Project 说明文档 – Information Retrieval

曹震东 3120104606 张瑞祥 3120104198 尹嘉权 3120000419 王宁 3120101836

功能列表

1. 倒排索引/向量空间模型
2. Top K
   - 堆（Heap）
   - 胜者表（Champion List）
   - 静态评分（Static Quality Score）
   - 簇剪枝（基于聚类，Cluster Pruning）
3. 布尔查询
4. 短语查询
5. 同义词查询
6. 拼写矫正

使用说明

编译说明

```
$ make
```

或

```
$ g++ Synonym.cpp Invertedindex.cpp Interpreter.cpp BoolQuery.cpp
PhraseQuery.cpp SpellingChecker.cpp StaticQualityScore.cpp TopK.cpp
ClusterPruning.cpp VectorSpaceModel.cpp ChampionList.cpp main.cpp -o main -
std=c++11
```

使用说明

运行编译出来的 exe 可执行文件 main.exe。

程序启动后会首先计算倒排索引，所以启动会慢一点。

```
E:\Code\IR\Information-Retrieval\SimpleInformationRetrievalTools\SimpleInformationRetrievalTools (master)
λ  main
Calculating Inverted Index...
Inverted Index Size: 53855
>
```

普通查询

计算完成后就可以进行查询操作了。在「>」符号后即可输入查询语句。若不带任何参数，则是默认按向量空间模型计算的余弦相似度排序。

```
> justice department
Search type: 0 TopK mode: 0 Synonym: 0
Time Elapsed: 22ms
---------------- TOP 50 RESULTS ----------------
docID: 10786.html score: 1 SQS: 0.162325
docID: 3984.html score: 1 SQS: 0.244404
docID: 6562.html score: 1 SQS: 0.197313
docID: 9470.html score: 1 SQS: 0.272835
docID: 2609.html score: 1 SQS: 0.183885
docID: 17700.html score: 1 SQS: 0.276567
docID: 2951.html score: 1 SQS: 0.174819
docID: 1920.html score: 1 SQS: 0.20607
docID: 3918.html score: 1 SQS: 0.183885
docID: 2475.html score: 1 SQS: 0.255023
docID: 5919.html score: 1 SQS: 0.185733
docID: 5517.html score: 1 SQS: 0.176343
docID: 1994.html score: 1 SQS: 0.223045
```

布尔查询

若语句中含有 AND、OR、NOT（大小写敏感），则认为是布尔查询。

```
> justice AND department
Search type: 1 TopK mode: 0 Synonym: 0
Time Elapsed: 85ms
---------------- TOP 50 RESULTS ----------------
docID: 6562.html score: 0 SQS: 0.197313
docID: 5919.html score: 0 SQS: 0.185733
docID: 1718.html score: 0 SQS: 0.244091
docID: 10895.html score: 0 SQS: 0.189209
docID: 3918.html score: 0 SQS: 0.183885
docID: 2951.html score: 0 SQS: 0.174819
docID: 17700.html score: 0 SQS: 0.276567
docID: 1994.html score: 0 SQS: 0.223045
docID: 2475.html score: 0 SQS: 0.255023
docID: 5643.html score: 0 SQS: 0.229885
docID: 2948.html score: 0 SQS: 0.225527
docID: 1920.html score: 0 SQS: 0.20607
docID: 3711.html score: 0 SQS: 0.224551
docID: 2609.html score: 0 SQS: 0.183885
docID: 2251.html score: 0 SQS: 0.252114
docID: 3984.html score: 0 SQS: 0.244404
```

短语查询

若要开启短语查询，在查询语句后加参数「–PHRASE_SEARCH」即可。

```
> justice department -PHRASE_SEARCH
Search type: 2 TopK mode: 0 Synonym: 0
Time Elapsed: 6ms
---------------- TOP 50 RESULTS ----------------
docID: 10786.html score: 0 SQS: 0.162325
docID: 10895.html score: 0 SQS: 0.189209
docID: 1718.html score: 0 SQS: 0.244091
docID: 17700.html score: 0 SQS: 0.276567
docID: 1920.html score: 0 SQS: 0.20607
docID: 1920.html score: 0 SQS: 0.20607
docID: 1920.html score: 0 SQS: 0.20607
docID: 1994.html score: 0 SQS: 0.223045
docID: 1994.html score: 0 SQS: 0.223045
```

Top K 查询

支持四种查询方式，类似的，只要在 query 后加参数指定即可。

```
> justice department -TOP_K_HEAP
Search type: 0 TopK mode: 1 Synonym: 0
Time Elapsed: 23ms
---------------- TOP 50 RESULTS ----------------
docID: 3984.html score: 1 SQS: 0.244404
docID: 17700.html score: 1 SQS: 0.276567
docID: 10786.html score: 1 SQS: 0.162325
docID: 10895.html score: 1 SQS: 0.189209
docID: 5919.html score: 1 SQS: 0.185733
docID: 2609.html score: 1 SQS: 0.183885
docID: 2475.html score: 1 SQS: 0.255023
docID: 2251.html score: 1 SQS: 0.252114
docID: 2376.html score: 1 SQS: 0.224304
docID: 2951.html score: 1 SQS: 0.174819
```

参数表为：

参数	开启的功能
-TOP_K_HEAP	堆
-TOP_K_CHAMPION_LIST	胜者表（Champion List）
-TOP_K_STATIC_QUALITY_SCORE	静态评分（Static Quality Score）
-TOP_K_CLUSTER_PRUNING	簇剪枝（基于聚类，Cluster Pruning）

同义词查询

使用参数「-SYNONYM_ON」即可开启同义词查询。

拼写矫正

程序自动检测，无手动开关。如下图所示：

如上图 justice 我故意打成了 jusitcem，拼写矫正程序会自动发现并矫正。

其他输出结果中，各参数的含义如下表所示：

Search Type	含义
0	普通查询
1	布尔查询
2	短语查询

TOP_K_MODE	含义
0	TOP K 关闭
1	堆
2	胜者表（Champion List）
3	静态评分（Static Quality Score）
4	簇剪枝（基于聚类，Cluster Pruning）

Synonym	含义
0	同义词检索关闭
1	同意词检索开启