实验报告

一、实验内容

利用 VSM 和 KNN 对文档进行处理、分类。

二、实验步骤

1、将文档分为测试集和训练集，比例为 1:4；

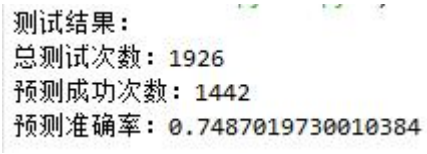
2、读取训练集与测试集文件，对文件进行分词、统一大小写等预处理，形成词典；

3、设置阈值，筛选具有代表性的数据，用 VSM 矢量化每个文档数据；

4、用 KNN 对测试集数据进行分类预测，与实际类别对比，计算准确率。

三、实验结果

k = 5　vector > 50 时



四、总结

1、实验的数据过大，选择了部分数据，并将权重的阈值调高一些，防止字典过大难以运行。

2、KNN 分类的时间比较长，这个问题还未解决，希望能获得好的方法。

3、分类准确率不高，需要继续调整参数以及优化程序来提高分类性能。