1. 项目名称：基于贝叶斯的中文垃圾邮件检测实现

2. 关键技术：文本分类

文本分类的关键是定义映射函数，其中映射函数既可以表示某些判别规则，也可以表示为机器学习或者深度学习模型组成的分类器。整个分类过程为根据已有标注好类别的分类文本，学习出映射函数，通过函数对未处理的文本进行预测，得到该文本数据的所属类别。

完整的文本分类的流程可分为以下 4 个部分:

(1) 文本预处理:将原始数据进行清洗，处理成实际应用中所需要的文本格式。

(2) 文本向量化表示:将文本中的词转换成实数向量，用于后续模型的计算。

(3) 分类模型的建立:建立合适的文本分类算法，在训练集上进行参数学习，得到模型。

(4) 分类效果评估:根据给定的评价指标得出分类效果的情况，判断算法的好坏。

2.1 文本预处理

(1) 在本模型中对数据集标记了两种标签，spam 是垃圾邮件,ham 是正常邮件，便于比较该模型在垃圾邮件识别上的好坏。

(2) 对空格以及非中文字符以及空格进行了过滤处理，使得进入模型参与训练的数据都是中文文本。

2.2 自然语言处理

无论是在汉语还是英语中，词一般都代表最小的语义单位，因此在研究中就需要将句子划分成词，才可转入后续的研究分析中。中文分词基本算法主要分类：基于词典的方法、基于统计的方法、基于规则的方法。本模型采用的是运用基于词典的方法。

2.2.1 基于词典方法的文本分词

(1) 按照一定策略将待分析的汉字串与一个“大机器词典”中的词条进行匹配，若在词典中找到某个字符串，则匹配成功。

(2) 按照扫描方向的不同：分为正向匹配和逆向匹配。

(3) 按照长度的不同：分为最大匹配和最小匹配。

在本模型中，词典来源为人民日报语料库，分词算法实现过程为，首先读取文件，按行读取得到要分词的句子将该句子代入 set 中，判断字典中是否包含，如果包含，则将该词放入 keyword，不包含检索 set 将去掉该词的句子，检索是否包含，循环直到 set 中包含该词并输出,如果子串只剩一个字符,那么就将这个字符输出将句子去掉已经输出的子串，剩余的句子为一个新的句子，进行 2,3,4,5,6，直到原句自全部输出。

2.3 文本分类模型

常用的文本分类算法有逻辑回归算法、朴素贝叶斯算法和神经网络算法。其中，逻辑回归是经典的二分类算法，既可用于预测，也能用于分类。研究选用 Sigmoid 函数，将输

入映射为概率值，实现预测功能，通过设置概率阈值实现分类功能。

逻辑回归不仅能够得到较好的分类效果，而且算法简单明了，在机器学习分类算法选择中，逻辑回归已然成为二分类任务首选算法，但是在数据特征有缺失或者特征空间很大时的运算效果并不好。

朴素贝叶斯算法是一种基于贝叶斯定理和特征条件独立假设的分类方法，其应用领域较为广泛。贝叶斯分类器所需估计的参数很少，对缺失数据不太敏感，算法也比较简单，可解释性强。

贝叶斯定理公式如下:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}. \qquad \text{公式(1)}$$

(1) P(A)是 A 的先验概率，它不考虑任何 B 方面的因素。

(2) P(A|B)是已知 B 发生后 A 的条件概率，也由于得自 B 的取值而被称作 A 的后验概率。

(3) P(B|A)是已知 A 发生后 B 的条件概率，也由于得自 A 的取值而被称作 B 的后验概率。

(4) P(B)是 B 的先验概率，也作标准化常量（normalizing constant）

2.4 贝叶斯模型价值体现

贝叶斯算法属于基于词频的二分类，用关键词是否出现，以及出现的频率来判断邮件是否为垃圾邮件。

数据集中每一行代表一封邮件。以 spam 开头代表是垃圾邮件，以 ham 开头代表是正常邮件。现在使用这个数据集训练出一个朴素贝叶斯模型。任意放入一封邮件，由模型判断出这封邮件是垃圾邮件的概率。如果这封邮件为垃圾邮件，则识别成功，如果为正常邮件，则识别错误。有两个标准来评价模型的价值——召回率 R(Recall Rate)和准确率 P(Precision Rate)

