

# 武汉大学国家网络安全学院

## 实验报告

课程名称 信 息 检 索

专业年级 2018 级网络安全

姓 名 李政

学 号 2018302060120

协 作 者 无

实验学期 2020-2021 学年 第二 学期

课堂时数 32 课外时数

填写时间 2021 年 6 月 15 日

## 实验介绍

【实验名称】：中文搜索引擎简易实现——爬虫，倒排索引，NLP

【实验目的】：

实现简易中文网页搜索引擎

- 1.爬取大量网站作为数据
- 2.分析网站文字并进行预处理
- 3.对获取的数据进行索引
- 4.对检索输入的语句进行分析纠错
- 5.在索引中检索并输出对应网站和标题

【实验环境】：

- 系统：

Edition	Windows 11 Pro
Version	Dev
Installed on	2021/6/17
OS build	21996.1
Experience	Windows Feature Experience Pack 321.14700.0.3
- 语言：Python 3.8
- IDE：VS Code
- 关键库：Jieba, Urllib, BeautifulSoup, xpinyin 等，详细见附件 requirements.txt

【参考文献】：

- [1] 倒排索引 <https://www.cnblogs.com/zlsich/p/6440114.html>
- [2] 索引与检索 <https://zhuanlan.zhihu.com/p/91732698>
- [3] 中文文本自动纠错 <https://blog.csdn.net/kobeyu652453/article/details/106905191>
- [4] BFS <https://blog.csdn.net/g11d111/article/details/76169861>

## 实验内容

【实验方案设计】：

### 1.1 爬虫部分

#### 1.1.1 广度优先链接爬取

通过输入初始 url，扫描页面中的可用链接，将扫描到的未访问过的链接加入待访问链接表，将深度 n 作为参数控制广度优先遍历的深度。

设置 LinkQuence 类来操作：

```

class linkQuence:
    def __init__(self):
        self.visted = []
        self.unVisited = []

    def getVisitedUrl(self):
        return self.visted

    def getUnvisitedUrl(self):
        return self.unVisited

    def addVisitedUrl(self, url):
        self.visted.append(url)

```

.....

通过循环实现广度优先遍历：

```

while self.current_deepth <= crawl_deepth:
    links = []
    while not self.linkQuence.unVisitedUrlsEnmpy():
        visitUrl = self.linkQuence.unVisitedUrlDeQuence()
        if visitUrl is None or visitUrl == "":
            continue

    .....

    print("遍历深度: "+str(self.current_deepth))
    for link in links:
        self.linkQuence.addUnvisitedUrl(link)
    self.current_deepth += 1

```

.....

### 1.1.2 页面解析和正文标题提取

采用 BeautifulSoup 对页面解析，默认网站正文处于<p>...</p>下，因此提取所有标签 p 下的文字作为正文，采用正则匹配匹配链接：

```

soup = BeautifulSoup(data[1])
title = soup.title.text
ps = soup.find_all("p")
for p in ps:
    text += p.text
    a_1 = soup.findAll("a", {"href": re.compile('^http|^/')})
    a_2 = soup.findAll("a", {"href": re.compile('^https|^/')})
for i in a_1:
    if i["href"].find("http://") != -1:
        links.append(quote(i["href"], safe=";/?:@#&=+$,"))
for i in a_2:
    if i["href"].find("https://") != -1:
        links.append(quote(i["href"], safe=";/?:@#&=+$,"))

```

.....

### 1.1.3 结果保存

将爬取结果保存到文件中，正文保存作为索引依据，标题和链接单独保存作为检索结果展示：

```

def save_results(self, path, url, title, text):
    md5hash = hashlib.md5(url.encode("utf-8"))
    md5 = md5hash.hexdigest()
    with open(path+"pages/"+md5, "w", encoding="utf-8") as f:
        f.write(text)
        f.close()
    with open(path+"links/"+md5, "w", encoding="utf-8") as f:
        f.write(url+","+title)
        f.close()

```

.....

