信息检索系统作业报告

2018211517 何泓川 2018211512 鲁嘉祺

实验目的

自己动手设计实现一个信息检索系统,数据源可以自选,数据通过开源的网络爬虫获取,规模不低于 100 篇文档,进行本地存储。中文可以分词(可用开源代码),也可以不分词,直接使用字作为基本单元。英文可以直接通过空格分隔。构建基本的倒排索引文件。实现基本的向量空间检索模型的匹配算法。用户查询输入可以是自然语言字串,查询结果输出按相关度从大到小排序,列出相关度、题目、主要匹配内容、URL、日期等信息。最好能对检索结果的准确率进行人工评价。

实验环境

python3.9

完成情况

本实验完成了 1-5 的全部要求。

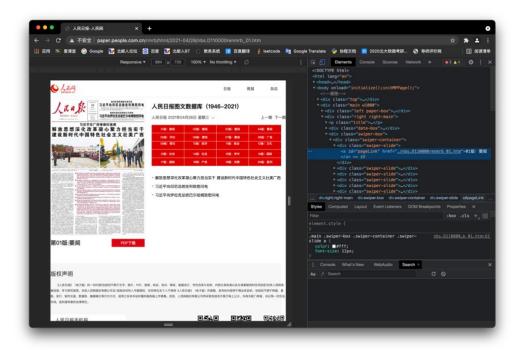
实验过程

(1) 爬取数据集

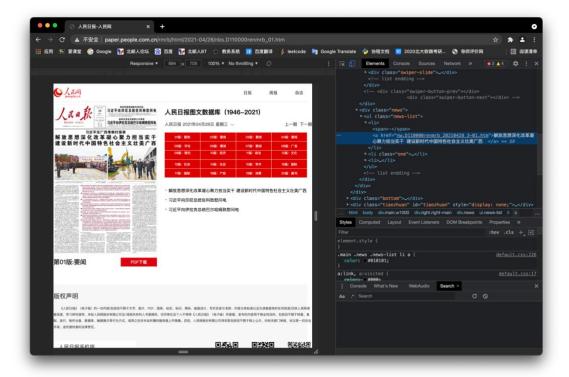
本实验共有20个分类,104篇文档,如下图:

mac@BruceHo info_search % python3.9 spider.py
20 kind_linklist finished
file count: 104

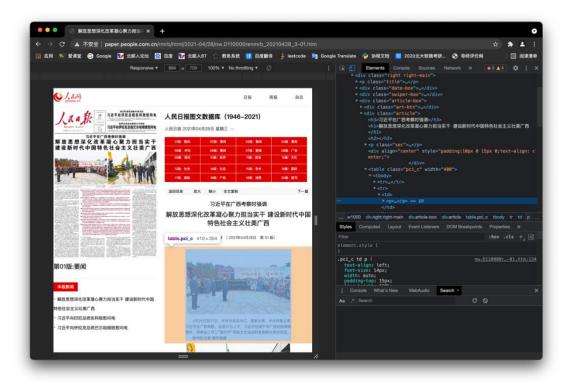
本实验中使用人民日报的文章作为数据集。先根据各类别的特征找到其对应的 url 并保存所有类别的 url,对应的函数为 spider.py 中的 getKindList 函数。下图中的类别特征为 class='swiper-container'的 div 标签下的 class='swiper-slide'的 div 标签。



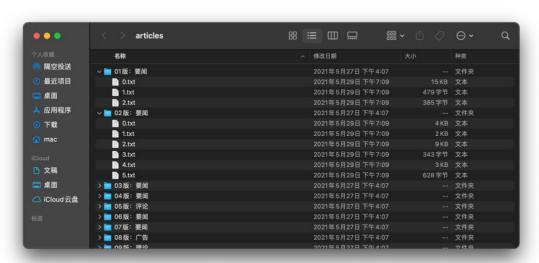
再根据一个类别中文章 url 的特征,保存所有类别所有文章的 url。相关的函数为 spider.py 中的 getTitleList 函数。下图的特征为 class='news'的 div 标签下的 li 标签。



再访问所有文章的 url,通过文章的特征找到文章的文本内容。下图的文章特征为id='ozoom'的 div 标签下的 p 标签。保存所有文章的文本内容到.txt 文件。相关的函数为 spider.py 中的 saveContent 函数。



爬取所有文章的结构如下:



所有 url 信息保存在文本文件中,内容如下:

```
01版: 要闻
解放思想求心政革凝心聚力担当实于 建设新时代中国特色社会主义社美广西
http://paper.people.com.cn/zmtb/html/2021-04/28/nw.D110000renmrb_20210428_3-01.htm
习近平向印尼总统佐科致影问电 http://paper.people.com.cn/zmrb/html/2021-04/28/nw.D110000renmrb_20210428_1-01.htm
习近平向伊拉克总统巴尔哈姆致慰何电 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_2-01.htm
V-NAV - NAM
李克强主持召开国务院常务会议 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_1-02.htm
一季度規上工业企业利润同比增1.37倍 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_2-02.htm
外贸外资双双"开门缸"(开局之年看外贸外资) http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000Tenmrb_20210428_3-02.htm
李克强将同德国总理默克尔以视频形式共同主持第六轮中德政府延商 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000Tenmrb_20210428_4-02.htm
栗战书同委内瑞拉全国代表大会主席罗德里格斯举行会谈 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_5-02.htm
薄涂变公园 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_6-02.htm
03版: 要闽
                             http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_1-03.htm
老中命运共同体的生动体现(国际论坛)
                                           http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_2-03.htm
http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_3-03.htm
"为推动亚洲和世界发展作出重要贡献" http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_5-03.htm 2021年中日韩合作国际论坛在首尔举行 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_6-03.htm
图片报道 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_7-03.htm
云蜡看限 感受魅力 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_8-03.htm
04版: 要闻
新闻发布工作呈现新亮点取得新成效 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_1-04.htm
"革命理想高于天"(人民论坛) http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_2-04.htm
消博会有序布展 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_3-04.htm
为庆祝建党100周年创造安全稳定环境 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_4-04.htm
国务院任务国家工作人员 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_5-04.htm
香港各界青年畅谈加强爱国主义教育、融入国家发展大局 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_6-04.htm
推动新职业更好更快发展(评论员观察)
                                          http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_1-05.htm
让健身成为一种生活方式(新论) http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_2-05.htm
提升对口理確綜合效益(现场评论) http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_3-05.htm
图片报道 http://paper.people.com.cn/rmrb/html/2021-04/28/nw.D110000renmrb_20210428_4-05.htm
```

(2) 信息检索

先进行分词。分词使用 jieba 库进行分词,分词过程在 divide.py 中。再把所有文章的词向量化,如下图:

```
def gen_vector(article_list):
    bag = CountVectorizer(token_pattern='\\b\\w+\\b')
    count = bag.fit_transform(article_list)
    return bag, count
```

其中 token_patter 的正则表达式为把两个空格之间的词作为向量一个单元。 再生成倒排索引,可以通过一个单词找到所有包括它的文章、这个词在该文章中的 出现次数、在该文章中的所有出现位置,如下图:

做完倒排索引后可以进行查询。先读入一个字符串,如果是 exit 则结束程序,否则进入 get_result 函数计算结果。先把字符串按空格分离出来,即对每个关键词依次计算。先计算所有关键词在多少不重复的文章中出现,记为 freq。再逐个分析每个关键词。遍历每一条该关键词的倒排索引,如果这条索引的文章不在结果列表中,则加入结果列表,该文章出现的关键词个数 keyword_num 加一,该文章中关键词出现的次数 keyword_times 加一,更新该文章中关键词出现次数占所有关键词出现的文章数的比例 ratio。结果列表中也要保存该词在该文章的出现位置。如果关键词的这条索引在结果列表中,则 keyword_times 和 keyword_num 加一,更新 ratio,保存出现位置。对所有关键词遍历后把关键词向量化,并与结果列表的每一条计算相似

度 correlation (余弦值)。此时结果列表信息已经完整,按照 ratio、keyword_num、manual_point、correlation 乘积倒序排列,即位最终结果。原因是该文章中关键词出现次数更多,所有关键词次数出现文章数更少,该文章命中的关键词个数更多,人工评价权值更高,相似度越高,则该文章更匹配。人工评价权值初始值都为 1,每次查询后人工对结果进行评价,选中的文章人工评价权值乘 1.1,即下次更容易被列在结果列表前部。信息检索部分实现如下:

```
result_dict = dict()
for i in word_result: # 一个关键词
   if not i:
    for info in i: # 遍历该关键词出现的所有文章
       if info[0] not in result_dict:
           item = resultItem(info[0], article_names[info[0]], article_list[info[0]])
           item.keyword_num += 1 # 该文章被多少关键词命中
           item.keyword_times += info[1] # 该文章中关键词出现次数
           item.ratio += info[1] / freq # 该文章中关键词出现次数 / 所有关键词出现的文章数
           item.occurence.extend(info[2])
           result_dict[info[0]] = item
           item = result_dict[info[0]]
           item.keyword_num += 1
           item.keyword_times += info[1]
           item.ratio += info[1] / freq
           item.occurence.extend(info[2])
           result_dict[info[0]] = item
result_list = [i for i in result_dict.values()]
search_vec = CountVectorizer(vocabulary=bag.get_feature_names()).fit_transform([search_str]
for i in result_list:
   i.correlation = calc_correlation(search_vec[0], array[i.index].A[0])
result_list.sort(key=lambda x: -x.ratio * x.keyword_num * x.manual_point * x.correlation)
```

实验结果

搜索一个关键词,可显示所有包含该关键词的文章,出现次数多的靠前,列出相关文章信息:

如果搜索多个关键词,同时包含所有关键词的文章排序更靠前:

