实验四报告

唐灵
519030910052
F1903002

2020 年 10 月 17 日

摘要

这是电工导 c 课程的第三次实验

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 建立索引并进行简单的搜索。

2 实验环境

本次实验的所有代码在电类工程导论 c 课程中在课程方统一给定的 “ee208” *Docker* 容器中运行并实现。

3 练习解决思路

3.1 爬虫的解决思路

爬虫的的话，直接采用 lab2 中的代码，需要注意的是关于网页编码的问题，一定要查看过后进行修改再进行爬取。由于我采用的是在建立索引的时候进行文档处理，所以这一步直接保留网页的源代码和文件名与网页链接的对应关系即可。

3.2 建立索引的解决思路

建立索引，需要建立五个字段，分别是文件名，本地路径，网页链接，文档标题，以及文档内容，我们只需要对于最后一个字段进行分析即可，因为前面的几个字段实际上是不可读的，添加倒排索引也并没有什么意义。

从爬取下来的页面中，分别对于五个字段进行读取：

1. 文件名（name）：直接读取
2. 本地路径（path）：将文件夹路径以及文件名组合形成本地路径
3. 网页链接（url）：将 index.txt 文件路径作为一个建立索引类的初始化参量，读取 index.txt 文件，形成一个文件名到网页链接的字典，通过文件名，得到网页链接。
4. 文档标题（title）：读取网页源文件，利用 BeautifulSoup 库找到 <title> 标签，读取其中的字符串作为文档标题。
5. 文档内容（contents）：利用 BeautifulSoup 库，去除文档中的所有标签，再利用 jieba 的 lcut 方法进行中文分词，再利用 join 方法并将间隔内容变为空格，自然的，我们的分词器应当使用 WhitespaceAnalyzer，利用空格分词，建立索引。

这样我们建立索引的工作就基本上完成了。

3.3 搜索结果解决思路

对于搜索，我们对于代码的修改主要集中在两个方面：

1. 第一是要将 url 等更多的信息给打印出来，这个是很好办的，只需要改变一下读取文档的字段以及打印的格式即可。
2. 第二个问题主要集中在如何对于中文语段达到比较好的搜索效果，如果输入的词较为短的话当然可以得到比较好的效果，直接搞进去就行。

   但是如果输入的是一句话，基本上不可能完全匹配（内容和顺序），所以考虑在进入搜索之前对于语句进行适当的处理，使得其成为一个布尔语句，用 AND 连接，这样至少突破了顺序完全匹配的限定，可以得到不错的效果。

   实现也是不难的，使用 jieba 分词，并将间隔词设定为" AND "，使用 join 方法即可获得一个布尔语句。

4 代码运行结果

4.1 爬虫的运行结果

```
https://m.guancha.cn/internation/2020_10_15_568166.shtml    httpsm.guancha.cninternation2020_10_15_568166.shtml
https://m.guancha.cn/internation/2020_10_15_568185.shtml    httpsm.guancha.cninternation2020_10_15_568185.shtml
https://m.guancha.cn/internation/2020_10_15_568193.shtml    httpsm.guancha.cninternation2020_10_15_568193.shtml
https://m.guancha.cn/internation/2020_10_15_568178.shtml    httpsm.guancha.cninternation2020_10_15_568178.shtml
https://m.guancha.cn/internation/2020_10_15_568181.shtml    httpsm.guancha.cninternation2020_10_15_568181.shtml
https://m.guancha.cn/internation/2020_10_15_568193.shtml    httpsm.guancha.cninternation2020_10_15_568193.shtml
https://m.guancha.cn/politics/2020_10_15_568207.shtml   httpsm.guancha.cnpolitics2020_10_15_568207.shtml
https://m.guancha.cn/internation/2020_10_15_568178.shtml    httpsm.guancha.cninternation2020_10_15_568178.shtml
https://m.guancha.cn/Shipin/2020_10_15_568175.shtml httpsm.guancha.cnShipin2020_10_15_568175.shtml
https://m.guancha.cn/politics/2020_10_15_568184.shtml   httpsm.guancha.cnpolitics2020_10_15_568184.shtml
https://m.guancha.cn/internation/2020_10_15_568187.shtml    httpsm.guancha.cninternation2020_10_15_568187.shtml
https://m.guancha.cn/internation/2020_10_15_568204.shtml    httpsm.guancha.cninternation2020_10_15_568204.shtml
https://m.guancha.cn/military-affairs/2020_10_15_568168.shtml   httpsm.guancha.cnmilitary-affairs2020_10_15_568168.shtml
https://m.guancha.cn/internation/2020_10_15_568194.shtml    httpsm.guancha.cninternation2020_10_15_568194.shtml
https://m.guancha.cn/politics/2020_10_15_568179.shtml   httpsm.guancha.cnpolitics2020_10_15_568179.shtml
https://m.guancha.cn/internation/2020_10_15_568203.shtml    httpsm.guancha.cninternation2020_10_15_568203.shtml
https://m.guancha.cn/politics/2020_10_15_568209.shtml   httpsm.guancha.cnpolitics2020_10_15_568209.shtml
https://m.guancha.cn/internation/2020_10_15_568181.shtml    httpsm.guancha.cninternation2020_10_15_568181.shtml
https://m.guancha.cn/politics/2020_10_15_568184.shtml   httpsm.guancha.cnpolitics2020_10_15_568184.shtml
https://m.guancha.cn/internation/2020_10_15_568157.shtml    httpsm.guancha.cninternation2020_10_15_568157.shtml
```

