电工导 C 第六次实验报告

姓名: 宋士祥　　学号:521030910013　　班级:F2103001

2022 年 10 月 26 日

1　实验概览

本次实验介绍了 TF-IDF。TF-IDF 是一种用于信息检索与文本挖掘的加权计数方法。
一般地，我们有

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D)$$

其中 t 为我们查询的单词，d 为某篇文档，D 为所有文档组成的语料库。
TF——“单词频率”：计算一个查询关键字中某一个单词在目标文档中出现的次数
IDF——“逆文档频率”：惩罚出现在太多文档中的单词。(e.g. a, an, the, of)
本次实验即是根据上述内容自我定义 Simiality 类，具体需要修改函数：

- `lengthNorm`
- `tf`
- `sloppyFreq`
- `idf`
- `idfExplain`

2　实验环境

本次实验采用所需的实验环境如下：

- Docker 中的`sjtumic/ee208` 镜像
- Python3（使用 VSCode 编译）
- `BeautifulSoup4` 扩展以及`lxml` 扩展。
- `java` 环境及`lucene` 扩展（在 SJTU EE208 中已经给出）。

3　问题重述与代码说明

3.1　lengthNorm

lengthNorm 要求我们返回长度范数。
我们根据 BM25 中的定义，直接返回：

$$f(distance) = \frac{1}{\sqrt{distance}}$$

3.2 tf

根据定义，我们是需要返回文档长度在某处出现的频率。直接返回查询到的数字个数即可。

然而对于比较长的文档，这样计算可能返回值过大，我们也可以采用 $f(x) = \log(1 + x)$ 的方式进行处理。

3.3 sloppyFreq

根据查询文档，长度范数要求控制在 $[0, 1]$ 的区间内，且长度越长，长度范数的值应当越小。以下两个函数均符合我们的要求：

$$f(x) = \mathrm{e}^{-x}$$

$$f(x) = \frac{1}{1 + x}$$

其中 BM25 采取的是后者。

3.4 idf

根据定义，idf 是对出现的过多频率文档的处罚。因此我们可以计算某个词的平均出现概率，如果这个词的平均出现概率很大，那说明这个词很有可能不包含什么信息量。我们有：

$$idf = \frac{numDoc}{docFreq}$$

该式也可直接取对数。

3.5 idfExample

为内置函数，从类中返回对应的值即可。

根据我们上述的说明，我们本次实验分别采取了使用上述函数组合后的 Similarity 类，分别代码如下：

Similarity1:

```
class SimpleSimilarity(PythonClassicSimilarity):

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

    def tf(self, freq):
        return math.log(1+freq)

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

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

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

Similarity2:

```python
class SimpleSimilarity(PythonClassicSimilarity):

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

    def tf(self, freq):
        return freq

    def sloppyFreq(self, distance):
        return math.exp(-distance)

    def idf(self, docFreq, numDocs):
        return (numDocs/docFreq)