人工评价: 开始时标题为《贵州兴义 服务消费亮点多(构建新发展格局·县域消费观察)》的文章在结果列表第二位,多次人工选择该文章后出现在了结果列表第一位。

```
终端 问题 输出 调试控制台

→ + → □ □ · 

  mac@BruceHo info_search % python3.9 test.py
type string you want to search
UI⊠
type 51/1/19 y/so-sen
山区
[0]:
村类: 14版: 生态
大寒味题: 树木成行 郁郁茎葱 (美丽中国)
虹/til http://paper.people.com.cn/msrb/html/2021-04/28/nw.D110000renmrb_20210428_1-14.htm
该文章中关键词出现及数: 3
光键词况取 (持高义准数: 1.5
正人写的故语: 26-2534137
正人写的故语: 26-2534137
用序码分: 0.116097348489012054
树近内容: 名字 祥瑞 意 嘉祥县 ゼ 盟 封山 生态 立县 嘉祥县 山区 绿色 发展 路 山体 黄白色 树木 山东省 济宁市 嘉祥县
树近内容: 化中 铁 锅 精酸 绿葱 无序 粗致式 嘉祥县 域 山体 山区 山体 以 庆 青山 嘉祥县 棋 联户 实力 公司 综合
树近内容: 也本 依 锅 精酸 绿葱 无序 粗致式 嘉祥县 域 山区 山体 以 庆 青山 嘉祥县 棋 联户 实力 公司 综合
树近内容: 专业 合作 山体 锅 锅 抽味 绘图 接触 经惠 无序 粗致式 嘉祥县 域 山体 山区 山体 以 庆 青山 嘉祥县 棋 联户 实力 公司 综合
树近内容: 专业 合作 山体 锅 锅 地址 当地人 號 锅 播撒 绿葱 嘉祥县 山区 封山育林 植树造林 生态 短板 发展 优势 全县 发展 思
 [5]:
中类: 18版: 統済
文章标题: 弗州兴义 服务消费系点多(构建新发展特局-县城消费观察)
文章标题: 弗州兴义 服务消费系点多(构建新发展特局-县城消费观察)
 、果都縣: 如州宋义 · 据为消费地应多 · (特殊前发操信用 · 鲁·瑞州姆殊龄)

UPI: http://pager.people.com.cn/mrb/html/2821-04/28/m⊿.D110900Fremrb_28210428_1-10.htm

该文章中关键问出数决数:3

长键问数数优易来文章数:1.5

匹配度:0.66175408271944637

人工评价权值:5.0

相序得分:0.9623407916955

相序得分:0.9623407916955

相好内容:世影 臨時 山区 小城 生活 溢味 贵州 兴义 服务 消费 亮点 发展 格局

附近内容:贵州 "兴、服务" 消费 亮点 发展 相同 表现 小城 "非

附近内容:为口卫生 消费 升级 步伐 县级市 约载 消费 业态 模式 山区 小城 相影 观影 人群 中老年人 电影院 年轻 D

耐近内容:打口卫生 消费 升级 步伐 县级市 约载 消费 业态 模式 山区 小城 电影 观影 人群 中老年人 电影院 年轻 D
1
type string you want to search
山区
[0]:
种类: 14版: 生态
    前五广记。
[1]:
种类: 10版: 经济
又草标题: 贵州兴义 服务消费亮点多 (构建新发展格局-最著消费观察)
又草标题: 贵州兴义 服务消费亮点多 (构建新发展格局-最著消费观察)
url: http://paper.people.com.cn/mrb/html/2821-84/28/mw.D118000remmrb_20210428_1-10.htm
级文章中关键询为规定数据: 3-5
区配层 0.60575402271944637
人工师仍设值: 1_21000000000000002
排斥师分: 0.11200555123579518
附近内容: 5人1200555123579518
附近内容: 贵州 兴义 服务 消费 表点 发展 格局 最短 消费 核心 山区 小城 用房 亮点 贵州 兴义市 老年人 电影院 年轻人 精
附近内容: 贵州 兴义 服务 消费 丢点 发展 格局 最短 消费 核心 山区 小城 电影 观影 人群 中老年人 电影院 年轻人 精
附近内容: 打口卫生 消费 升级 步伐 县级市 封整 消费 业态 模式 山区 小城 电影 观影 人群 中老年人 老人 电影院 年轻
```

实验结论

本实验可以完成信息检索系统的功能,根据输入的关键词找到相关文章及其信息,并按相关度从大到小排列。考虑到对环境和社会可持续发展影响的考虑,本实验数据集的选择中选择

了有生态版块的人民日报网的文章。

创新性思考和多媒体信息搜索

本实验中,由于考虑到实际的搜索引擎可以对多个关键词进行搜索,增加了多个关键词搜索的功能,对所有结果排序方法使用多个指标乘积,使相关度更高的结果排序更靠前,而不是仅考虑向量间余弦值,使相关度计算方法更完善。

result_list.sort(key=lambda x: -x.ratio * x.keyword_num * x.manual_point * x.correlation)

在多媒体(如视频、图片)信息搜索中,可以先把所有图片向量化组成图片向量库,再把输入图片也进行向量化,计算输入图片与所有图片向量库中向量的余弦值,把结果倒序排列。 在本程序中,只需在 get_result 函数中直接把输入图片向量化,并计算与所有图片向量的余弦,结果倒序排序,即位图片或视频的信息检索。