程序说明

Training data 用的是mini\_newsgroups，每个类里都只有100个文件。这样在调试的时候速度比较快，自然准确性有所下降。

0到19的文件是我从完整的20\_newsgroups里随便抽取的20个文件，不过自然是跟mini——newgroups里面的不一样。

方法用的是naïve Bayes，根据

其中***d***为检测数据的特征向量，为第i类，这里i=0,1…19。因为是常数，可以忽略，这次的测试数据中每个类别的文档数也相同，所以的值夜是一样的。最终可以得到

计算的值可以有词频法跟文档法，也就是计算词的频率以及只记录词的出线与否。这里我采用了更加简单的文档法。假设每个词相互独立，可以得到

其中为***d***中得一个元素，也就是一个单词。为了避免，可以利用拉普拉斯概率估计得到

其中表示类中出现的文档数量。为总的文档数。但是因为为常数，所以事实上我们可以为了计算的方便而更变分母。

最后得到计算得到的也就是得到的预测。

程序中是全部打印出来了，也就是0.out到19.out，里面的数字就是计算得到的每个，观察argmax跟真实的分类，分类情况还可以。虽然不一定完全正确，但是基本能分到大类中，数字较大的项也是与真实分类属于一个大类。

不过可以看得出，如果文章比较长，能获得更好的结果。

补充：

Rate1.out表示100~200号文件的正确率，里面包括了所有的group分别对应的正确率，同理有rate2.out到rate9.out。