README

Group 2

组员：刘佳润 陈诚

工程环境说明

- 编写语言：Python 3.7
- 操作系统：Windows 10 或 CentOS 7 均可运行
- 实验用CPU：AMD Ryzen 7 5800H

工程结构说明

```
IRProject
|  en_thesaurus.jsonl   // 未经处理的同义词表
|  BooleanQuery.py      // 布尔查询模块
|  GlobbingQuery.py     // 通配符查询模块
|  InvertedIndex.py     // 预处理模块
|  main.py              // 主程序入口
|  PhraseQuery.py       // 词语查询模块
|  README.md            // 本文件
|  SpellingCorrect.py   // 拼写检查模块
|  synonymslist.jsonl   // 经处理过的同义词表
|  SynonymsWords.py     // 同义词扩展模块
|  topk.py              // Top-K排序显示模块
|  utils.py             // 一些工具类函数
|
├─index
|      doc_size.json    // 预处理判断的有效文件目录
|      index.json       // 倒排索引
|      VSM.json         // VSM
|      wordlist.json    // 词表（词典）
|
└─Reuters               // 语料库
```

部署与运行方法

- 下载依赖库

```
$ pip install nltk
$ pip install chardet
$ pip install jsonlines
```

- 因为已经构建好索引，所以直接运行main.py，不需要任何改动

```
$ python main.py
```

功能测试与说明

构建倒排索引与VSM

这一部分构建倒排索引、构建词表、构建空间向量模型。为了加速后续的检索、排序等，我们在预处理阶段将所有文档的空间向量全部计算好。空间向量的每一项与词表中的位置对应。

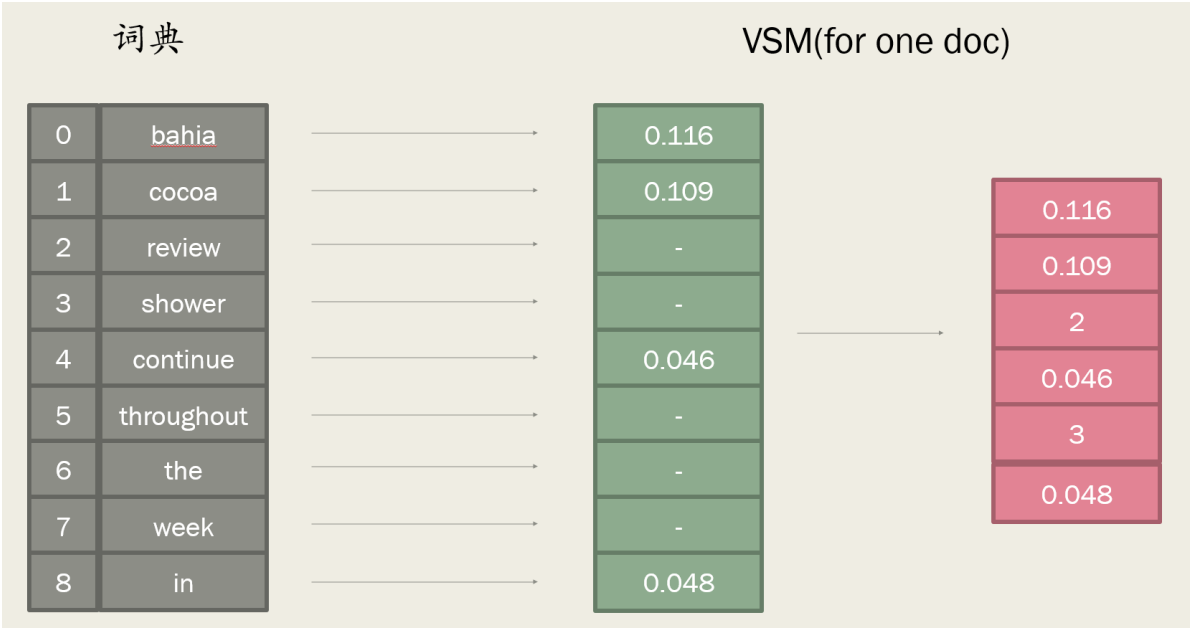
倒排索引结构如下：

```
{word1: {docID1: [pos1, pos2, …, posn], docID2 : [pos1, pos2, …, posn], …},
word2:…}
```

向量空间结构如下：

```
{docID1: [tfidf1, tfidf2, …, tfidfn], docID2 : [tfidf1, tfidf2, …, tfidfn], …}
```

向量空间压缩方法如下：



本来我们希望使用gzip进行压缩，以二进制的格式储存，但是发现python对于gzip解压缩的速度太慢，与我们的预期效果不符，因此我们最终放弃了这个方法，仅使用json来进行存储和读取。

