实验报告

一、 实验内容

首先读取每个文本的内容，进行分词、去停用词等操作，同时按一定比例划分训练集和测试集。在这里我是把每个文件夹下前 90%的文本作为训练集，剩下的作为测试集。程序中函数 data_process 完成这些工作。函数返回两个字典：dict_train,dict_test。dict_train 中 key 为文件夹名（类名），value 值是个二级字典，二级字典 key_2:value_2 对应该文件夹下所有单词及其频率。dict_test 的 key 是每个文本的路径，value 也是个二级字典，二级字典的 key_2:value_2 对应该文本所有的单词及其频率。

然后，对测试集进行分类，具体就是计算测试文本分到每个类的概率，将该文本分到概率最高的那个类。我分别使用了多项式模型和伯努利模型，分类的正确率分别为 82.8%和 84.3%。

二、 实验日志

NBC 相对于 KNN 简单些，但我却用了很长时间才完成它。因为使用字典的 clear()不当，导致程序的运算结果及其不合理。时间都用在定位这个 bug 上面了。最后发现 python 中有个比较坑人的部分：一个对象同时又充当了类似于指针或引用的功能。所以 clear 一个字典的同时，会影响之前的赋值。之所以出现这些情况，还是因为对 python 了解的不够。所以今后还用继续深入的学习 python。