实验报告

姓名：段新朋
学号：201814810

实验要求

使用朴素贝叶斯的方法为实验数据集文档分类。

朴素贝叶斯

利用朴素贝叶斯进行文档分类的主要思想是：通过比较某个文档属于哪一类的概率的大小，实现文档分类，假设变量 X 表示文档，变量 Y 表示类别，则文档 x 属于类别 y 的概率为：

$$P(Y = y|X = x)$$

若要求出该概率，需要两个关键点：

1、 **贝叶斯定理**

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$
$$= [P(X = x|Y = y) * P(Y = y)]/P(X = x)$$

利用贝叶斯定理，将 P(Y=y|X=x)的计算转换为 P(Y=y)和 P(X=x|Y=y)的计算。

2、**朴素**

朴素的意思是向量 x 的每个分量之间相互独立，于是:

$$P(X = x|Y = y)$$
$$= P(X1 = x1|Y = y) * P(X2 = x2|Y = y) \ldots * P(Xn = xn|Y = y)$$

朴素贝叶斯的三种模型

朴素贝叶斯主要有三种模型即，伯努利模型，多项式模型，高斯模型；在本次实验中，实现了伯努利模型和多项式模型；其中，伯努利模型不考虑词频，只考虑每个单词是否出现，多项式模型需要考虑词频；高斯模型针对连续的特征变量。

拉普拉斯平滑

拉普拉斯平滑指的是在计算某个类下的某个特征的概率的时候，如果测试数据中出现某个特征在训练数据集中没有出现过，那么，该特征的概率会是 0，又因为计算 P(x|y)的时候是将所有的特征的概率乘起来，最后会导致该概率为 0；拉普拉斯平滑就是这种问题的一个解决方法，也就是把分子分母同时加一个数字，使得每个特征的概率都不为 0。伯努利模型在计算概率的时候，分子加 1，分母加上某个类中单词个数。多项式模型计算时，分子加 1，分母加 2。

代码实现

代码中主要包括加载数据，计算 P(y)，计算 P(x|y)，得到预测结果，模型评估等五部分；data_namager.py 主要用于加载数据，naïve_bayes.py，naïve_bayes_multipoly.py 分别是伯努利模型，多项式模型的计算 P(y)，计算 P(x|y)，得到预测结果三部分，metrice.py 主要用于模型评估。

在代码实现的过程中遇到两个问题：

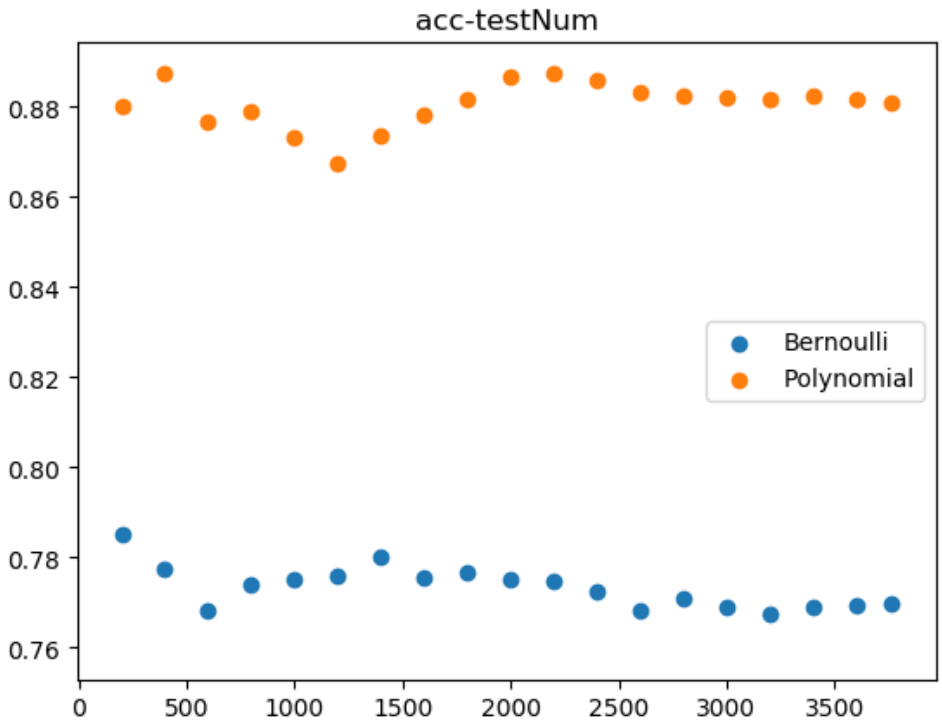
1、 在 python 中，以[[0]]*10 的形式初始化二维数组时，得到的二维数组中的每个一维数组其实是指向的同一个一维数组。

