Naïve Bayes 实验报告

一、项目描述

此次实验主要是实现 Naïve Bayes 分类器，过程如下：首先预处理给定的新闻数据集，包括分词，去标点符号等，去除低频词，清洗文件；其次对测试集进行处理，得到最终的分类；最后计算准确率。

二、过程实现

（1）多项式模型实现原理（公式计算）：

①每个测试样例属于某个类别的概率 = 某个类别中出现样例中词的概率的乘积（类条件概率） * 出现某个类别的概率(先验概率)

②类条件概率 p(word | cate) = （类 cate 下单词 word 出现在所有文档中的次数之和 + 1） / （类 cate 下单词总数 + 训练样本中不重复的特征词总数）

③先验概率 p(cate) = 类 cate 单词总数 / 训练样本中的特征词总数

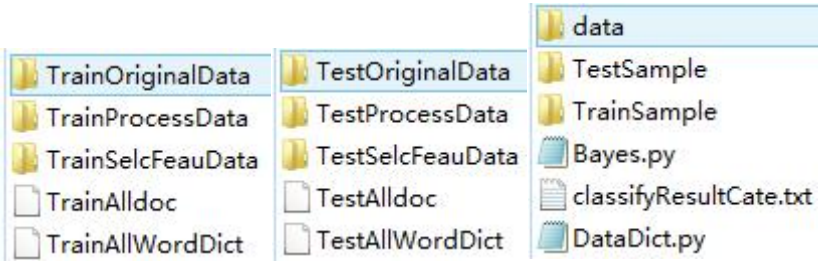
（2）DataDict.py:

与 KNN 的 VSM.py 实现功能一样，包括划分数据集，训练集为 80%，数据集为 20%；对文档内容进行分词，去标点，小写化，去停用词，词干化等处理；遍历文档统计词频，删除小于 4 的低频词，并获取字典；根据字典清理文件并删除字典中未包含的单词。

（3）Bayes.py:

首先得到每个类下每个单词的出现次数和每个类包含的单词总数；其次采取多项式模型，计算条件概率和先验概率；接着求测试样本在每个类别的概率；最后计算准确率。

三、运行结果




```
cate1 sci.space contains 129519
cate2 rec.sport.baseball contains 99656
cate3 talk.politics.mideast contains 194210
cate4 talk.religion.misc contains 93756
cate5 rec.sport.hockey contains 119155
cate6 soc.religion.christian contains 150452
cate7 comp.sys.ibm.pc.hardware contains 87509
cate8 sci.crypt contains 145027
cate9 sci.electronics contains 88841
cate10 sci.med contains 125806
cate11 rec.autos contains 93974
cate12 comp.windows.x contains 131949
cate13 talk.politics.guns contains 139563
cate14 rec.motorcycles contains 85211
cate15 comp.os.ms-windows.misc contains 279834
cate16 comp.sys.mac.hardware contains 72627
cate17 alt.atheism contains 112459
cate18 comp.graphics contains 117999
cate19 misc.forsale contains 64776
cate20 talk.politics.misc contains 146680
cate-word size: 145734
trainTotalNum: 2479003
rightCount : 3192  rightCate: 3772
accuracy is : 0.846235
```

四、问题分析及解决

1. 编码问题：服务器上的程序复制到自己电脑运行，在读取文件进行处理时会出现编码问题，显示 gbk 无法解码，判断是因为之前的数据集保存格式不同，所以统一在服务器上面运行

2. 对公式②中分子求对数避免很多很小的数相乘下溢出。

3. 原本采取将正确分类结果和得到的结果分别存储在两个文件中来计算准确率，发现比较过程中读取文件麻烦，所以写在将正确分类结果和得到的结果写在一个文件中。