山东大学 计算机科学与技术 学院

信息检索与数据挖掘 课程实验报告

学号：201600130032	姓名： 陆禹承	班级： 智能 16

实验题目：朴素贝叶斯分类器

实验内容：

根据给定的数据，使用朴素贝叶斯分类器进行分类，并测试结果。

实验过程中遇到和解决的问题：

实验时使用上一次实验处理过的数据。

读入上一次已经做好分词工作的数据。

大概有 20+个分类，根据分类统计每个类别的词出现次数

然后计算概率：当前词所在类别出现次数/类别总词数

为了方便计算，实际程序中用了平滑技术（避免 0 概率），并且取对数进行运算（避免下溢）。使用的是多项式模型。

类别 t_i

词编号 w_i

词 w_i 在 t_j 中出现次数 c_{ij}

t_i 中总词数 a_i

总单词种类数 k

$m[t_i][w_i]=\log((1+c_{ij})/(a_i+k))$

分类判别的时候，将每个类对应词的矩阵值相加，可以获得属于该类别的权值（通常是一个负数）

然后取最大权值，表示分类结果。

训练和测验时，90%作为训练集，10%作为测试集，一个文档属于训练集还是测试集靠随机数判定。

最后测试结果，平均准确度大约为 85.8%

多次测试，准确度浮动在 84.8% ~ 86.8%

使用不同比例的训练集和测试集对最终结果影响不大（测试集占比 10%~20%）。

除了朴素贝叶斯分类器，似乎还有对其改进的方法。

https://www.jianshu.com/p/46f9e837a43c

参考 4.3-4.5

结论分析与体会： 对朴素贝叶斯分类器有了基本认识