第六周实验报告

1. 实验概览

本实验旨在于理解 TD-IDF 技术如何评估特定字词对于一个文件集或一个语料库中的其中一份文件的重要程度，了解 IDF 系数的多种常见形式，理解 BM25 的原理和基本算法，基于 Lucene 中的 Similarity 类构建自定义相关性函数。

2. 实验环境

Docker: SJTU-EE208

3. 解决思路

设计一：

```python
class SimpleSimilarity(PythonClassicSimilarity):

    def lengthNorm(self, numTerms):
        return 1 / math.sqrt(numTerms)

    def tf(self, freq):
        return freq

    def sloppyFreq(self, distance):
        return 1 / (distance + 1)

    def idf(self, docFreq, numDocs):
        return math.log((numDocs - docFreq + 0.5) / (docFreq + 0.5))

    def idfExplain(self, collectionStats, termStats):
        return Explanation.match(1.0, "inexplicable", [])
```

其中，将词频直接设置为 TF 得分，IDF 得分的计算公式基于 $\log\frac{N-n_t}{n_t}$对数公式，分子分母分别增加一个常数防止出现零分母情况

设计二：

```python
class SimpleSimilarity2(PythonClassicSimilarity):

    def lengthNorm(self, numTerms):
        return 1 / numTerms

    def tf(self, freq):
        return math.sqrt(freq)

    def sloppyFreq(self, distance):
        return 1 / (distance + 1)

    def idf(self, docFreq, numDocs):
        return math.log((numDocs + 1) / (docFreq + 1))

    def idfExplain(self, collectionStats, termStats):
        return Explanation.match(1.0, "inexplicable", [])
```

整体设计和一类似，微调了部分计算公式（是否添加开方运算等）

4. 实验结果

default:

```
Searching for: london author:shakespeare
6 total matching documents.
path: testfolder/pg1342.txt name: pg1342.txt score: 0.41727691888809204
path: testfolder/pg1661.txt name: pg1661.txt score: 0.41477009654045105
path: testfolder/pg16328.txt name: pg16328.txt score: 0.368636816740036
path: testfolder/pg27827.txt name: pg27827.txt score: 0.36713215708732605
path: testfolder/pg76.txt name: pg76.txt score: 0.33648228645324707
path: testfolder/pg30601.txt name: pg30601.txt score: 0.21114353835582733
```

Sim1:

```
Searching for: london author:shakespeare
6 total matching documents.
path: testfolder/pg30601.txt name: pg30601.txt score: -0.0024969379883259535
path: testfolder/pg76.txt name: pg76.txt score: -0.010966422036290169
path: testfolder/pg16328.txt name: pg16328.txt score: -0.014577941969037056
path: testfolder/pg27827.txt name: pg27827.txt score: -0.014981627464294434
path: testfolder/pg1661.txt name: pg1661.txt score: -0.07776540517807007
path: testfolder/pg1342.txt name: pg1342.txt score: -0.097117580473423
```

Sim2:

```
Searching for: london author:shakespeare
6 total matching documents.
path: testfolder/pg1342.txt name: pg1342.txt score: 2.152226079488173e-05
path: testfolder/pg1661.txt name: pg1661.txt score: 2.14404226426268e-05
path: testfolder/pg16328.txt name: pg16328.txt score: 1.7691869288682938e-05
path: testfolder/pg27827.txt name: pg27827.txt score: 1.4214364455256145e-05
path: testfolder/pg76.txt name: pg76.txt score: 7.616218681505416e-06
path: testfolder/pg30601.txt name: pg30601.txt score: 5.802989562653238e-06
```