(1) 精度

精度也叫精确率，代表的是所有被分为正样本中实际也是正样本的样本个数占所有被分为正样本的比例，计算公式如下:

$$precision = \frac{TP}{TP + FP} \qquad \text{公式(2)}$$

(2) 召回率

召回率是覆盖面的度量，代表所有分为正样本中被正确识别的样本个数占所有正样本的比例。计算公式如下:

$$recall = \frac{TP}{TP + FN} \qquad \text{公式(3)}$$

在此研究中用 A，B，C，D 代表对应的数量。

-	实际为垃圾邮件	实际为正常邮件
识别为垃圾邮件	A	B
识别为正常邮件	C	D

则召回率为：

$$R = \frac{A}{A + C}$$ 公式(4)

准确率为：

$$P = \frac{A}{A + B}$$ 公式(5)

此外，准确率和召回率是互相影响的，理想情况下肯定是做到两者都高，但是一般情况下准确率高、召回率就低，召回率低、准确率高。如果是做搜索，那就需要保证召回的情况下提升准确率；如果做疾病监测、反垃圾信息，则是保准确率的条件下，提升召回。

3. 模型识别流程

使用 JAVA 语言，先对文本进行分词，根据频率筛出关键词，计算求包含关键词的邮件是否是垃圾邮件的概率，文本分类流程（1.预处理 2.文本表示及特征选择 3.构造分类器 4.分类），结果分析主要是检验分类的准确性，要计算预测结果的正确率。

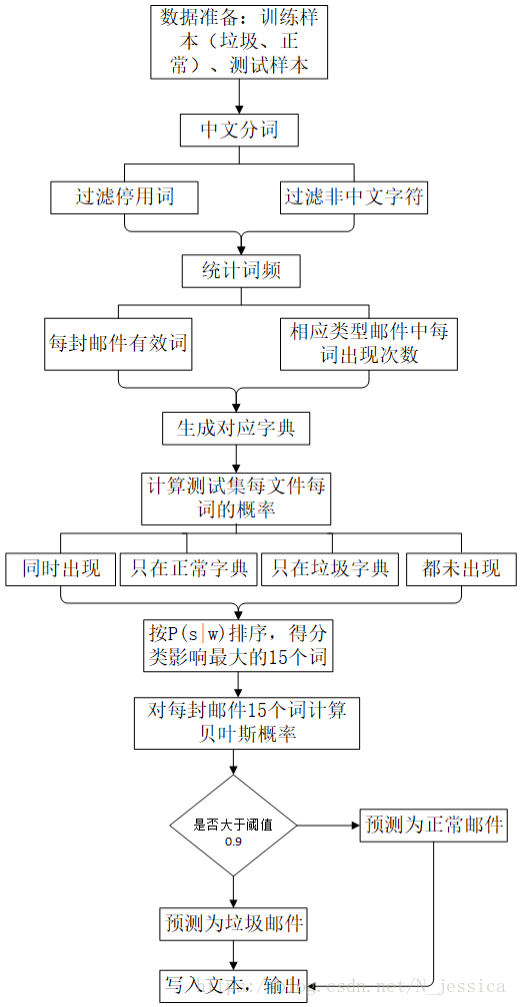

图 3 贝叶斯算法识别垃圾邮件流程图

4. 实验

4.1 实验的设备及环境

实验设备：联想笔记本电脑

实验环境：windows10+8GB 内存+64 位操作系统

实验软件：eclipse

编程语言：java

4.2. 测试数据

（1）数据集来自于互联网各个渠道，其中包含有一万多条邮件数据，每一条数据为一封邮件的内容，一封邮件存入一行，每一行的头部标注了 spam 或 ham 代表垃圾邮件或正常