图 1: index.txt 文件内容（存储文件名与网页链接的关系）

```
<!DOCTYPE html>
<html>
<head>
<meta charset="utf-8">
<meta name="format-detection" content="telephone=no" />
<meta name="viewport" content="width=device-width,initial-scale=1.0,minimum-scale=1.0,maximum-scale=1.0,user-scalable=no">
<meta name="applicable-device"  content="mobile">
<title>一位澳籍华裔遭“忠诚度”测试：敢不敢谴责中国？-观察者网</title>
<link rel="stylesheet" href="../css/public.css?20181218">
<link rel="stylesheet" href="../css/main.css?v20200211">
<link rel="canonical"  href="http://www.guancha.cn/internation/2020_10_15_568157.shtml" >
<link rel="apple-touch-icon-precomposed" href="https://i.guancha.cn/news/2020/10/15/20201015120313151.jpg!m.jpg">
<meta name="keywords" content="" />
<meta name="description" content="【编译/观察者网 童黎】作为澳大利亚“人均水平（Per Capita）”智库的华裔研究员，奥斯蒙德·邱（Os" />
<script>var _API_URL="http://api.guancha.cn";</script>
<script>
    var id="475707";
    var _DOC_ID="475707";
    var making = 'yuanchuang';
</script>
<script src="../js/jquery.min.js" type="text/javascript"></script>
</head>
<body>
    <div class="downloadBtn-box">
        <div class="downloadBtn" id="goapp" style="display:none">
            <a href="javascript:;" onclick="Xbutton()" id="Xbutton"></a>
            <img src="../images/ico.png" alt="观察者App">
            <span>观察者App</span>
            <button onclick="appstore()" id="appstore">点击下载</button>
        </div>
    </div>
```

图 2: 存下来的网页源代码（部分）

4.2 建立索引的运行结果

4.3 搜索结果的运行结果

```
lucene 8.6.1
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 0.738 seconds.
Prefix dict has been built successfully.
warning: no content in httpsm.guancha.cninternation2020_10_15_568193.shtml
warning: no content in httpsm.guancha.cninternation2020_10_15_568204.shtml
warning: no content in httpsm.guancha.cnpolitics2020_10_15_568202.shtml
commit index
.done
0:00:02.721865
```

图 3: 建立索引过程中终端打印信息

```
Searching for: 战争 AND 与 AND 和平

5 total matching documents.

path:   html/httpswww.guancha.cnOuShuJun2020_10_10_567613.shtml
title:  欧树军：冷战后失去外敌的美国，陷入国家认同危机
url:    https://www.guancha.cn/OuShuJun/2020_10_10_567613.shtml
score:  3.220496654510498

path:   html/httpswww.guancha.cnbaoer2020_10_13_567880.shtml
title:  保尔：明清易代对中国历史走向的影响，真有那么大?
url:    https://www.guancha.cn/baoer/2020_10_13_567880.shtml
score:  2.7937066555023193

path:   html/httpswww.guancha.cnQuora2020_10_13_567885.shtml
title:  外网讨论“中美脱钩”：美国企业不是讨厌中国，是讨厌自己本国人-Quora
url:    https://www.guancha.cn/Quora/2020_10_13_567885.shtml
score:  2.33467173576355

path:   html/httpswww.guancha.cnzhangjun32020_10_08_567492.shtml
title:  张军大使：我正告你们，趁早悬崖勒马
url:    https://www.guancha.cn/zhangjun3/2020_10_08_567492.shtml
score:  2.3337409496307373

path:   html/httpswww.guancha.cnMaryMcCord2020_10_11_567707.shtml
title:  玛丽·麦考德：绑架州长背后，是全美越来越多不服管束的武装分子
url:    https://www.guancha.cn/MaryMcCord/2020_10_11_567707.shtml
score:  2.231516122817993
```

图 4: 搜索内容为“战争与和平”的搜索结果