境

Docker: sjtumic/ee208

3. 解决思路

组合搜索：

索引建立过程整体类似于 lab4，添加 site 域提取网站的域名

```
contents = re.sub("[^\u4e00-\u9fa5]", "", contents)
site = urls[i].split('/')[2]
site = ' '.join(site.split('.'))

doc = Document()
doc.add(Field("name", filename, t1))
doc.add(Field("path", path, t1))
doc.add(Field("title", title, t1))
doc.add(Field("url", urls[i], t1))
doc.add(Field("site", site, t2))
```

查找的过程中加入代码按照格式分析输入命令

```
def parseCommand(command):
    '''
    input: C title:T author:A language:L
    output: {'contents':C, 'title':T, 'author':A, 'language':L}

    Sample:
    input:'contenance title:henri language:french author:william shakespeare'
    output:{'author': ' william shakespeare',
                    'language': ' french',
                    'contents': ' contenance',
                    'title': ' henri'}
    '''
    allowed_opt = ['title', 'name', 'site', 'url']
    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 != '':
                if opt == 'site':
                    value = value.split('.')
                    value = ' AND '.join(value)
                command_dict[opt] = (command_dict.get(opt, '') + ' ' + value).strip()
        else:
            command_dict[opt] = command_dict.get(opt, '') + ' ' + i
    return command_dict
```

图片索引：

实验以网易新闻作为根节点，根据对于网页的分析，得到图片和对应标题的结构；基于 bs4 中的 parent、next_sibling 等针对树结构的分析，得到图片和对应的文字描述。基于这些信息创建索引，实现搜索

```
spans = soup.findAll("span", {"class": "bg"})
body = soup.find("div", {"class": "post_body"})
if body != None:
    imgs = soup.findAll("img")
    for img in imgs:
        src = img.get("src")
        title = img.get("alt")

        if title == None:
            try:
                title = img.parent.find("a").get("title")
            except:
                continue

        title = " ".join(jieba.cut(title, cut_all=True))

        doc = Document()
        doc.add(Field("contents", title, t2))
        doc.add(Field("imgurl", src, t1))
        doc.add(Field("url", urls[i], t1))
        doc.add(Field("title", site_title, t1))
        writer.addDocument(doc)

        print(src, title, urls[i])
        print("------------------------------")
else:
    for span in spans:
        try:
            hyper = span.parent
            img = hyper.find("img")
            if img.get("alt") != None:
                title = img.get("alt")
            else:
                if hyper.find("h3") != None:
                    if hyper.find("h3").find("a") == None:
                        title = hyper.find("h3").string
                    else:
                        title = hyper.find("h3").find("a").string
                elif hyper.find("h2") != None:
                    if hyper.find("h2").find("a") == None:
                        title = hyper.find("h2").string
                    else:
```

```
    src = img.get("src")
    if src == None:
        src = img.get("data-original")
    if src == None:
        src = img.get("lazy-src")

    title = " ".join(jieba.cut(title, cut_all=True))
    doc = Document()
    doc.add(Field("contents", title, t2))
    doc.add(Field("imgurl", src, t1))
    doc.add(Field("url", urls[i], t1))
    doc.add(Field("title", site_title, t1))
    writer.addDocument(doc)
    print(src, title, urls[i])
    print("------------------------------")
except:
    continue
```

4. 代码运行结果

组合搜索：

```
Hit enter with no input to quit.
Query:体育 site:sports.163.com

Searching for: 体育 site:sports.163.com
10 total matching documents.
-----------------------
name: httpssports.163.comallsports
path: html/httpssports.163.comallsports
title: 综合体育_网易体育
url: 'https://sports.163.com/allsports'
-----------------------
name: httpssports.163.comyj
path: html/httpssports.163.comyj
title: 意甲|意甲联赛|意甲新闻|意甲直播|意甲转会_网易体育
url: 'https://sports.163.com/yj'
-----------------------
name: httpssports.163.comxj
path: html/httpssports.163.comxj
title: 西甲|西甲联赛|西甲新闻|西甲直播|西甲转会_网易体育
url: 'https://sports.163.com/xj'
-----------------------
name: httpssports.163.comyc
path: html/httpssports.163.comyc
title: 英超|英超联赛|英超新闻|英超直播|英超转会_网易体育
url: 'https://sports.163.com/yc'
-----------------------
name: httpssports.163.comworld
path: html/httpssports.163.comworld
title: 国际足球|国际足球新闻|国际足球转会|国际足球明星_网易体育
url: 'https://sports.163.com/world'
```

图片索引：

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

Searching for: 足球
50 total matching documents.
----------------------
imgurl: https://nimg.ws.126.net/?url=http://bjnewsrec-cv.ws.126.net/three559ef27908ej00s2o9pe000wd000hs00bcp.jpg&thumbnail=140y88&quality=80&type=jpg
url: 'https://www.163.com/sports/article/IH865IPQ00059A7T.html'
title: 中国足协大换血:总局空降监管 选帅足协或说了不算|李颖川|中国足球|孙雯_网易体育
----------------------
imgurl: https://nimg.ws.126.net/?url=http://cms-bucket.ws.126.net/2023/1018/670f1f7cp00s2pf1a0019c0009c0070c.png&thumbnail=140y88&quality=80&type=jpg
url: 'https://www.163.com/sports/article/IH865IPQ00059A7T.html'
title: 中国足协大换血:总局空降监管 选帅足协或说了不算|李颖川|中国足球|孙雯_网易体育
----------------------
imgurl: https://nimg.ws.126.net/?url=http://dingyue.ws.126.net/2023/1014/56ad4110j00s2iqte008gd000sw00lod.jpg&thumbnail=140y88&quality=80&type=jpg
url: 'https://www.163.com/jiankang/article/IGEBF5TV00388045.html'
title: 玉米须煮水降血糖？吃腐乳会致癌？假的|桑葚|解酒_网易健康
----------------------
imgurl: https://nimg.ws.126.net/?url=http://dingyue.ws.126.net/2023/1017/c5ff49baj00s2oeqs000kc000hh00asm.jpg&thumbnail=140y88&quality=80&type=jpg
url: 'https://www.163.com/sports/article/IGUF2SF800058781.html'
title: 看台上的“孤勇者”，8年后带队夺冠了|圣克鲁斯|足球|桑托斯|格雷米奥_网易体育
----------------------
imgurl: https://nimg.ws.126.net/?url=http%3A%2F%2Fcms-bucket.ws.126.net%2F2023%2F1017%2F11470db7j00s2oc8z000oc000cl0069c.jpg&thumbnail=453y225&quality=100&type=jpg
url: 'https://www.163.com/'
title: 网易
----------------------
imgurl: https://nimg.ws.126.net/?url=http%3A%2F%2Fcms-bucket.ws.126.net%2F2023%2F1017%2F11470db7j00s2oc8z000oc000cl0069c.jpg&thumbnail=453y225&quality=100&type=jpg
url: 'https://www.163.com/#f=topnav'
title: 网易
----------------------
imgurl: https://nimg.ws.126.net/?url=http://dingyue.ws.126.net/2023/1018/ff4c9efbj00s2per9001dc000l200dmm.jpg&thumbnail=140y88&quality=80&type=jpg
url: 'https://www.163.com/sports/article/IHB35QJQ00058781.html'
title: 世预赛-梅西双响独享南美射手王 阿根廷2-0四连胜|里奥梅西|阿尔瓦雷斯|格雷罗|冈萨雷斯|罗梅罗_网易体育
----------------------
```