作业报告

学　号：201814841　　姓　名：徐　强

经过三周的学习和实践，基本完成了作业要求的内容。报告内容主要分为作业完成情况、学习收获和几点思考三个部分。

一、作业完成情况

按照 NBC 处理流程，将程序划分为三个步骤编写，具体实现过程如下：

（一）文档读取并划分训练集和测试集

利用 sklearn.model_selection 模块实现了随机划分 80%训练集和 20%测试集，并根据后面 NBC 需要将文档用字典结构存储，键为文档分类，值为每个元素为一篇文档的列表。

（二）文档预处理并去除高低频词

1.基本和 KNN 的处理过程相同，使用 String 模块替换了文档中其它字符内容，使用 Textblob 模块进行了 tokenization，使用 NLTK 模块进行了 lemmatize、steming 和去掉 stopwords 和单词长度小于 3 的单词，使字典结构里的每个文档成为一个由一系列单词组成的列表。

2.使用 collections 模块的 Counter 类统计了每个类的单词在每类文档中的词频，并根据统计出的词频，设置了高频词和低频词上下限，去除了每类中的 non-informative words and rare words。

（三）进行 NBC 并计算正确率

1.根据 Bayes 公式进行分类的基本思想是：

$$P(\text{“属于某类”}|\text{“具有某特征”}) = \frac{P(\text{“具有某特征”}|\text{“属于某类”})P(\text{“属于某类”})}{P(\text{“具有某特征”})}$$

计算已知测试数据的特征的条件下要计算测试数据属于某类的最大概率（后验概率），等于计算已知训练数据属于某类的条件下所具有特征的概率（相似度）乘以属于训练数据中某类的概率的最大值。

2.按照 NBC，假设一个数据中的特征之间独立：

$$P(x|c)=P(x1,x2\ldots xn|c)=P(x1|c)*P(x2|c)\ldots P(xn|c)$$

3.按照平滑后的多项式模型分别计算测试数据文档中每个单词在一个类出现的概率：

$$\hat{P}(w_i \mid c) = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right) + |V|}$$

4.将上述概率值的连乘取对数改为连加，计算测试集中每篇文档属于训练集中每个类的概率，划分至最大概率值的那一类，完成 NBC。

5.对比测试集原本属于的类和 NBC 后的类，在考虑权重的情况下，计算 5 次后的平均正确率约为 89.57%。

二、学习收获

1.重新学习了 Bayes 公式并将其用在具体问题的解决中，理解了 naive 的本质就是将数据的各个特征值近似看做相互

独立，进而将条件联合概率近似为条件独立概率的连乘。

2.完成了上次作业中随机划分训练集和测试集的任务。

三、几点思考

1.NBC 基本没有可以调整的参数，优化的方法不多。经过权重计算正确率提升效果不是很明显，且计算权重的时间复杂度有点高，程序运行时间较长。

2.NBC 对于数据的预处理程度要求比较高，特征选取的不同也会影响 NBC 的效果。通过调整设置类词典上下限（具体记录数据见表格）观察 NBC 的正确率，感觉去除高低频词对整个 NBC 结果影响不大。

3.调整了训练集和测试集的大小（具体记录数据见表格），在训练集为 10%的情况下 NBC 还能得到约 70% 的正确率，感觉 NBC 方法还是效果不错的。