Homework 2: NBC

姓名：史金婉　　**学号：**201834874

教师：尹建华　　**项目：**朴素贝叶斯进行文本分类

一、实验目的

1. 进一步熟悉预处理文本的方法
2. 理解并掌握朴素贝叶斯的原理和应用
3. 利用朴素贝叶斯对文本进行分类

二、实验任务

使用朴素贝叶斯分类器实现文档的分类

三、实验数据

20news-18828.tar.gz (http://qwone.com/~jason/20Newsgroup)

四、实验步骤

1. 预处理

本次实验中一共有 20 类，根据实验要求，我们需要统计每个类别所包含的单词以及该单词在本类中出现的次数和在本类文档中出现的次数，以便之后的概率计算。

2. 模型

计算待分类文本可能属于每个类别的概率，将其分类为概率最大的类别，即

$$\hat{c} = \arg\max_{c \in C} P(c \mid d) = \arg\max_{c \in C} \frac{P(d \mid c)P(c)}{P(d)}$$

d 是待分类文本，由单词组成；P(c)为文本类别的先验概率，即每个类别的文本个数占总文本个数的比例； P(d | c)为每类文本中出现待分类文本的概率，待分类文本由单词构成，假设每个单词出现相互独立，则有：

$$\hat{c} = \arg\max_{c \in C} P(d \mid c)P(c) = \arg\max_{c \in C} P(d_1 \mid c) \cdot P(d_2 \mid c) \cdot \ldots \cdot P(d_n \mid c)P(c)$$

$$= \arg\max_{c \in C} P(c) \cdot \prod_{d_i \in d} P(d_i \mid c)$$

3. 应用对数函数

为了避免计算过程出现下溢，引入对数函数 log，得到如下公式：

$$\hat{c} = \arg\max_{c \in C} \left(\log P(c \mid d) + \sum_{d_i \in d} \log P(d_i \mid c) \right)$$

五、实验结果

Accuracy:0.83