**深 圳 大 学 实 验 报 告**

**课程名称：­ Probability & Statistics**

**实验项目名称：Naive Bayes for Text Classification**

**学院： College of Electronic and Information Engineering**

**专业： Electronic Information Engineering**

**指导教师： Changsheng CHEN**

**报告人： Zuyi LIAO 学号： 2022110131**

**班级： Wenhua Honors Class**

**实验时间： 2023.11.23**

**实验报告提交时间： 2023.11.23**

**教务处制**

|  |
| --- |
| Aim of Experiment:   * Familiar with the Bayes theorem. * Understand the implementation of the Bayes theorem in python. * Know how to use Naive Bayes for a practical task, e.g., text classification. |
| Experiment Content:  In this review, we initially examined Bayes Theorem and illustrated its application through an example. Subsequently, we reinterpreted the theorem using the concepts of hypotheses and evidence, rather than merely events A and B, tailoring it more specifically to the task of spam detection. This led to the derivation of the Naive Bayes method, underpinned by the assumption that each word in a document appears independently of all others. Following this theoretical foundation, we formulated a prediction rule for practical application. Utilizing the Enron dataset, we developed a binary Naive Bayes classifier dedicated to the detection of spam emails. |
| Experiment Process：    First, we can compute the log class priors by counting up how many spam/ham messages are in our dataset and dividing by the total number.    Calculate the conditional probability of a word given its category    Cumulative conditional probability score    Plus a priori probability    Predictive labelling    Compute the accuracy rate |
| Data Logging and Processing:      Recording of data |
| Experimental Results and Analysis:  **Results：**  The results are exactly what the fact is  **Analysis:**   1. **Naive Bayes Classifier Implementation:**    * Successfully implemented a spam detector using the Naive Bayes algorithm. 2. **Data Preprocessing:**    * Processed the Enron dataset, including text tokenization and frequency counting. 3. **Probability Calculation and Laplace Smoothing:**    * Calculated conditional probabilities of words in spam and non-spam emails and applied Laplace smoothing for unseen vocabulary. 4. **Independence Assumption:**    * Adhered to the Naive Bayes assumption that each word's occurrence is independent. 5. **Model Training and Prediction:**    * Trained the model to differentiate between spam and non-spam emails and predicted classifications for new email data. 6. Understanding the steps of Naive Bayes for Text Classification 7. Knowing the importance of data preprocessing in machine leaning tasks |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后10日内。