Data Mining 实验报告

Homework NBC

朴素贝叶斯分类

201834889

张玉卉

一实验要求

实现朴素贝叶斯分类器

测试其在 20 Newsgroups 数据集上的效果

二实验步骤

1、数据读取并且处理，遍历文件夹读取全部数据，并进行大写转小写、分词、词干还原、去停用词、去标点符号和与单词无关的字符、将每一篇处理好文档保存起来，并且为文档建一个标签 list，并生成词汇表、处理文档同时记录每一个词汇出现过所有的次数，并且过滤出现次数低于 20 高于 800 的单词生成词典。

2、将保存好的文档划分为测试集和训练集，比列为 20%和 80%

3、训练过程，计算词频，对存放词频向量的训练集进行训练，首先计算先验概率：P(S)=某类文档数/文档总数。其次每个类中每个词出现的概率：P(word|S)=(类 S 中 word 出现总数+1)/(类 S 中出现的单词总数+词典长度)。

4、测试过程，用朴素贝叶斯公式计算测试文档属于某个类的概率，哪个概率最大就属于哪个类，最后依据测试集的标签统计正确率。

三实验结果

20news 中随机挑选 20%测试集实验输出的分类正确率为 86.7%