

本科毕业论文(设计)

**论文题目：** 基于机器学习的垃圾邮件检测分类系统

学生姓名： 方阳

学 号**：** 2131053129

专 业**：** 网络工程

班 级： 网络2101

指导教师**：** 殷娅铃

完成日期**：2024年3月30日**

**西安财经大学**

**本科毕业论文（设计）独创性及知识产权声明**

本人郑重声明：所呈交的毕业论文是本人在导师的指导下取得的成果，论文写作严格遵循学术规范。对本论文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本毕业论文引起的法律结果完全由本人承担。

本毕业论文成果归西安财经大学所有。

特此声明

毕业论文签名：

专业：

学号：

年 月 日

**基于机器学习的垃圾邮件检测分类系统**

摘 要

当今时代是数字化、网络化、信息化蓬勃发展的时代，人们越来越多地依赖于网络进行交流和沟通。在校大学生尤其需要一个共享平台来进行日常生活和学习上的交流，以便更好地适应大学生活。随着各类软件的发展，微信和QQ等社交软件虽已非常流行，但是这些软件实现共享交流的范围有限，已经不能满足当代大学生的交流需求了。因此，大学生群体需要一个更专业全面、更适合校园生活和学习的交流平台。

本研究以当代大学生校园生活各类需求为根据，结合校园热点讨论话题，设计出一款具有生活分享、学习交流、交友告白和二手交易等多功能模块的交流平台。在本系统的开发中，应用了Spring Boot、Vue、Redis、MySQL等多种技术。一方面，在前台页面实现发布话题、点赞评论等功能，便于大学生进行各类话题的讨论交流。另一方面，在后台实现系统管理、平台管理等功能，便于管理员对系统进行规范管理。最后对系统进行功能测试和性能测试，分析和优化测试结果。最终形成了一个符合大学生校园生活日常实际需求的交流平台。

**关键词**：交流平台；校园生活；Vue；Spring Boot

**Design and Implementation of Campus Life and Learning Communication Platform**

Abstract

Today's era is one of flourishing digitalization, networking and informatization, and people are increasingly relying on the Internet for communication and exchange. College students especially need a shared platform to communicate in daily life and study in order to better adapt to college life. With the development of various software, social software such as WeChat and QQ have become very popular, but the scope of shared communication by these software is limited and can no longer meet the communication needs of contemporary college students. Therefore, college students need a more professional and comprehensive communication platform that is more suitable for campus life and study.

Based on the various needs of contemporary college students' campus life, this study designs a communication platform with multi-functional modules such as life sharing, study exchange, dating confession and second-hand trading by combining hot discussion topics on campus. In the development of this system, various technologies such as Spring Boot, Vue, Redis and MySQL are applied. On the one hand, the front page realizes the functions of posting various topics and commenting, so as to facilitate the discussion and exchange of various topics among college students. On the other hand, the system management and platform management functions are implemented in the background, which is convenient for the administrator to manage the system in a standardized way. After that, the system was tested for function and performance, and the test results were analyzed and optimized. Finally, a communication platform that meets the actual needs of college students' campus daily was formed.

