实验报告

学号：201814849　　姓名：张延超　　班级：2018 级学硕

1、 实验任务

本次实验的任务是使用朴素贝叶斯分类器实现文档的分类。朴素贝叶斯分类器有 3 种模型：伯努利模型（Bernoulli Model），多项式模型（Binomial Model）以及混合模型（Joint Model）

2、 实验步骤

（1） 数据预处理

在上一个实验中，我们使用 knn 算法来对文档进行分类，在此期间，对文档内容进行了预处理，并生成了文档集的词表、词项频率和文档频率。根据生成的词表，我们可以对文档进行预处理，过滤不在词表中的单词，重新生成新的文档数据集。实验结果存储在 20new_18828_preprocess_filter 文件夹中。本次实验中一共有 20 类，根据实验要求，我们需要统计每个类别所包含的单词以及该单词在本类中出现的次数和在本类文档中出现的次数，以便之后的概率计算。类别统计信息存储在 class_word_info 文件夹中。

（2） 基础模型

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j \in V} P(v_j | x_1, x_2, \ldots, x_n) \\ &= \arg\max_{v_j \in V} \frac{P(x_1, x_2, \ldots, x_n | v_j) P(v_j)}{P(x_1, x_2, \ldots, x_n)} \\ &= \arg\max_{v_j \in V} P(x_1, x_2, \ldots, x_n | v_j) P(v_j) \end{aligned}$$

得到最后的公式

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(x_i | v_j)$$

在模型中，有两点是在实现的时候需要注意的：（1）有时候，条件独立这个前提是不满足的（2）在计算概率的时候，需要对概率进行平滑处理，因为在测试过程中存在这样一种情况：测试集中存在词汇在训练集中并没有出现。

（3） 伯努利模型

在统计与判断时，将重复的词语都视为其只出现 1 次。这种方式更加简化与方便，当然它丢失了词频信息，因此效果可能会差一些。

（4） 多项式模型

重复的词语视为其出现多次，在统计与判断时，都关注重复次数。

（5） 混合模型

在计算句子概率时，不考虑重复词语出现的次数，但是在统计计算词语的概率时，却考虑重复词语的出现次数。这种方式更加简化与方便，当然它丢失了词频的信息，因此效果可能会差一些。

（6） **以上 3 种模型都在 naïve_bayes.py 中得以实现。**

（7） **平滑技术、取对数来转换权重**

3、 实验结果

	伯努利模型	多项式模型	混合模型
Accuracy	**86.75%**	**86.83%**	**86.51%**