**需求**

题目：论文查重

描述如下：

设计一个论文查重算法，给出一个原文文件和一个在这份原文上经过了增删改的抄袭版论文的文件，在答案文件中输出其重复率。

* 原文示例：今天是星期天，天气晴，今天晚上我要去看电影。
* 抄袭版示例：今天是周天，天气晴朗，我晚上要去看电影。

要求输入输出采用文件输入输出，规范如下：

* 从**命令行参数**给出：论文原文的文件的**绝对路径**。
* 从**命令行参数**给出：抄袭版论文的文件的**绝对路径**。
* 从**命令行参数**给出：输出的答案文件的**绝对路径**。

我们提供一份样例，课堂上下发，上传到班级群，使用方法是：orig.txt是原文，其他orig\_add.txt等均为抄袭版论文。

注意：答案文件中输出的答案为浮点型，精确到小数点后两位

说明一下程序流程，有三个函数，分别是文本切分、特征构建和量度距离，主函数获得文件内容后，先将两个字符串进行文本切分（省略空白字符），然后对得出来的结果放入特征构建之中，再把得到的数组分别进行量度，最后进行简单的运算即可得出相似度。

**文本切分**

预计原型：std::vector<std::string> textSegmentation(std::string s)

传入字符串，返回其文本切分后的结果，暂定使用unigram进行分词，视情况也会使用bigram。

**特征构建**

预计原型：std::vector<int> vecConstruct(std::vector<std::string>)

传入文本切分后的数组，返回一个向量，暂定使用独热表示方法，即每一个词使用一个独特的长度为1的向量，各向量正交，也就是无关。

**量度距离**

预计原型：double calcDistance(std::vector<int>)

传入特征向量，返回其离原点的距离，欧拉距离或者余弦距离都行。

**主函数**

主函数需要处理参数的文件名，读入字符串，文本切分后对全部文本进行离散化，还有最后通过距离计算相似度。