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

4 运行结果

我们利用我们刚才所写的内容来运行 Lab4 的代码。我们分别对原始代码和重写的两个类的代码进行运行，结果如下：

```
Hit enter with no input to quit.
Query:
root@66fb19921937:/workspaces/lab4-Lucene/codes# python SearchFiles.py
lucene 8.6.1

Hit enter with no input to quit.
Query:上海交大

Searching for: 上海交大
50 total matching documents.
path: html/httpswww.163.comeduarticleHAHLEB3M00297VGM.html.txt
name: httpswww.163.comeduarticleHAHLEB3M00297VGM.html.txt
url: https://www.163.com/edu/article/HAHLEB3M00297VGM.html
title: C9、E9、华五、中九、五院四系、二龙四虎……这些黑话你懂吗？|志愿填报|大学_网易教育
------
path: html/httpbiz.finance.sina.com.cnfuturesask.txt
name: httpbiz.finance.sina.com.cnfuturesask.txt
url: http://biz.finance.sina.com.cn/futures/ask/
title: 期货专家坐堂_首页_财经频道_新浪网
------
path: html/httpswww.163.comeduarticleH9DILVH200297VGM.html.txt
name: httpswww.163.comeduarticleH9DILVH200297VGM.html.txt
url: https://www.163.com/edu/article/H9DILVH200297VGM.html
title: 清华调整强基计划方案：初试分省计算机考试，复试分区域|清华大学_网易教育
------
path: html/httpedu.sina.com.cn.txt
name: httpedu.sina.com.cn.txt
url: http://edu.sina.com.cn/
title: 教育频道_新浪教育_新浪网
------
path: html/httpzhiyuan.edu.sina.cnvt4pos108.txt
name: httpzhiyuan.edu.sina.cnvt4pos108.txt
url: http://zhiyuan.edu.sina.cn/?vt=4&pos=108
title: 教育频道_新浪教育_新浪网
------
```

```
Hit enter with no input to quit.
Query:上海交大

Searching for: 上海交大
50 total matching documents.
path: html/httpswww.163.comeduarticleHAHLEB3M00297VGM.html.txt
name: httpswww.163.comeduarticleHAHLEB3M00297VGM.html.txt
url: https://www.163.com/edu/article/HAHLEB3M00297VGM.html
title: C9、E9、华五、中九、五院四系、二龙四虎……这些黑话你懂吗？|志愿填报|大学_网易教育
------
path: html/httpbiz.finance.sina.com.cnfuturesask.txt
name: httpbiz.finance.sina.com.cnfuturesask.txt
url: http://biz.finance.sina.com.cn/futures/ask/
title: 期货专家坐堂_首页_财经频道_新浪网
------
path: html/httpswww.163.comeduarticleH9DILVH200297VGM.html.txt
name: httpswww.163.comeduarticleH9DILVH200297VGM.html.txt
url: https://www.163.com/edu/article/H9DILVH200297VGM.html
title: 清华调整强基计划方案：初试分省计算机考试，复试分区域|清华大学_网易教育
------
path: html/httpedu.sina.com.cn.txt
name: httpedu.sina.com.cn.txt
url: http://edu.sina.com.cn/
title: 教育频道_新浪教育_新浪网
------
path: html/httpzhiyuan.edu.sina.cnvt4pos108.txt
name: httpzhiyuan.edu.sina.cnvt4pos108.txt
url: http://zhiyuan.edu.sina.cn/?vt=4&pos=108
title: 教育频道_新浪教育_新浪网
------
path: html/httpsv.qq.comxpagep004418tr7x.html.txt
name: httpsv.qq.comxpagep004418tr7x.html.txt
url: https://v.qq.com/x/page/p004418tr7x.html
title: 上海交响乐团2022-23乐季开幕！埃尔加刮起英伦风_高清1080P在线观看平台_腾讯视频
root@66fb19921937:/workspaces/lab4-Lucene/codes# python SearchFiles2.py
lucene 8.6.1

Hit enter with no input to quit.
Query:上海交大

Searching for: 上海交大
50 total matching documents.
path: html/httpsv.qq.comxpagep004418tr7x.html.txt
name: httpsv.qq.comxpagep004418tr7x.html.txt
url: https://v.qq.com/x/page/p004418tr7x.html
title: 上海交响乐团2022-23乐季开幕！埃尔加刮起英伦风_高清1080P在线观看平台_腾讯视频
------
path: html/httpswww.163.comeduarticleHAHLEB3M00297VGM.html.txt
name: httpswww.163.comeduarticleHAHLEB3M00297VGM.html.txt
url: https://www.163.com/edu/article/HAHLEB3M00297VGM.html
title: C9、E9、华五、中九、五院四系、二龙四虎……这些黑话你懂吗？|志愿填报|大学_网易教育
------
path: html/httpbiz.finance.sina.com.cnfuturesask.txt
name: httpbiz.finance.sina.com.cnfuturesask.txt
url: http://biz.finance.sina.com.cn/futures/ask/
title: 期货专家坐堂_首页_财经频道_新浪网
------
path: html/httpswww.163.comnewsarticleHC64334D00018AP2.html.txt
name: httpswww.163.comnewsarticleHC64334D00018AP2.html.txt
url: https://www.163.com/news/article/HC64334D00018AP2.html
title: 谁会成为英国新首相？|工党|英国议会|党首|竞选人_网易新闻
------
path: html/httpsv.qq.comxcover21z0kvwqvopae1s.html.txt
name: httpsv.qq.comxcover21z0kvwqvopae1s.html.txt
url: https://v.qq.com/x/cover/21z0kvwqvopae1s.html
title: 记忆大师_电影_高清1080P在线观看平台
------
```

5 问题探究与拓展思考

5.1 搜索结果的差异及其成因

这三种写法的搜索结果均不完全一致。其中，我们的第一种写法和 CJK 的自带类比较接近，而第二种写法则体现出来一定的差距。

我们对比时会发现，第二种写法放大了 idf 而减小了 tf。因此这一组会对出现较多的词比较敏感。我们发现，第二组对于“交大”这两个词较为敏感，第一个结果和第三个结果均含“上海交”词组，这一点可以说明我们第二种算法对一些词组的敏感性过强。这种适合一些专业词的搜索。而对于第一种搜索方式，精确度会相对较好，但是会引入一些不必要的答案。这比较适合大型的搜索引擎。

5.2 BM25 的计算方法

我们知道，lucene 默认采用了 BM25 作为处理。BM25 相较于我们的计算方法，有以下优点：

- 其 idf 的计算避免了特殊性，函数：

$$IDF(q_i) = \log\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}\right)$$

- 其采用了加权计算的方法，使得权重更准确。

5.3 修改 Similarity 的意义

很多时候，原生的 Similarity 类并不符合我们的需求，例如我们可能想要避免“谷歌炸弹”的情形出现，此时我们需要修改 Similarity 来适应不同的需求。