## 1.2 预处理和索引

### 1.2.1 分词，去停用词处理

由于面向中文数据，因此需要进行分词，部分词无意义并且可能影响结果，因此提前构造了停用词表：

```

f = codecs.open(participle_tmp_path, 'w', encoding="UTF-8-SIG")
    for line in open(page_path+"\\ "+file_name, encoding='utf-8', errors='ignore').readlines():
        line = re.sub(
            r"[0-9\s+\.!\|/_,$%^*()?;,:-【】+\"'\"]+|[+—! , ;:。 ? 、~@#¥%……&* ( ) ]+", " ", line)
        seg_list = jieba.cut(line, cut_all=True)
        seg_list = movestopwords(seg_list, stopwords)
        if len(seg_list) == 0:
            continue
        f.write(" ".join(seg_list)+"\n")
    .....
```

### 1.2.2 倒排索引

建立词表并计算词频：

```

# 建立词表
sp_data = src_data.split()
set_data = set(sp_data) # 去重复
word = list(set_data) # set 转换成 list, 否则不能索引
# 词频
words = list(sp_data)
dic_word_count = Counter(words)
for word in dic_word_count.keys():
    dic_word_count[word] = [page, dic_word_count[word]]
    if word in index.keys():
        index[word].append(dic_word_count[word])
    else:
        index[word] = [dic_word_count[word]]
    .....
```

记录位置并建立索引：

```
# 词位置
src_list = src_data.split("\n") # 分割成单段话 vv
# 建立索引
for w in range(0, len(word)):
    para_pos = [] # 记录段落及段落位置 [(段落号, 位置), (段落号, 位置)...]
    for i in range(0, len(src_list)): # 遍历所有段落
        sub_list = src_list[i].split()
        for j in range(0, len(sub_list)): # 遍历段落中所有单词
            if sub_list[j] == word[w]:
                para_pos.append((i, j))
        result[word[w]] = para_pos

.....
```

统计词频并保存

```
for key in index.keys():
    df = len(index[key])
    for file_tf in index[key]:
        tf = file_tf[1]
        w = (1.0 + math.log(tf)) * math.log10(N / df)
        file_tf.append(w)
    with codecs.open(index_path, 'w', encoding="gb18030") as i:
        i.write(json.dumps(index))

.....
```

## 1.3检索部分

### 1.3.1 输入语句处理

为了提升搜索引擎的鲁棒性,设置搜索语句的检测,避免出现用户简单的拼写错误导致检索结果不准确,比如百度搜索也有这样的功能:



因此我们需要借助中文语法纠错实现此功能

1. 本地构建一个正确的字词库
2. 构建一个文档用于编辑距离操作
3. 输入一个错误单词 (句子分词得到的单词, 或者单独一个错误单词), 计算编辑距离, 生成编辑距离词集。编辑距离需要比对 数据库.txt 的单词,计算距离然后对错误单词进行删除字, 增加字, 修改字, 替换字。增加删除替换的字, 需要从编辑距离文档里选取

字插入或替换到错误单词里。最后生成编辑距离词集。

4. 生成的编辑距离词集可能含有一些错误单词，找出同时在编辑距离词集和字词数据库的单词，即为我们候选正确词集。
5. 对候选正确词进行分级。首先 `pinyin.get` 得到错误词的拼音。然后遍历候选正确词集的单词，求取得拼音。我们根据候选词的拼音对其重要性进行排序如果候选词的拼音与错误词完全匹配，则将候选词放入一级数组。如果候选词的第一个词的拼音与错误词的第一个词匹配，我们将其按二级数组。否则我们把候选短语放入三级数组。
6. 找到正确单词，如果一级数组存在，得到的正确字词是在字词数据库中的。考虑到得到的词可能有多个，前文提到字词数据库第一列是词，第二列是词频。我们应该返回一级数组中词在字词数据库中词频最大的那个单词。如果一级数组不存在，二级数组存在，返回词频最大的那个单词。否则，返回三级数组词频最大的那个单词。

### 1.3.2 检索并输出结果

对输入的句子进行关键词提取，统计出现符合关键词的数目和关键词出现的频率，找到最优结果。这里采用 `jieba` 的 `tf-idf` 方法进行关键词抽取：

```
tfidf = analyse.extract_tags
keywords = tfidf(query)
pages = {}
for word in keywords:
    if word in index:
        for page in index[word]:
            if page[0] in pages:
                pages[page[0]] += page[2]
            else:
                pages[page[0]] = page[2]
page_list = sorted(pages.items(), key=lambda item: item[1], reverse=True)
```

.....

获取匹配的页面的名称（md5 值），找到存放对应链接的文件并打印：

```
def print_info(i, index_name, datapath="data/database/links/"):
    with open(datapath+str(index_name), encoding="utf-8") as f:
        li = f.readline().split(",")
        print(str(i)+": "+str(li[1])+"\""+str(li[0])+"\"")
```

.....