邮件。将数据集划分为 1000 量级和 2000 量级，以下为 1000 封邮件的数据集如图：

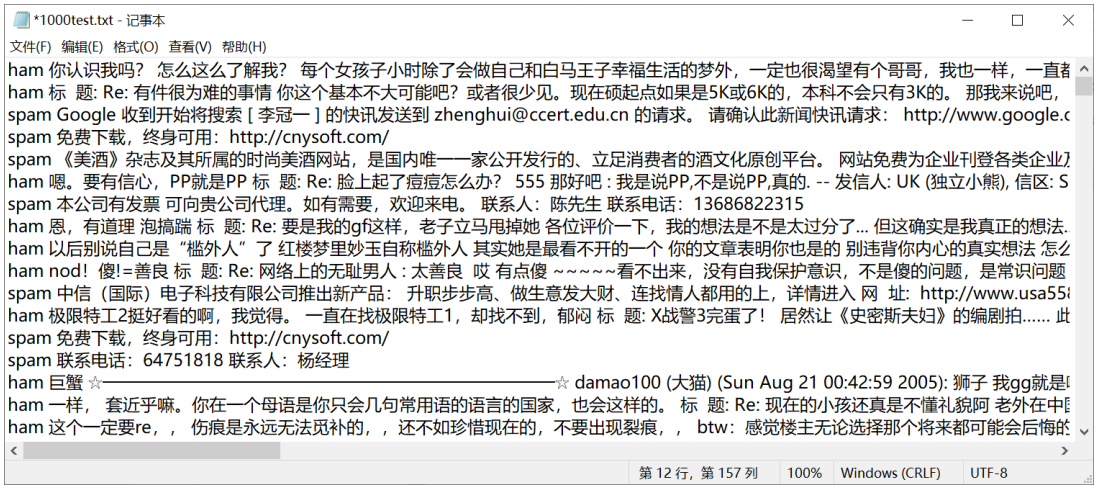

图4. 2. 1训练数据集

（2）测试数据集选取数据集的 1000 邮件，如图：

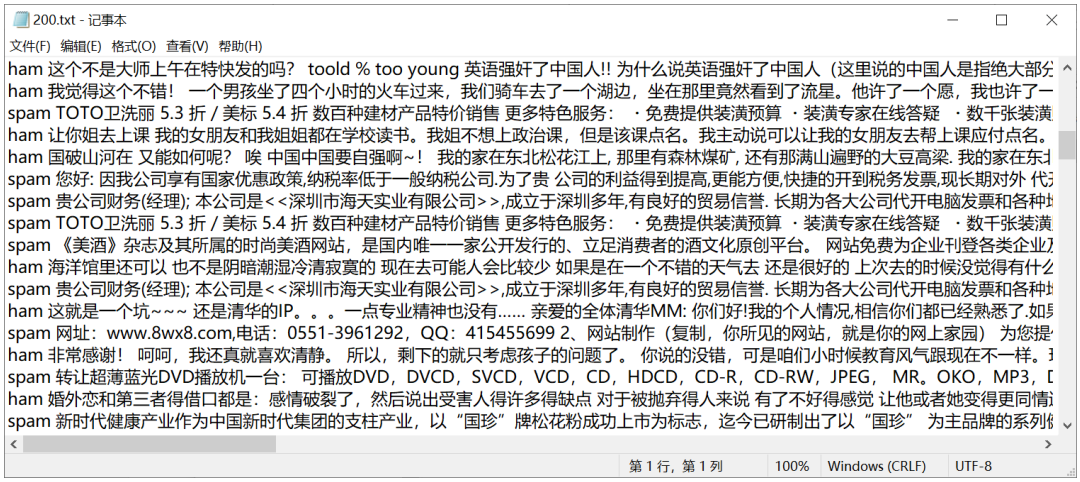

图 4.2.2 测试训练集

4.3 实验结果

结果分析主要是检验分类的准确性，要计算预测结果的正确率。

（1） 关键词筛选：

垃圾邮件关键词有采购、免费挂机、免费等有垃圾代表意义的词，实验初步效果成功。

```
[关键词=深层卸妆油, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=陈先生, 在垃圾邮件中出现的次数=14, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=信件已收到, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=全部正规产品, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=元一分钟, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=超大免费邮箱, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=内河运输发票, 在垃圾邮件中出现的次数=2, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=请联系, 在垃圾邮件中出现的次数=9, 垃圾邮件的总数=471, 在正常邮件中出现的次数=1, 正常邮件的总数=529, 联合概率=0.9]
[关键词=北京市朝阳区十里堡恒泰大厦, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=尊敬的商家客户, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=希望合作, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=级会员的, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=张源, 在垃圾邮件中出现的次数=2, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=无效退款, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=发广告服务等, 在垃圾邮件中出现的次数=2, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=免费挂机网, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=海外公司注册权威机构, 