1.任务目标

给定一篇（些）论文，判断其属于哪一领域的论文

2.数据预处理

(一) 获取训练样本

样本来源：网上收集的四大类计算机领域的论文

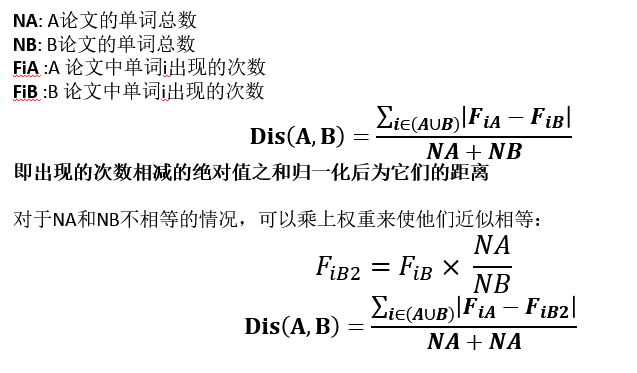
（二）转换pdf

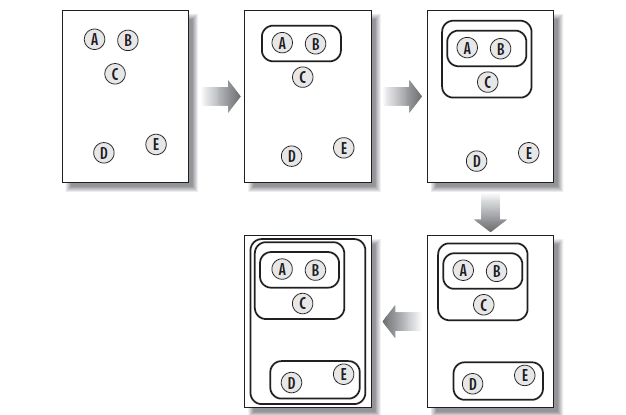
用linux自带的pdftotext将pdf转化为txt

在pdf转化为txt已经自动进行了去除图和表格

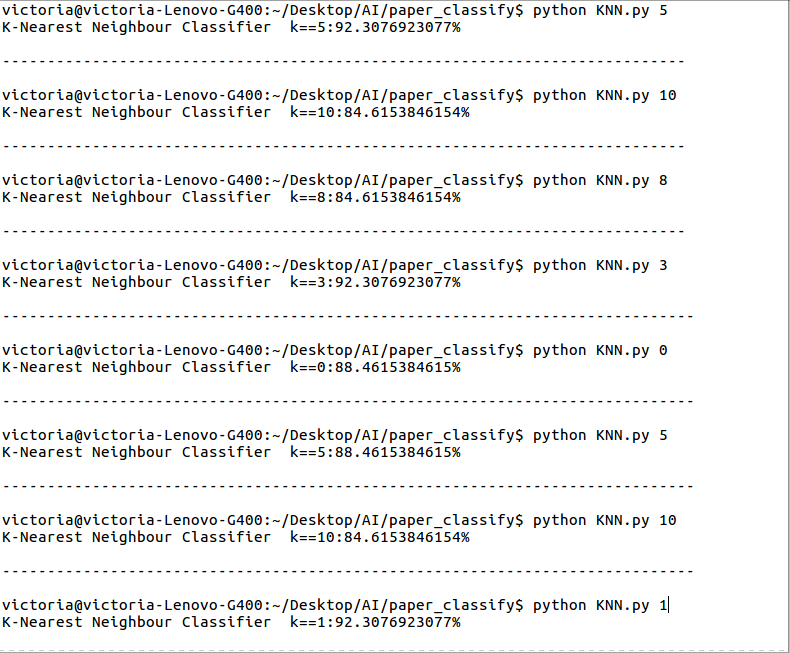
3.判定

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：KNN.py +最低被计入词汇表的单词频率（不加默认为0）



可以看出效果已经非常的好