**任务三：使用朴素贝叶斯对电影评论分类**

**（1）数据集讲解**：

该数据集是IMDB电影数据集的一个子集，已经划分好了测试集和训练集，训练集包括25000条电影评论，测试集也有25000条，该数据集已经经过预处理，将每条评论的具体单词序列转化为词库里的整数序列，其中每个整数代表该单词在词库里的位置。例如，整数104代表该单词是词库的第104个单词。为实验简单，词库仅仅保留了10000个最常出现的单词，低频词汇被舍弃。每条评论都具有一个标签，0表示为负面评论，1表示为正面评论。

训练数据在train\_data.txt文件下，每一行为一条评论，训练集标签在train\_labels.txt文件下，每一行为一条评论的标签；测试数据在test\_data.txt文件下，测试数据标签未给出。

**（2）步骤提示：**

**1.训练思路：**每个文档可以看成由n个特征构成的文档向量，每个单词表示一个特征维度。统计正样本数和负样本数，可以得到文档分布的先验概率；统计每类样本中，某个单词出现的次数和总单词数的比值，可以得到该特征的条件概率。

**2.拉普拉斯平滑：**朴素贝叶斯用各个特征的条件概率连乘表示某个样本在某个类别的条件概率。然而，如果一个单词没有出现在某个类别的样本中，那么它的条件概率就是0，导致最后的连乘结果也为0，从而将不再有文档被分到这一类。因此在训练的过程中，注意使用拉普拉斯平滑处理。

**（3）提交要求**：

将测试数据预测结果，与训练数据标签存储方式相同，存储为txt文件，每一行为一条评论的标签。实验报告中（和任务一任务二写在同一份里）需要写明具体实验流程，思路。

需要在系统中上传测试集预测结果的txt文件。