2、 在 python 中概率如何保存的问题；在多项式中由于是每个类中每个单词对应一个概率，所以采用 directory 保存某个类中的所有单词的概率，用 list 保存 20 个类对应的字典。

结论

1、伯努利模型和多项式模型的准确率分别为：0.7697822623473181 和 0.8810408921933085。

2、伯努利模型和多项式模型的准确率随着测试文档个数的增多的变化如下图所示：



3、伯努利模型对于每个类的 precision，recall，f1-score 如下图所示：

	precision	recall	f1-score	support
alt.atheism	0.84	0.80	0.82	148
comp.graphics	0.78	0.69	0.74	199
comp.os.ms-windows.misc	0.69	0.74	0.72	195
comp.sys.ibm.pc.hardware	0.71	0.73	0.72	191
comp.sys.mac.hardware	0.54	0.90	0.67	183
comp.windows.x	0.89	0.70	0.79	195
misc.forsale	0.35	0.93	0.51	191
rec.autos	0.91	0.76	0.83	205
rec.motorcycles	0.90	0.93	0.91	222
rec.sport.baseball	0.94	0.83	0.88	209
rec.sport.hockey	0.99	0.89	0.93	193
sci.crypt	0.91	0.80	0.85	191
sci.electronics	0.83	0.67	0.74	221
sci.med	0.95	0.71	0.81	201
sci.space	0.94	0.72	0.81	184
soc.religion.christian	0.91	0.82	0.86	222
talk.politics.guns	0.86	0.80	0.83	158
talk.politics.mideast	0.96	0.73	0.83	194
talk.politics.misc	0.77	0.65	0.70	141
talk.religion.misc	0.88	0.42	0.57	123
avg / total	0.83	0.77	0.78	3766

4、多项式模型对于每个类的 precision，recall，f1-score 如下图所示

	precision	recall	f1-score	support
alt.atheism	0.84	0.93	0.88	148
comp.graphics	0.75	0.82	0.78	199
comp.os.ms-windows.misc	0.94	0.65	0.77	195
comp.sys.ibm.pc.hardware	0.69	0.80	0.74	191
comp.sys.mac.hardware	0.74	0.92	0.82	183
comp.windows.x	0.89	0.84	0.86	195
misc.forsale	0.78	0.84	0.81	191
rec.autos	0.94	0.91	0.93	205
rec.motorcycles	0.94	0.98	0.96	222
rec.sport.baseball	0.96	0.96	0.96	209
rec.sport.hockey	0.97	0.95	0.96	193
sci.crypt	0.95	0.93	0.94	191
sci.electronics	0.85	0.85	0.85	221
sci.med	0.98	0.89	0.93	201
sci.space	0.93	0.93	0.93	184
soc.religion.christian	0.94	0.91	0.92	222
talk.politics.guns	0.87	0.95	0.91	158
talk.politics.mideast	0.98	0.96	0.97	194
talk.politics.misc	0.89	0.86	0.87	141
talk.religion.misc	0.84	0.67	0.75	123
avg / total	0.89	0.88	0.88	3766

总结

1、 理解了朴素贝叶斯算法，以及其伯努利模式，多项式模式，高斯模式三种模型。

2、 理解了在文档分类里，为什么要加拉普拉斯平滑，如何加拉普拉斯平滑；并验证了若不加平滑，对结果的巨大的影响。尤其是在多项式模型中，若不加平滑，准确率只有 0.004，加上平滑之后，准确率可达 0.88。

3、 熟悉了 python 中的一些数据结构(directory, generator 等), 更正了对 python 中二维数组初始化的一个误解([[0]*3]*5)。