**【实验结果分析】：**

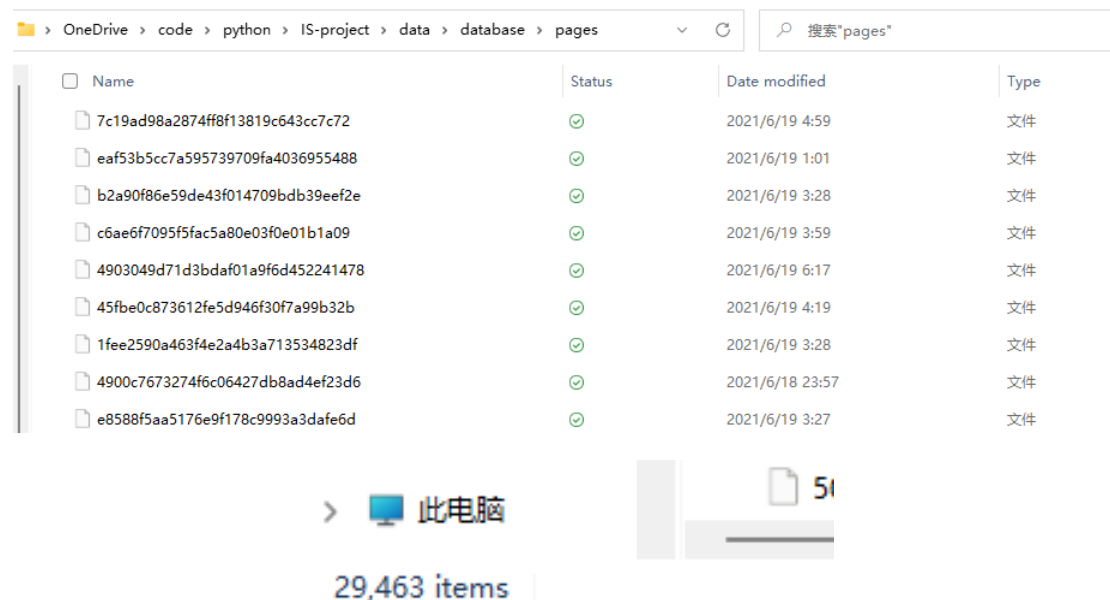
## 2 结果演示

## 2.1准备工作

以百度，武汉大学首页，hao123 起始，层数为 3 进行爬取（如果出现报错会跳过网站并打印信息）：

```
已经爬取链接数: 13173
遍历深度: 2
已经爬取链接数: 13174
遍历深度: 2
已经爬取链接数: 13175
遍历深度: 2
已经爬取链接数: 13176
遍历深度: 2
已经爬取链接数: 13177
遍历深度: 2
HTTP Error 404: CHttpException
已经爬取链接数: 13178
遍历深度: 2
已经爬取链接数: 13179
遍历深度: 2
已经爬取链接数: 13180
遍历深度: 2
```

最终爬取到了 3 万个网站及其信息作为数据集：



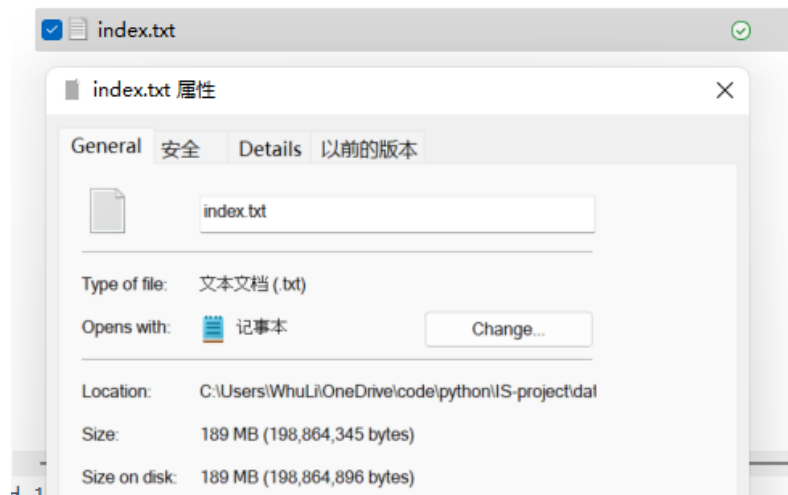
Name	Status	Date modified	Type
7c19ad98a2874ff8f13819c643cc7c72	✓	2021/6/19 4:59	文件
eaf53b5cc7a595739709fa4036955488	✓	2021/6/19 1:01	文件
b2a90f86e59de43f014709bdb39eef2e	✓	2021/6/19 3:28	文件
c6ae6f7095f5fac5a80e03f0e01b1a09	✓	2021/6/19 3:59	文件
4903049d71d3bda01a9f6d452241478	✓	2021/6/19 6:17	文件
45fbc0c873612fe5d946f30f7a99b32b	✓	2021/6/19 4:19	文件
1fee2590a463f4e2a4b3a713534823df	✓	2021/6/19 3:28	文件
4900c7673274f6c06427db8ad4ef23d6	✓	2021/6/18 23:57	文件
e8588f5aa5176e9f178c9993a3dafa6d	✓	2021/6/19 3:27	文件

