第六周实验报告

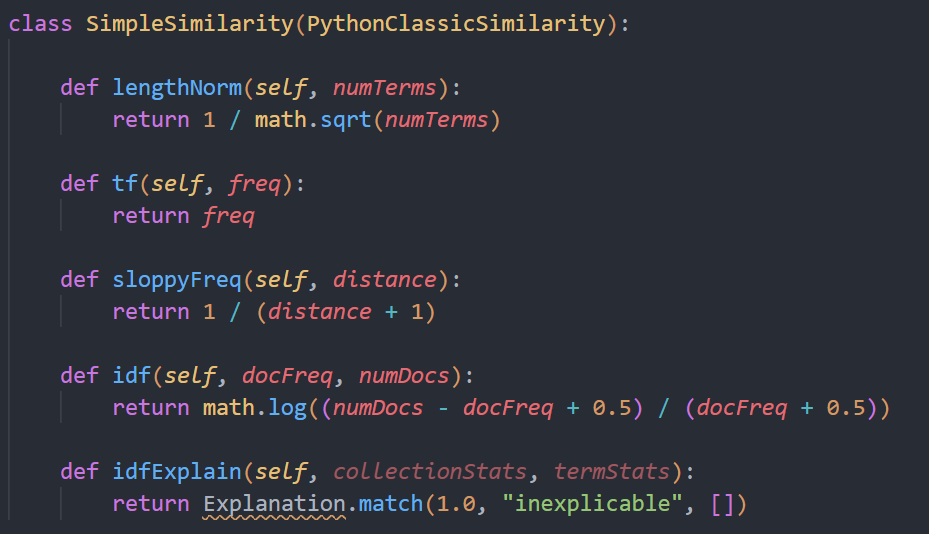
1. 实验概览

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。

1. 实验环境

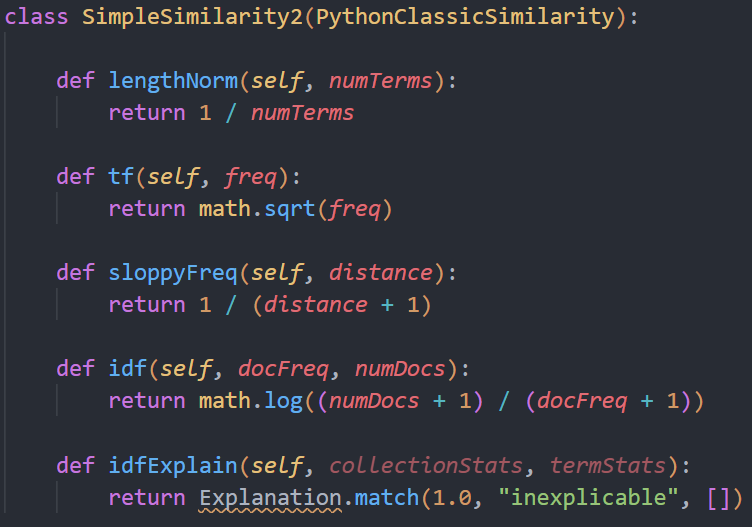
Docker: SJTU-EE208

1. 解决思路

**设计一：**

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

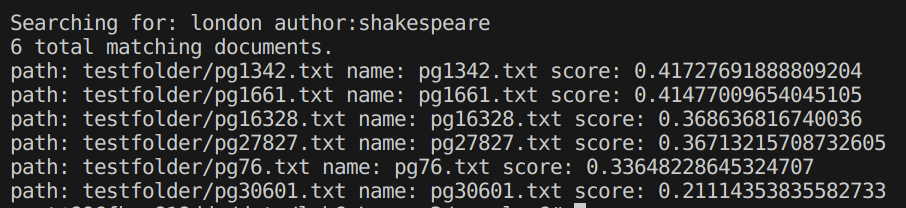
**设计二：**

****

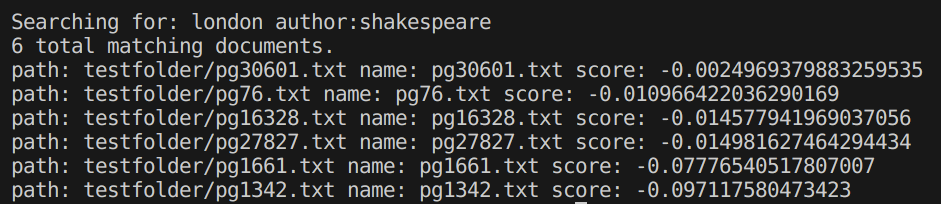
整体设计和一类似，微调了部分计算公式（是否添加开方运算等）

1. 实验结果

default:



Sim1:



Sim2:

