HW3 向量空间模型 (Vector Space Model)

作业描述

给定查询文档集合（**诗词**txt文件），完成向量空间模型并对文档集合实现查询功能。

作业要求

1. 实现带域的查询功能，具体为（**诗**名、作者、**诗句**）三个域，要求实现自定义组合域中的查询。如只在歌名中进行检索，或者只在歌名和歌词中进行检索；
2. 完成代码后编写实验报告，主要阐述代码细节以及原理；
3. 录制演示视频，举例演示查询功能，1-5分钟即可；
4. 编程语言不做限制；
5. 完成作业后将源码、实验报告以及演示视频放到一个文件夹下打包，命名格式为学号_姓名_hw3。并于2022.11.9日晚23:59之前提交到邮箱nkulxb2022@163.com。

实验讲解

给定文档集合如下：

文档1：d b a

文档2：c a e

查询：a b

首先对词项进行排序（字母序列），统一成小写，单词变成原型（这里不做强制要求，因为大家可能对NLP的包不太了解，知道即可）：a b c d e

然后计算每个词项的在不同文档中的词频tf(Term Frequency)如下表： 这里为了减少文本长度带来的影响，使用log来减小词频的影响。$TF = log_{10}(N+1)$，其中N表示词项在对应文档中出现的次数。

term	**文档一**	**文档二**	**查询**
a	$log_{10}(2)$	$log_{10}(2)$	$log_{10}(2)$
b	$log_{10}(2)$	0	$log_{10}(2)$
c	0	$log_{10}(2)$	0
d	$log_{10}(2)$	0	0
e	0	$log_{10}(2)$	0

计算每个词项在整个文档集合的逆向文件频率idf(Inverse Document Frequency)

$IDF_t = log_{10}(\frac{N}{df_t})$，其中$N$表示文档总数，$df_t$表示包含词项$t$的文档数。

ps：因为我们是根据文档集合进行统计的，因此df（包含词条w的文档数）不会为0，因此分母不需要加1。

term	df	idf
a	2	0
b	1	$log_{10}(2)$
c	1	$log_{10}(2)$
d	1	$log_{10}(2)$
e	1	$log_{10}(2)$

因此如果用向量来表示文档，其中向量的每一项用词项的tf*idf来表示，可以得到文档以及查询的向量表示如下:
文档一：(0, $log_{10}(2) * log_{10}(2)$, 0, $log_{10}(2) * log_{10}(2)$, 0)

文档二：(0, 0, $log_{10}(2) * log_{10}(2)$, 0, $log_{10}(2) * log_{10}(2)$)

查询：(0, $log_{10}(2) * log_{10}(2)$, 0, 0, 0)

将查询语句的向量分别与每个文档的向量计算余弦相似度，相似度越高则说明越相关，返回的结果也更靠前。