**华东师范大学数据科学与工程学院实验报告**

|  |  |  |
| --- | --- | --- |
| **课程名称**：信息检索 | **年级**：2019 | **上机实践成绩**： |
| **指导教师**：张蓉 | **姓名**：张雯怡 |  |
| **上机实践名称**：倒排索引构建 | **学号**：10195501425 | **上机实践日期**： |
| **上机实践编号**：2 | **组号**： | **上机实践时间**： |

1. **实验目的**

学习构建基于倒排索引的检索系统

1. **实验任务**

子任务1：建索引

子任务2：布尔查询

输入：

-给定100000个文挡

-一个布尔查询Q （And Or 操作）

输出：

-倒排索引表（存成文件dict.index）dict.index.txt

-Q的查询结果（界面输出文档名列表）

文本处理要求：

-不要求做词条变化如friends -> friend等，直接用空格作为分割符

-都转成小写A->a

-把出现次数排名Top 100的词去掉(a, an, the 等等)

要求：

-至少完成1000个文档索引

-把索引写入dict.index

-从文件中读取索引用于查询

-支持连续输入查询，即，输入Q1后，得到结果；接着输入Q2

1. **使用环境**

Python 3.0

1. **实验过程**

Task 1：构建倒排索引

首先读取文件，将其分行后再对每行进行分词，把分词结果保存至一个元组中。

path = './数据/'

# filelist = [path + i for i in os.listdir(path)]

for i in range(1,10001):

    file = os.path.join(path, str(i) + ".txt")

    tmp = []

    with open(file, 'r') as file\_to\_read:

        while True:

            lines = file\_to\_read.readline()

            lsplit = lines.split( )

            if (len(lines) == 0): break

            for word in lsplit:

                tmp.append(word)

        files.append(tmp)

    file\_to\_read.close()

然后构建倒排索引表，这里要建立两个列表dict1和dict2，dict1{}用来存储每个单词以及包含它的所有文件编号的链表，dict2{}用来存储每个单词出现的频次。

for i in range(10000):

    file = files[i]

    #print(file)

    for word in file:

        if (word == ' '): continue

        if word.lower() not in dict1:

            dict1[word.lower()] = set()

            dict2[word.lower()] = 1

        else:

            dict2[word.lower()] += 1

        dict1[word.lower()].add(i+1)

再对dict1和dict2进行一定处理，首先按单词出现的频次从高到低排序，再删除top100的单词（去除掉类似the的无效信息），得到最终的倒排索引表。

output\_list = sorted(dict2.items(),key=lambda d:d[1], reverse=True)

output\_list\_delete\_top100 = output\_list[100:]

output\_sort = sorted(output\_list\_delete\_top100, key=lambda x:x[0])

Task 2：实现布尔查询

布尔查询依旧需要用到词典，刚才的倒排索引表可以直接作为词典使用。

with open("./dict.index.txt", 'r', encoding="UTF-8") as f:

    for line in f.readlines():

        line\_data = line.split( )

        query\_dict[line\_data[0]] = line\_data[1:]

布尔查询只要在单个单词查询的基础上，对两个及以上的查询结果进行交&或者并|的操作即可。

def query():

    while True:

        x = input("Search for：").split( )

        #print(len(x))

        if (len(x) == 1):

            x = x[0]

            if (x == "q"):

                print("------------END-----------")

                break

            if (x in query\_dict):

                q = query\_dict[x]

                print\_query\_ret(q)

            else:

                print("none")

        else:

            if (x[1] == 'and'):

#                 print(x)

                setx = set(query\_dict.get(x[0], []))

                sety = set(query\_dict.get(x[2], []))

                q = list(setx & sety)

                print\_query\_ret(q)

            elif (x[1] == 'or'):

                setx = set(query\_dict.get(x[0], []))

                sety = set(query\_dict.get(x[2], []))

                q = list(setx | sety)

                print\_query\_ret(q)