**Key Words:** Communication Platform, Campus Life, Vue, Spring Boot

目 录

1．绪 论 1

1.1 开发背景 1

1.2 研究现状及意义 1

1.3 研究内容 2

1.4 论文组织结构 3

1.5 本章小结 4

2．关键技术介绍 5

2.1 Spring Boot 5

2.2 Vue 5

2.3 Redis 5

2.4 MySQL 6

2.5 本章小结 6

3．系统分析 7

3.1 可行性分析 7

3.1.1 技术可行性 7

3.1.2 经济可行性 7

3.1.3 操作可行性 7

3.1.4 社会和法律可行性 7

3.2 需求分析 7

3.2.1 用户需求 7

3.2.2 功能性需求 8

3.2.3 非功能性需求 9

3.3 系统需求模型 10

3.3.1 学生用户用例 10

3.3.2 管理员用例 11

3.4 本章小结 12

4．系统设计 13

4.1 设计原则及运行环境 13

4.1.1设计原则 13

4.1.2运行环境描述 13

4.2 系统体系结构设计 13

4.3 系统建模 15

4.3.1 活动图模型 15

4.3.2 顺序图模型 20

4.3.3状态机模型 22

4.4 数据库设计 23

4.4.1 E-R图设计 23

4.4.2 数据库表设计 24

4.5 本章小结 26

5．系统实现 27

5.1 前台子系统的实现 27

5.1.1 用户注册功能 27

5.1.2 用户登录功能 28

5.1.3 系统首页界面 29

5.1.4 发布话题功能 29

5.1.5 发表评论功能 30

5.1.6 修改个人信息功能 31

5.1.7 话题管理功能 32

5.2 后台管理系统的实现 32

5.2.1 管理员登录功能 33

5.2.2 后台管理系统主页面 33

5.2.3 用户信息管理功能 34

5.2.4 话题类别管理功能 35

5.2.5 话题内容管理功能 36

5.2.6 评论管理功能 38

5.3 本章小结 38

6．系统测试 39

6.1 登录功能测试 39

6.2 发布话题功能测试 40

6.3 个人信息修改功能测试 40

6.4 发表评论功能测试 41

6.5 性能测试 41

6.6 测试结果分析 41

6.7 本章小结 42

7．总结与展望 43

参考文献 44

致 谢 45

1．绪 论

1.1 开发背景

随着数字通信技术的不断发展，电子邮件已稳居现代社会核心沟通工具之列，成为个人、企业及机构广泛依赖的信息传递方式。因其操作便捷、响应迅速且成本低廉，电子邮件广泛渗透至工作协同、商业营销、教育共享与个人交流等多个领域。然而，随着用户数量急剧上升与网络环境愈发开放，垃圾邮件泛滥的问题日益突出，严重扰乱了正常通讯秩序，并引发诸多安全与社会隐患。

对普通用户而言，大量未经许可的垃圾邮件早已司空见惯。这类邮件常含虚假广告、诈骗信息、恶意链接乃至病毒附件，不仅耗费邮箱空间、拖慢处理速度，更可能泄露隐私、造成财产损失。企业层面，员工为辨别垃圾邮件而耗费精力，效率受限，若误点其中恶意内容，更可能威胁企业数据与系统安全。加之，随着垃圾邮件伪装手法不断演进，如假冒正常邮件或借用社会工程学诱导点击，传统基于关键词或规则的拦截方式日渐无效，用户对更智能、精准的应对方案愈发依赖。

因此，开发一套高效且稳定的垃圾邮件智能识别与分类系统已成当务之急。此类系统可协助用户迅速从庞杂邮件中识别并拦截垃圾内容，有效减轻干扰、提升沟通质量。对个人而言，其不仅可提升使用体验，更保障信息安全；对企业与组织，则有助于控制风险、构建良好的网络安全生态。借助机器学习技术，系统能从既有数据中提取垃圾邮件特征，不断适应其演变趋势，从而提供持久而可靠的防护支持。这不仅是技术层面的进步，更是当前网络通信实际需求下的现实选择。

1.2 研究现状及意义

电子邮件广泛应用的同时，垃圾邮件问题也日益严峻，成为全球网络通信中的顽疾。为此，国内外学界和业界纷纷提出多种过滤方案，但实际效果仍有差异。以下是当前主要的垃圾邮件检测技术现状：

（1）基于规则的过滤：早期系统主要依靠人工设定关键词、黑名单等规则拦截可疑邮件，尽管实现简单，但难以应对内容伪装、关键词变体等新型手法[6]。

（2）传统机器学习方法：随着技术发展，贝叶斯分类器、SVM等被用于提取邮件文本特征并进行分类。这类方法通过统计分析提升识别率，但依赖人工设计特征，难以涵盖复杂语义，对新型垃圾邮件的泛化能力也有限[7]。

（3）深度学习模型：近年来，神经网络（如RNN、Transformer）被用于学习邮件深层语义，在处理多语言和多模态内容时表现突出。但这类模型资源消耗大，样本不足时易过拟合，部署难度高[8]。

（4）集成化商用系统：如Gmail、Outlook等平台集成多重过滤机制，结合规则引擎、黑名单和用户反馈形成动态防护。但这类系统多为“黑箱”，用户难以自定义策略，对小语种和垂直领域适应性也有限[9]。

