实验五报告

唐灵
519030910052
F1903002

2020 年 10 月 29 日

摘要

这是电工导 c 课程的第四次实验

1 实验概览

本次实验是对于爬虫与文本检索的一个综合简单应用，重点在于应用 lucene 建立索引实现简单的组合搜索，以及图像搜索。

2 实验环境

本次实验并未在课程方统一给定的“ee208” *Docker* 容器中运行并实现。

本次实验由于环境问题, 没有办法在个人电脑上运行 indexfile 文件，上次实验选择在别的同学电脑上完成程序的编写，并迟交了一段时间，而这次实验，可能由于学业繁忙，没有找到合适的同学。

所以最终选择在无参考方式下阅读周毓 519030910049 同学的代码，并独立完成报告，对于文本检索的实际应用进行学习。

3 练习解决思路

3.1 练习一的解决思路

针对爬虫模块，解决思路与我个人的实验四不同，采用了在爬取的过程中，固定的将其特定属性写入 html 文档中，也就是事先进行信息处理和分析，将文本信息，开头，url，文件名等等信息按照一定的顺序存储在开头，在建立索引的过程中可以直接使用。

针对建立索引模块，为了能够事先组合搜索，我们需要将 url，title 等等信息预先进行分词，方便之后的组合查询。

```
doc = Document()
doc.add(StringField("name", filename, Field.Store.YES))
doc.add(StringField("path", path, Field.Store.YES))
if len(contents) > 0:
    doc.add(TextField('title', title, Field.Store.YES))
    doc.add(TextField('site', page, Field.Store.YES))
    doc.add(TextField('url',url,Field.Store.YES))
    doc.add(TextField('contents', contents, Field.Store.YES))
```

图 1: 加入字段

针对与搜索模块，我们对于输入的字段进行自定义的处理，构造字典，使得其能够满足布尔查询的方法。

```
def divide_by_jieba(content):
    content = content.split()
    command = [] #返回的命令
    for i in range(len(content)):
        s = jieba.lcut_for_search(content[i])
        if s[0]==content[i]:
            command.append(s[0])
            continue
        command.append(''.join(['(',' AND '.join(s),')']))

    return ' '.join(command)

def divide_site(page):
    seg_url = jieba.lcut_for_search(page)
    return ' AND '.join(list(set(seg_url)-set(['.','http','https','/',':','?','=',''])))

def parseCommand(command):
    #解析命令行，形成一个字典
    allowed_opt = ['site']
    command_dict = {}
    opt = 'contents'
    #以空格分割
    for i in command.split(' '):
        #对于一个小结构，以：为分割
        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, '') + ' ' + divide_site(value)
        else:
            command_dict[opt] = command_dict.get(opt, '') + ' ' + divide_by_jieba(i)
    return command_dict
```

图 2: 对于搜索字段进行处理

3.2 练习二的解决思路

对于练习二的解决思路和一般的检索思路相同。

关键点在于爬取的时候，从图片的周围抽取图片信息来进行字段的搜索存储，用于字段的搜索。

4 代码运行结果

4.1 练习一的运行结果

图 3: 对于域名为 sina 的搜索结果

图 4: 对于域名为 guancha 的搜索结果

4.2 练习二的运行结果


Searching for: 东风
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 0.723 seconds.
Prefix dict has been built successfully.
11 total matching documents.
img_url: https://i.guancha.cn/news/mainland/2020/10/21/20201021154449230.jpg title: 台媒炒作东风-17部署东南沿海，解放军高调"插播"东风-17受阅照
url: https://www.guancha.cn/politics/2020_10_21_568801.shtml
img_url: https://i.guancha.cn/news/mainland/2020/10/21/20201021154411557.jpg title: 台媒炒作东风-17部署东南沿海，解放军高调"插播"东风-17受阅照
url: https://www.guancha.cn/politics/2020_10_21_568801.shtml
img_url: https://i.guancha.cn/news/social/2020/10/18/20201018140857678.png title: 《南华早报》称东风17部署东南沿海，台媒慌了
url: https://www.guancha.cn/military-affairs/2020_10_18_568451.shtml
img_url: https://i.guancha.cn/news/social/2020/10/18/20201018142615919.png title: 《南华早报》称东风17部署东南沿海，台媒慌了
url: https://www.guancha.cn/military-affairs/2020_10_18_568451.shtml
img_url: https://i.guancha.cn/news/social/2020/10/18/20201018143315608.png title: 《南华早报》称东风17部署东南沿海，台媒慌了
url: https://www.guancha.cn/military-affairs/2020_10_18_568451.shtml
img_url: https://i.guancha.cn/news/hmt/2020/10/21/20201021152718382.jpg title: 台媒炒作东风-17部署东南沿海，解放军高调"插播"东风-17受阅照
url: https://www.guancha.cn/politics/2020_10_21_568801.shtml

图 5: 搜索图片结果