在垃圾邮件中出现的次数=2, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=不要手机, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=属正规发票, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=月圆人圆事事团圆, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=或是各地有分公司的企业, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=本公司作风凌厉, 在垃圾邮件中出现的次数=2, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=手机数据线, 在垃圾邮件中出现的次数=1, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=采购, 在垃圾邮件中出现的次数=2, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
[关键词=有商品销, 在垃圾邮件中出现的次数=2, 垃圾邮件的总数=471, 在正常邮件中出现的次数=0, 正常邮件的总数=529, 联合概率=1.0]
```

图 4.3.1 基于词频摘要生成的短文本结果

（2）垃圾邮件分类结果以及模型训练的结果：

准确率达到 0.99，召回率为 0.63。垃圾邮件检测初步实现，后续研究还可从如何提高召回率入手。

```
这份邮件是垃圾邮件的概率是: 0.9999629612145945, 实际是否为垃圾邮件: true
这份邮件是垃圾邮件的概率是: 0.9971778652734934, 实际是否为垃圾邮件: true
这份邮件是垃圾邮件的概率是: 0.5, 实际是否为垃圾邮件: false
这份邮件是垃圾邮件的概率是: 0.9646227796027926, 实际是否为垃圾邮件: false
这份邮件是垃圾邮件的概率是: 0.999999998695185, 实际是否为垃圾邮件: true
这份邮件是垃圾邮件的概率是: 0.9999999220934693, 实际是否为垃圾邮件: true
这份邮件是垃圾邮件的概率是: 0.9343734590783699, 实际是否为垃圾邮件: false
这份邮件是垃圾邮件的概率是: 0.48859934853420195, 实际是否为垃圾邮件: false
这份邮件是垃圾邮件的概率是: 0.8878314537346205, 实际是否为垃圾邮件: true
这份邮件是垃圾邮件的概率是: 0.9999999868634825, 实际是否为垃圾邮件: true
这份邮件是垃圾邮件的概率是: 0.8840716190387334, 实际是否为垃圾邮件: true
这份邮件是垃圾邮件的概率是: 0.9999881293087495, 实际是否为垃圾邮件: true
这份邮件是垃圾邮件的概率是: 0.9999999999849865, 实际是否为垃圾邮件: true
垃圾邮件总数为: 496, 正确识别了319封垃圾邮件,召回率0.6431451612903226, 准确率: 0.996875
```

图 4.3.2-模型训练结果

4.4 实验结果分析

（1）不同量级的数据集对比：

	准确率	召回率
1000 封邮件训练集	0.996875	0.6431
2000 封邮件训练集	0.994366	0.71313

通过比较，可知增大训练集的数据量，可提高算法有效性。