1.任务目标

给定一篇（些）论文，判断其是否为真论文，即非scigen生成的论文。

2.数据预处理

(一) 获取训练样本

正样本来源：由老师给定的三千多论文样本

负样本来源：于MIT scigen上爬取下来的论文，原打算爬三千多篇论文下来，由于有大量重复，最后只有一千六百多篇。

代码：crawelFalse.py

（二）转换pdf

用linux自带的pdftotext将pdf转化为txt

在pdf转化为txt已经自动进行了去除图和表格

（三）处理txt

去除标题和作者信息

去除Reference 信息

去除论文引用的标号 []

将其他括号，标点符号替换为空

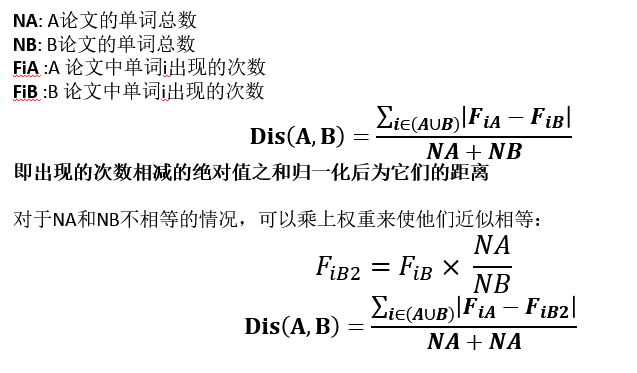
所有单词转化为小写

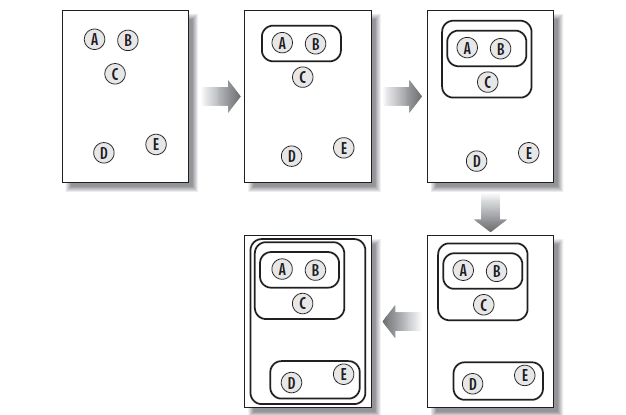
数字归一化处理

代码：processTxt.py

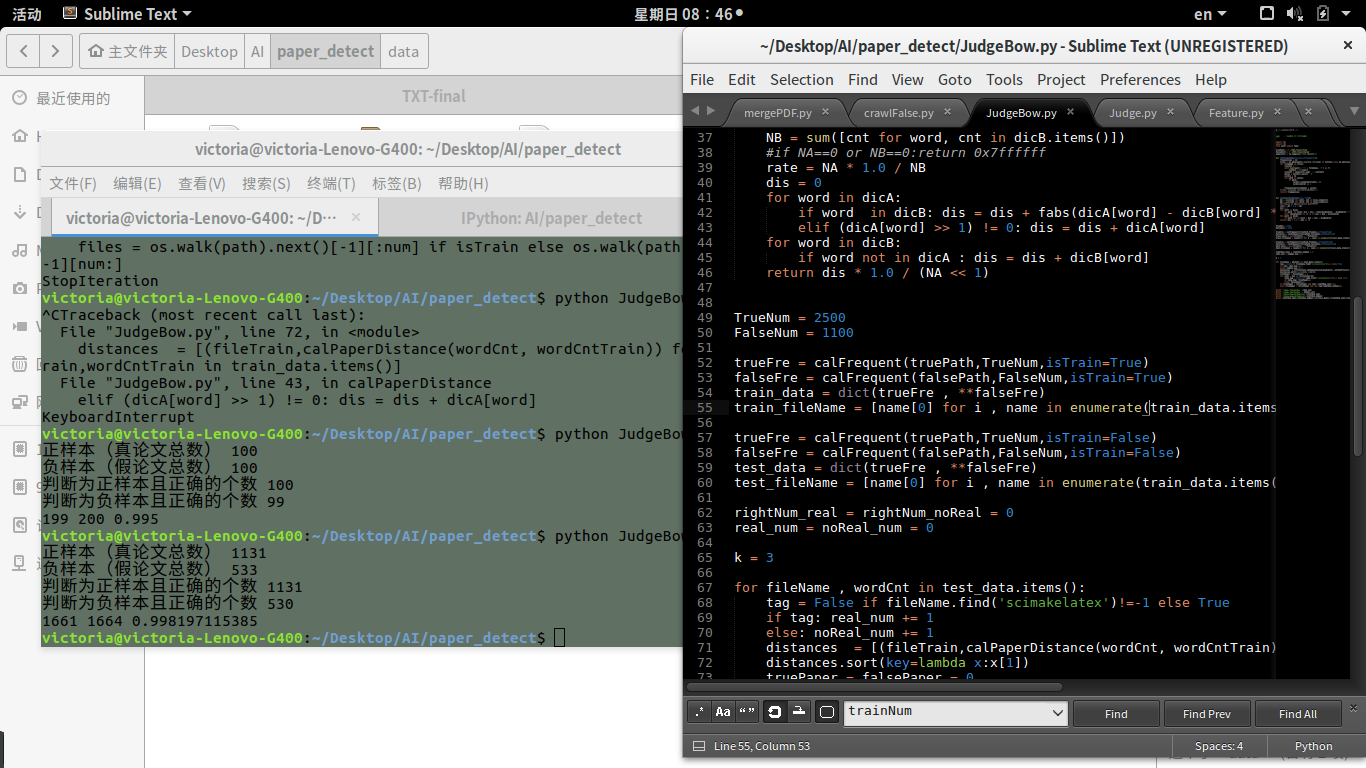
3.判定

利用文档距离作为聚类标准，对于一篇论文直接找最近距离的效果并不是很好，而是用类似KNN的思想，求出距离后，进行排序，看更接近哪个类。





代码：JudgeBow.py



可以看出效果已经非常的好