测试结果：预处理时间总共1~2分钟

单词查询

进入方式：使用2.短语查询功能即可

```
************************
  欢迎使用文档搜索引擎!
      Ver. 1.1
************************
  请输入指令:
      1.布尔查询
      2.词语查询
      3.通配符查询
      0.退出系统
2
请输入查询:
education

是否需要进行同义词搜索？[y/n]
n
不进行同义词搜索!

1
显示排序前多少名的文档?
输入-1显示全部
15


************ 查询结果 ************

共找到  17  篇文档

显示前 15  篇文档

[10573, 1665, 9479, 5231, 10261, 1407, 7633, 7537, 11984, 12261, 18009, 21423, 246, 21368, 4552]

按Y查看文档，任意键跳过查看...
```

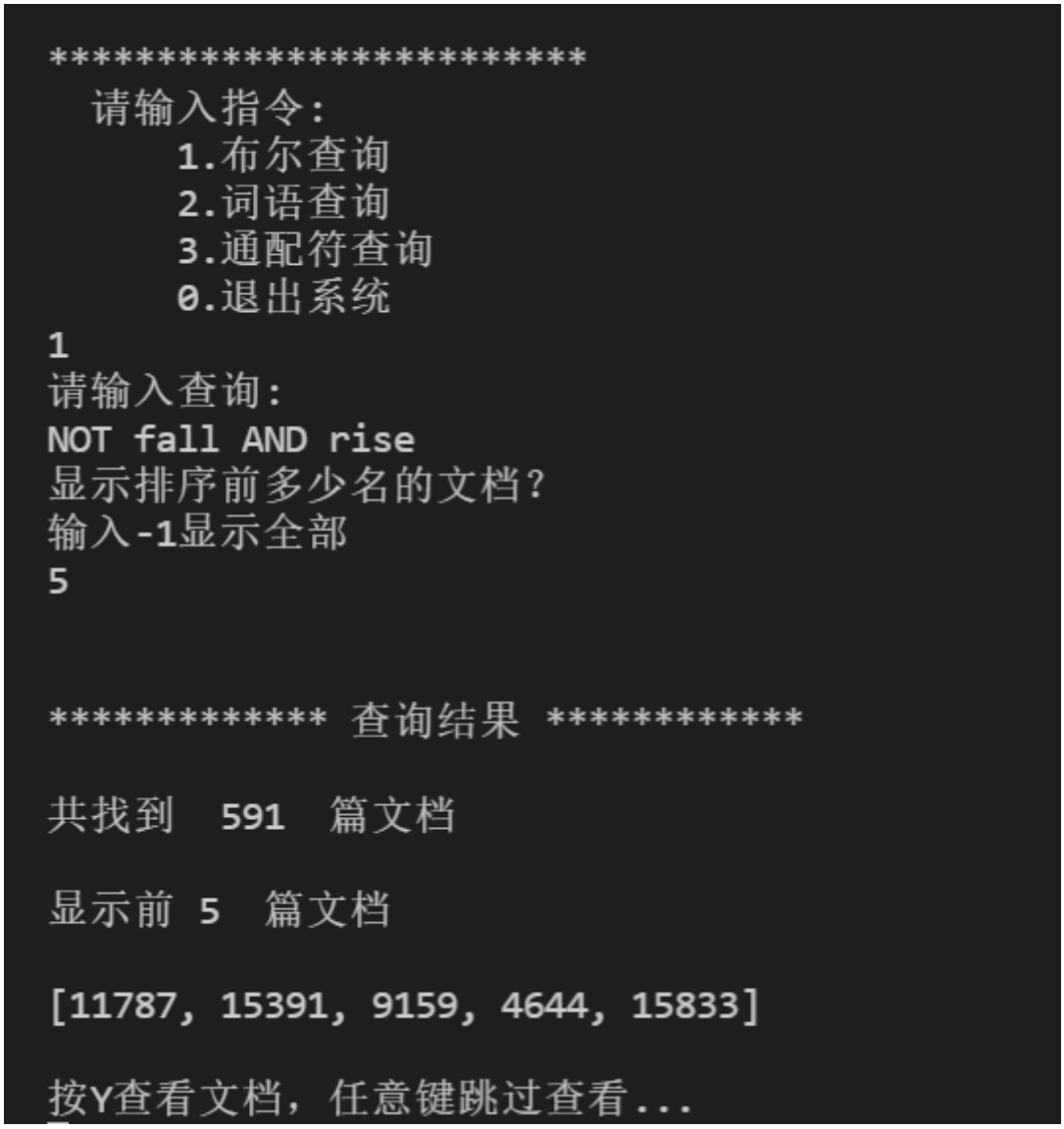
布尔查询

进入方式：使用1.布尔查询

语法：查询词使用正常拼写，布尔关系词使用全大写的AND/OR/NOT

```
************************
  欢迎使用文档搜索引擎!
      Ver. 1.1
************************
  请输入指令:
      1.布尔查询
      2.词语查询
      3.通配符查询
      0.退出系统
1
请输入查询:
government AND policy
显示排序前多少名的文档?
输入-1显示全部
15


************ 查询结果 ************

共找到  135  篇文档

显示前 15  篇文档

[6869, 338, 19982, 20013, 2709, 3449, 3528, 9293, 3372, 19500, 12121, 21497, 19947, 18095, 1948]

按Y查看文档，任意键跳过查看...
```

```
************************
  请输入指令:
      1.布尔查询
      2.词语查询
      3.通配符查询
      0.退出系统
1
请输入查询:
billion OR million
显示排序前多少名的文档?
输入-1显示全部
5


************ 查询结果 ************

共找到  1713  篇文档

显示前 5  篇文档

[3341, 17440, 17887, 19045, 7541]

按Y查看文档，任意键跳过查看...
y
************** 文档显示 **************
3341.html
U.S. CONSUMER CREDIT ROSE 536 MILLION DLRS IN JAN VS 144 MILLION DEC GAIN

  U.S. CONSUMER CREDIT ROSE 536 MILLION DLRS IN JAN VS 144 MILLION DEC GAIN



************** 文档显示 **************
17440.html
```



通配查询

进入方式：使用3.通配符查询

语法：使用*代替模糊部分

```
************************
  欢迎使用文档搜索引擎!
      Ver. 1.1
************************
  请输入指令:
      1.布尔查询
      2.词语查询
      3.通配符查询
      0.退出系统
3
请输入查询:
technol*
您可能想要找的是:
 ['technolgies', 'technolgy', 'technological', 'technologically', 'technologies', 'technology']
您是否要查看包含这些词的文档？ [y/n]
n
************************
  请输入指令:
      1.布尔查询
      2.词语查询
      3.通配符查询
      0.退出系统
3
请输入查询:
*formatio*
您可能想要找的是:
 ['formation', 'information', 'transformation', 'informations', 'formations', 'reformation']
```

短语查询

进入方式：使用2.短语查询

```
2
请输入查询:
commodity company

是否需要进行同义词搜索？[y/n]
n
不进行同义词搜索！

1
显示排序前多少名的文档？
输入-1显示全部
-1


************ 查询结果 ************

共找到  1  篇文档

显示前 1  篇文档

[1960]

按Y查看文档，任意键跳过查看...
y
************** 文档显示 **************
1960.html
