The Second Experimental Report

一、 实验内容

---实现朴素贝叶斯分类器，测试其在 20 Newsgroups 数据集上的效果。

二、数据集

用上次实验中经过处理后的测试集、训练集，还有分类的文本文档。

三、 实验过程

（1） 先统计训练样本中每个目录下每个单词的出现次数，以及每个目录下的单词总数。

（2） 用贝叶斯对测试文档进行分类，输出每个类别的单词数目以及单词总数

```
cate 17 contains 208047
cate 18 contains 143570
cate 19 contains 96559
cate-word size: 211949
trainTotalNum: 2613870
```

条件概率 =（类 k 中单词 i 的数目+0.0001）/（类 k 中单词总数+训练样本中所有类单词总数）

先验概率 =（类 k 中单词总数）/（训练样本中所有类单词总数

（3） 计算准确率

四、 实验结果

```
The category talk.religion.misc contains 92864 words.
Words size in this cate is 202337
Total words num in train set is 2481065
The rightCount is : 2815 The rightCate is : 3496
The accuracy of NB classifier is : 0.805206
```

五、 实验心得

由于有了前一次 KNN 实验的基础，可以使用前次实验处理好的数据集，因此就省去了数据处理这一部分，实验也相对简单，只需要搞懂贝叶斯相关的算法就可以。课堂上只是学到了贝叶斯算法的基础，由于编程基础较弱，所以从网上找了些贝叶斯算法相关的代码进行学习，在他们的基础上加以优化，最终完成了本次实验。

本次实验心得：coding 能力有待加强。