最后输出查询结果：

def print\_query\_ret(query\_ret):

    if (len(query\_ret) == 0):

        print("0")

        return

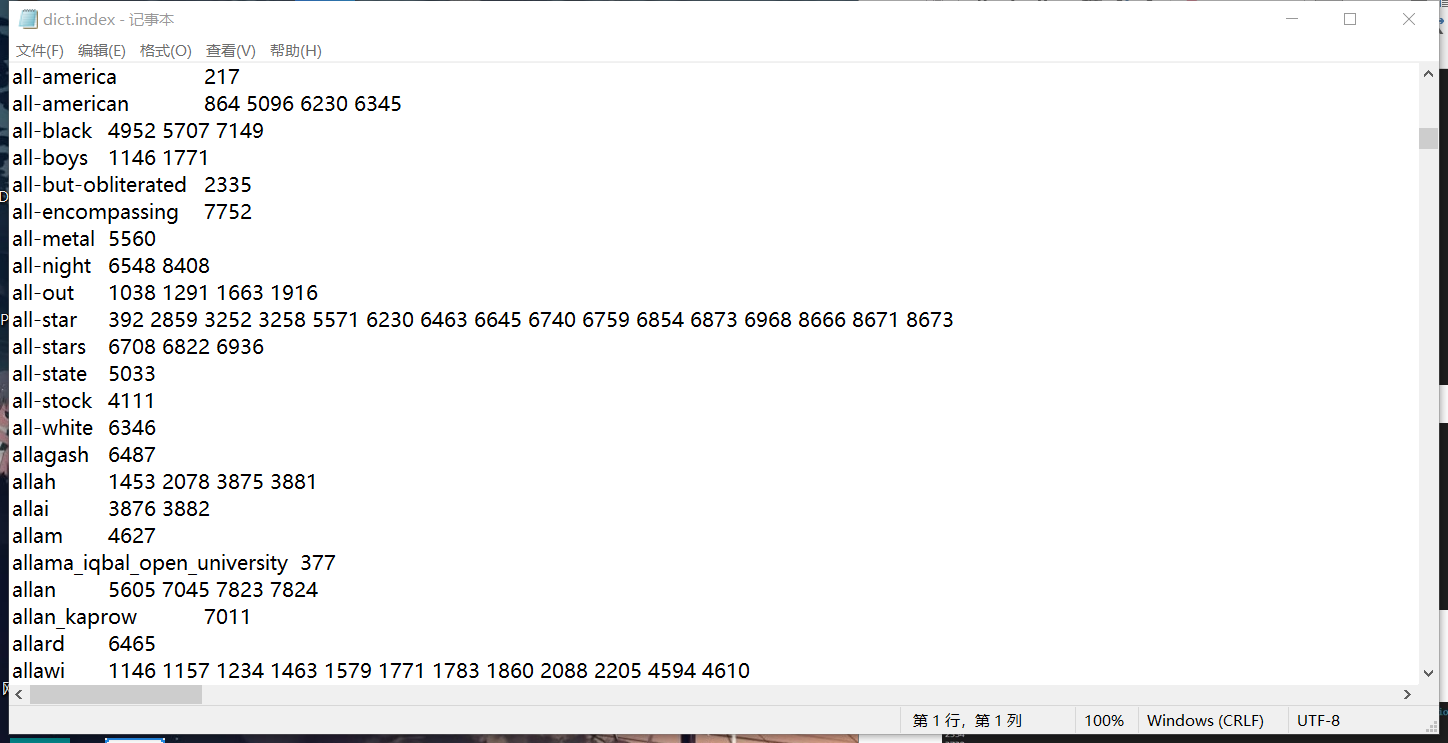
    query\_ret = sorted(query\_ret)

    for docID in query\_ret:

        print(docID.strip())

    print('\n')

运行结果：





1. **总结**

倒排索引应用于给定几个关键词，找出包含关键词的文档的场景。

倒排索引是实现“单词-文档矩阵”的一种具体存储式，通过倒排索引，可以根据单词快速获取包含这个单词的文档列表。

倒排索引主要由两个部分组成：“单词词典”和“倒排文件”。

单词词典(Lexicon)：搜索引擎的通常索引单位是单词，单词词典是由文档集合中出现过的所有单词构成的字符串集合，单词词典内每条索引项记载单词本身的一些信息以及指向“倒排列表”的指针。

倒排列表(PostingList)：倒排列表记载了出现过某个单词的所有文档的文档列表及单词在该文档中出现的位置信息，每条记录称为一个倒排项(Posting)。根据倒排列表，即可获知哪些文档包含某个单词。