CREDITOR BANKS MAY BUY INTO SINGAPORE COFFEE FIRM
  The nine creditor banks of the
  Singapore coffee trader <Teck Hock and Co (Pte) Ltd> are
  thinking of buying a controlling stake in the company
  themselves, a creditor bank official said.
      Since last December the banks have been allowing the
  company to postpone loan repayments while they try to find an
  overseas commodity company to make an offer for the firm.
      At least one company has expressed interest and
  negotiations are not yet over, banking sources said.
      However, the banks are now prepared to consider taking the
  stake if they find an investor willing to inject six to seven
  mln dlrs in the company but not take control, the banking
  sources said.
```

```
2
请输入查询:
information technology

是否需要进行同义词搜索？[y/n]
n
不进行同义词搜索！

1
显示排序前多少名的文档？
输入-1显示全部
-1


************ 查询结果 ************

共找到  2  篇文档

显示前 2  篇文档

[6537, 9734]

按Y查看文档，任意键跳过查看...
y
************** 文档显示 **************
6537.html
PEAT MARWICK AND NOLAN NORTON TO MERGE
  <Peat Marwick>, an accounting and
  management consulting firm, and <Nolan, Norton and Co>, an
  information and technology planning concern, said they have
  merged.
      The companies said with the merger Nolan now will be known
  as Nolan, Norton and Co-partners, the information technology
  arm of Peat Marwick.
      Also as part of the merger, Nolan's 21 principals have
  become Peat Marwick partners, the companies said.
```

拼写校正

进入方式：使用2.短语查询，对于可能输错的词，自动进行判断与提示

```
2
请输入查询:
govrnment
您要找的可能是: ['government', 'govenment']
您是否要使用替换词项进行搜索？[y/n]:
```

```
2
请输入查询:
gurantee
您要找的可能是: ['gaurantee', 'gurantees', 'guarantee']
您是否要使用替换词项进行搜索? [y/n]:
```

```
2
请输入查询:
kindom
您要找的可能是: ['kingdom']
您是否要使用替换词项进行搜索? [y/n]:
```

同义词查询

进入方式：使用2.短语查询，用户可以选择是否进行同义词查询

```
************************
  请输入指令:
      1.布尔查询
      2.词语查询
      3.通配符查询
      0.退出系统
2
请输入查询:
automobile

是否需要进行同义词搜索? [y/n]
y
开始执行同义词搜索:
发现同义词:  machine
发现同义词:  car
发现同义词:  auto
同义词搜索结束!

显示排序前多少名的文档?
输入-1显示全部
15


************ 查询结果 ************

共找到  116  篇文档

显示前 15  篇文档

[3125, 12580, 4820, 17688, 5500, 16152, 1263, 10960, 8998, 11909, 16868, 16868, 13711, 13711, 5559]

按Y查看文档，任意键跳过查看...
y
************** 文档显示 **************
```