使用采集到的数据进行预处理和索引：

```
对29463条数据进行索引
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\WhuLi\AppData\Local\Temp\jieba.cache
Loading model cost 0.916 seconds.
Prefix dict has been built successfully.
>>>建立索引完成！用时1015.404458秒
.....
```

最终生成索引文件：





此后可直接加载此文件进行数据检索，如果有新的数据，才需要重新进行索引

## 2.2使用展示

使用前需要初始化，将索引加载到内存：

```
>>>初始化完成！用时5.770063秒
```

请输入搜索内容：

n: 当前文件 (IS-project)

尝试简单搜索（结果包含标题和链接）：

```
请输入搜索内容：航空
搜索【航空】，为您找到327条结果(用时0.002992秒):
1: 首页-中原人才网(http://www.zyrc.com.cn/)
2: 武汉长投航空路壹号·长投航空路壹号户型图_地址_长投·航空路壹号房价-武汉房天下(https://www.fang.com/xinfang/wuhan-2610157034/)
3: 多地绘出航空航天产业发展蓝图_专家看好三大领域_财经_中国网(http://finance.china.com.cn/industry/20210618/5595753.shtml)
4: 驴妈妈社会招聘-驴妈妈旅游网(http://www.lvmama.com/public/jobs/)
5: cjp1(http://guba.eastmoney.com/news/)
6: 太诡异！突发全球性断网，知名金融机构、航空公司集体宕机，港交所受波及，什么情况？|网络安全|黑客攻击|网易财经(https://www.163.com/money/article/GC05G4P600259DLP.html?clickfrom=we_money)
7:
8: 航空旅游网(http://www.cnair.com/)
9: 美欧航空补贴争端“休战”难掩深层经贸分歧_新闻新闻(https://news.sina.com.cn/w/2021-06-18/doc-ikqcfnc1036812.shtml)
10: 武汉悦荟天地_悦荟天地户型图_地址_悦荟天地房价-武汉房天下(https://www.fang.com/xinfang/wuhan-2610159006/)
已显示前10条，是否显示所有？[y/n]
```

尝试完整句子搜索：

```
请输入搜索内容：小米手机多少钱啊？
搜索【小米手机多少钱啊】，为您找到4416条结果(用时0.012970秒):
1: 手机主页-中关村在线手机频道(http://mobile.zol.com.cn/)
2: 手机主页-中关村在线手机频道(https://mobile.zol.com.cn/)
3: 手机_手机通讯_苏宁手机_正品行货【价格_品牌_行情_评价测评】-苏宁易购网上商城(https://sucs.suning.com/visitor.htm?userId=32258543&webSiteId=)
4: 新闻中心_IT新闻资讯_IT新闻动态-中关村在线新闻中心频道(https://news.zol.com.cn/)
5: 手机通讯_手机通讯【价格_品牌_推荐_正品折扣】-当当网(http://category.dangdang.com/cid4006497.html)
6: 小米MIX Fold_小米MIX Fold 报价、参数、图片、怎么样_太平洋产品报价(https://product.pconline.com.cn/mobile/miui/1373696.html)
7: 小米小爱音箱HD - 小米商城(https://www.mi.com/aispeaker-hd/)
8: 咪咕隐私政策(https://passport.migu.cn/portal/privacy/protocol?sourceid=220001)
9: 小米太憋屈了！刚把华为熬下去，又迎来了合并加强版的新劲敌_手机(https://www.sohu.com/a/472843531_121031229)
10: 小米手机 MIUI 专场解答：8 月份上线均衡/性能模式，小米 11 Ultra 对不兼容充电器做了限制 - IT之家(https://www.ithome.com/0/557/838.html)
已显示前10条，是否显示所有？[y/n]
```

