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、进行 Naïve Bayes 分类

三、使用方法介绍

1、朴素贝叶斯算法对条件概率分布作出了独立性的假设，通俗地讲就是说假设各个维度的特征 x1,x2,...,xn 互相独立，朴素贝叶斯分类器可表示为：

$$f(x) = argmax_{y_k} P(y_k|x) = argmax_{y_k} \frac{P(y_k)\prod_{i=1}^{n} P(x_i|y_k)}{\sum_k P(y_k)\prod_{i=1}^{n} P(x_i|y_k)}$$

多项式模型在计算先验概率 P(yk)和条件概率 P(xi|yk)时，会视重复的词语为出现多次，统计判断时，重视重复词语。

2、平滑技术

四、实验结果

```
测试结果：
总测试次数：1926
预测成功次数：1309
预测准确率：
0.6796469366562825
```

五、总结

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。