实验报告

学号：201814817　　姓名：连艺翔　　班级：2018 级学硕

一、 简介

本实验是基于贝叶斯分类器在数据集 20news-18828 上的实现实验，本实验任务分为两部分：1）将 20news-18828 上的数据进行预处理，将文本构成词典。2）用贝叶斯分类器对 18828 中的文档进行分类。

本实验分为两个部分:1)实现从 20news-18828 中读取数据，并通过文档预处理，将 string 处理为单词构成的 list。主要使用 NLTK、Textblob 等工具。2）计算先验概率和后验概率，通过贝叶斯公式计算文档属于某一个类的概率，取出概率的最大值。

二、 数据集

20news-18828数据集一共包含20个类，共18828个文档，均来自各个不同的新闻评论，各类数据分布平衡，文档编码少部分采用ISO格式，大部分采用ASCII编码。文档平均长度在1000词左右，因为取自新闻有部分网页格式。

三、 方法步骤

3.1数据集预处理

数据集预处理分为：1）去符号2）分词3）词形还原4）大小写转换5)去停用词这五个部分组成。使用str类的方法maketrans对文档进行去符号，这里我们除去所有非英文字母的符号。使用

textblob对文档进行分词，使用nltk套件中的Snowballstemmer进行词形还原，使用nltk中的stopwords作为停用词表。

3.2贝叶斯分类器

根据贝叶斯分类器，设文章内容为D，D = {word1，word2，……}，类别为h，h ∈ {class1，class2，……，class20}则根据贝叶斯公式：

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

分类问题即为，求：

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j|x_1, x_2, \ldots, x_n)$$

根据条件独立，即可将问题转化为求：

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(x_i|v_j)$$

四、 实验结果