实验报告

1. 数据预处理

创建新的文件夹 data_make 存放数据预处理之后的文档。读取 20news-18828 文件夹中的文档，利用 nltk 相关工具对文档进行分词处理，然后将大写变为小写，去停用词，并对单词进行词干提取操作（不考虑顺序）。

2. 统计词频并建立字典

从已经预处理过的文档中读取数据并统计词频，由于数据量过大，只统计词频大于 4 的单词。建立字典 allword。

3. 划分数据集

将数据按照 8:2 的比例将原数据集划分成训练集与测试集，percent=0.8。随机生成 0～5 的数字 index，利用 index*(文档数目*（1-percent）)来确定测试集选取的起始位置，以此保证每次选取的测试集都不同，最后运行的结果也不同。文件夹 Test 存放测试集，文件夹 Train 存放训练集。

4. bayes 算法

4.1 算法思想

设 x={a1,a2,…,am}为一个待分类项，而每个 a 为 x 的一个特征属性。

有类别集合 C={y1,y2,…,yn}。

计算 P(y1|x) ,P(y2|x),…, P(yn|x)。

如果 P(yk|x) =max{P(y1|x),P(y2|x),…, P(yn|x)}，则 x∈yk

4.2 程序运行

1.计算条件概率与先验概率：

条件概率=类 k 中单词 c 的个数/（类 k 中的单词总数+训练样本中总的单词数目）

先验概率=类 k 中单词总数/训练样本中总的单词数目

2.对划分好的测试集进行分类，将结果存在 result.txt 中。

4.3 计算正确率

正确率=分类正确的测试集数目/进行分类的总的数目

5. 实验总结

在实验一中的划分数据集没有做到很灵活，在这里通过设置一个随机数来确定划分测试集的起始位置，每次程序运行保证测试集的不同。

由于数据量过大，通过限制词频来控制字典的大小。

贝叶斯分类算法明显比之前的 KNN 分类算法在运行时间上有改观。