实验报告

姓名： 刘兵 学号： 201814819

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

一、 工具平台

python3.7.0

二、 实现方法

1.**准备数据集**（在“20news-18828”文件夹内，20 个类 18828 个文档）；

2.**数据划分**。将数据集分为两部分，每个类前 80%文档构成训练集，后 20%构成测试集；

3.**文件预处理**。Txt2wordLst(txt): # 对单个文本进行处理：转为小写字母、分割、去掉无用的词、词形还原、去掉带数字的、去停用词。

4.**去除低频词**（判断标准为大于 2 或大于最大词频的一半的值），返回选出来的每个词构建的 list。

5. **加载文档**，loadDataSet(dataSourcePath, part)，2-5 项处理后返回文档的 list 形式和每个类的文档数 list。

6.**构建词典**。根据训练集全部文档内容构建词典（createVocabList）。

7.**词频统计**。各词在该类文档中总的出现总次数，将值表示为 list。

8.**文档分类。**调用 classifyNB 函数进行分类：对测试集每个文档进行向量化，然后再训练集中各类每个文档的向量分别相乘再取对数，

计算出每个类下的概率，取最大的概率的类为识别出的类。

三、 调试总结

1.前期调试每轮需要800多秒，为了节约时间，将训练集生成的数据保存到硬盘，下次调试后面程序直接从硬盘读取，每次调试时间节约了1/3。

2.训练集所得的参数都保存到硬盘，对测试进行测试时，直接调用该数据即可。将训练过程和测试过程独立开，不相互干扰。

四、实验结果

1.各阶段结果详见各txt文档。

2.测试正确率与字典有一定关系。

3.测试正确率为86.042412193505633%。