尝试冗余信息搜索：

<p>请输入搜索内容：请问大家，华为手机价格和性能怎么样呢？？？          搜索【请问大家 华为手机价格和性能怎么样呢】，为您找到5990条结果(用时0.023719秒)：          1: 手机主页-中关村在线手机频道\(<a href="http://mobile.zol.com.cn/">http://mobile.zol.com.cn/</a>)          2: 手机主页-中关村在线手机频道\(<a href="https://mobile.zol.com.cn/">https://mobile.zol.com.cn/</a>)          3: 新闻中心-IT新闻资讯-IT新闻动态-中关村在线新闻中心频道\(<a href="https://news.zol.com.cn/">https://news.zol.com.cn/</a>)          4: 手机_手机通讯_苏宁手机_正品行货【价格 品牌 行情 评价测评】 - 苏宁易购网上商城\(<a href="https://sucs.suning.com/visitor.htm?userId=3225854">https://sucs.suning.com/visitor.htm?userId=3225854</a>)          5: \(<a href="https://enterprise.pconline.com.cn/">https://enterprise.pconline.com.cn/</a>)          6: 华为WATCH 3体验：全景设备、一表掌控，真旗舰-中关村在线综合论坛\(<a href="https://bbs.zol.com.cn/quanzi/d14_29600.html">https://bbs.zol.com.cn/quanzi/d14_29600.html</a>)          7: 618买手机先别急 预算2k到3k看完横评再下手_一加 9R_手机评测-中关村在线\(<a href="https://mobile.zol.com.cn/770/7702023.html">https://mobile.zol.com.cn/770/7702023.html</a>)          8: 太平洋汽车网_精准报价_专业评测_以车会友\(<a href="http://www.pcauto.com.cn/">http://www.pcauto.com.cn/</a>)          9: 华为三款“新机”上市：现货不用抢，4G+鸿蒙系统买不买？_手机\(<a href="http://www.sohu.com/a/472796890_120597085">http://www.sohu.com/a/472796890_120597085</a>)          10: 手机厂商来风_手机新品_手机_太平洋电脑网手机频道\(<a href="https://mobile.pconline.com.cn/news/changshang/">https://mobile.pconline.com.cn/news/changshang/</a>)          已显示前10条，是否显示所有？[y/n]</p>	<p>尝试刻意构造错误搜索检验鲁棒性：</p> <p>请输入搜索内容：砖家团队          即将显示【专家团队】的搜索结果。仍然搜索【砖家团队】？[y/n]          搜索【专家团队】，为您找到1713条结果(用时0.010280秒)：          1: 大学报志愿咨询收费10万元天价 所谓规划专家身份真假难辨 高考_新浪财经_新浪网\(<a href="http://finance.sina.com.cn">http://finance.sina.com.cn</a>)          2: 军事视频 军事新闻 - 中国军视网 最大的军事视频网站\(<a href="http://www.js7tv.cn/">http://www.js7tv.cn/</a>)          3: 2021年中甲联赛_新浪网\(<a href="http://sports.sina.com.cn/zt_d/zhongjia/">http://sports.sina.com.cn/zt_d/zhongjia/</a>)</p>
<p><b>【实验总结】：</b></p> <p>通过本次实验，对倒排索引等信息检索方法进一步学习和实验，并且结合爬虫和自然语言处理，简易的实现了一个带有自动数据爬取的中文网页搜索引擎，并且对用户输入的搜索语句有一定的语法纠正功能。在本次实验中，爬取数据的处理，倒排索引以及搜索和语法纠错部分为重点部分，也遇到一定的困难，最后通过网上查阅参考文献得到解决。经过本次实验，实现了信息检索的功能并且将其融入到搜索引擎中，收获颇多。</p>	
<p><b>评语及评分（指导教师）</b></p>	
<p><b>【评语】：</b></p> <p>该实验报告设计并实现了基于 TF-IDF 实现的文档检索系统，并在自己爬取的数据集上进行了定性评价，基本达到设计目标。报告关于爬虫部分说明详细，虽然关于输入字符串进行了鲁棒性考量，但是实现的查询方式较为单一且缺少评价算法。</p> <p style="text-align: right;">评分：81 马越          日期：2021-6-26</p>	

附件：

## 实验报告说明

- 1. 实验名称：**要用最简练的语言反映实验的内容。
- 2. 实验目的：**目的要明确，要抓住重点。
- 3. 实验环境：**实验用的软硬件环境（配置）。
- 4. 实验方案设计（思路、步骤和方法等）：**这是实验报告极其重要的内容。包括概要设计、详细设计和核心算法说明及分析，系统开发工具等。应同时提交程序或设计电子版。

对于**设计型和综合型实验**，在上述内容基础上还应该画出流程图、设计思路和设计方法，再配以相应的文字说明。

对于**创新型实验**，还应注明其创新点、特色。

- 5. 实验结果分析：**即根据实验过程中所见到的现象和测得的数据，进行对比分析并做出结论（可以将部分测试结果进行截屏）。
- 6. 实验总结：**对本次实验的心得体会，所遇到的问题及解决方法，其他思考和建议。
- 7. 评语及评分：**指导教师依据学生的实际报告内容，用简练语言给出本次实验报告的评价和价值。