## 算法方面：

思路1：文本长度相差较大时默认相似度很低。

若两字符串分词的词语数组中两者相差一定比例以上时默认相似度很低。

在实验中比例设为较长者的1/5

max\_len = max(len(words1), len(words2))

var\_len = 0.2

var\_sim = 0.1

if abs(len(words1) - len(words2)) > int(max\_len\*var\_len):

return var\_sim

思路2：不输出完全相同的语句，取相似度的下界和上界之间的语句

lower\_bound = 0.7

upper\_bound = 0.99

## 输出文件格式

将12个excel文件进行审查内容比较并将相似度满足要求的语句导出在文件比较汇总.xlsx

中

## 针对excel表格的处理和分析

1. 打开后设置行高和列宽分别为50
2. 设置自动换行
3. Excel格式分析



语句包含相似率的语句是与下面相邻的语句做对比的

包含相似率的语句是在行名所在文件的

不包含相似率的语句是在列名所在文件的

## 待完善方面

1. 分词不会自动去除空格等字符
2. 分词的策略有待优化
3. 审查内容解析不彻底，混入一些目录的内容