在此背景下，我们决定采用朴素贝叶斯算法进行垃圾邮件检测系统的算法，其优势主要体现在以下几个方面：

（1）提升处理效率：自动化分类可减少用户手动筛查时间，让个人和企业聚焦重要信息，提升通信效率。

（2）增强安全防护：有效拦截带有恶意链接或病毒的邮件，防止用户因误点导致隐私或财产受损，尤其对中小企业和普通用户更具防护价值。

（3）节约通信资源：减少垃圾邮件占用的服务器存储和带宽，减轻运维负担，推动绿色信息基础设施建设。

（4）适应动态演化：朴素贝叶斯模型轻量、具备增量学习能力，能快速适应垃圾邮件策略变化，解决静态规则频繁失效的问题。

（5）推动技术普及：相比深度学习模型，其部署成本低、解释性强，适用于中小机构与个人开发者，助力技术下沉。

1.3 研究内容

本研究旨在提升电子邮件通信的效率与安全性，围绕现有垃圾邮件检测技术在模型适应性、用户可控性和资源消耗方面的短板，设计并实现一种基于朴素贝叶斯算法的轻量级垃圾邮件检测系统。该系统通过构建多层次架构与动态优化机制，实现高精度识别与灵活管理。研究内容主要包括以下三个核心部分：

（1）系统架构设计

系统采用分层结构，分为用户交互层、算法处理层和数据管理层，并通过标准化API接口实现高效协同：

用户交互层：为终端用户提供分类服务接口，支持邮件输入、结果反馈和规则自定义。

算法处理层：集成朴素贝叶斯分类器与动态优化引擎，完成特征提取、概率计算和模型更新。

数据管理层：负责模型存储、历史数据归档与安全策略配置，采用轻量JSON格式确保兼容性与快速部署。

（2）核心功能模块

邮件分类检测模块

基于朴素贝叶斯构建概率模型，分类流程包括：

数据预处理：去除HTML标签、停用词过滤、分词及词频统计。

特征向量化：提取主题、正文关键词、发件人信誉等特征，生成高维稀疏向量。

概率计算与判定：应用贝叶斯定理计算后验概率，并设定动态阈值进行分类。

最终输出为垃圾/非垃圾标签及置信评分，并支持JSON结构化返回。

模型动态优化模块

通过增量学习机制，实现模型的持续更新，依据新数据动态更新先验概率，提升模型对新型垃圾邮件的适应能力。

（3） AIGC辅助数据集构建

由于现有公开数据集（如trec06c）存在时间滞后与语义偏差问题，本研究引入AIGC生成技术，构建了15,000条高质量合成样本：

数据生成策略：结合真实垃圾邮件模板（如钓鱼、促销等），利用大语言模型生成语义自然的内容变体。

1.4 论文组织结构

根本文围绕基于朴素贝叶斯的垃圾邮件检测分类系统的研究与实现展开，共分为七章，遵循“理论分析—技术改进—实验验证—总结展望”的逻辑脉络，具体章节内容如下：

第一章：绪论。阐述垃圾邮件检测的研究背景与意义，分析现有技术的局限性，明确研究目标与内容，并概述论文的整体框架。

第二章：算法描述。介绍朴素贝叶斯算法的基本原理与数学模型，重点解析其在文本分类任务中的适用性及核心假设条件。

第三章：相关技术介绍。简述系统实现所依赖的技术工具，包括数据预处理方法、特征提取技术、API接口规范及JSON模型存储机制。

第四章：算法改进。针对朴素贝叶斯算法的不足（如特征独立性约束），提出改进策略，包括动态特征权重优化与增量学习机制设计。

第五章：算法实现。描述改进后算法的工程化实现方案，涵盖数据预处理流程、模型训练框架、API接口开发及轻量化模型部署。

第六章：测试与分析。通过实验验证算法性能，包括分类准确率、误判率等指标对比，分析改进算法的有效性及资源消耗表现。

第七章：总结与展望。总结研究成果与创新点，反思当前方案的局限性，并提出未来研究方向。

1.5 本章小结

本章的主要内容是首先对本算法系统的开发背景进行调研和了解，在此背景下深入调查该系统的研究现状和研究意义，继而采用模块化研究方法，结合领域本体构建与用例分析技术，确定核心研究对象，界定系统功能模块的边界范围。

2. 算法描述

2.1 贝叶斯原理

贝叶斯原理（Bayes' Theorem）是概率论中的核心定理之一，用于在已知某些证据（或条件）的情况下，更新对某一事件发生的概率估计。它的核心思想是通过新信息动态调整原有的概率判断，体现了“逆概率”的逻辑。

贝叶斯公式：



其中：

P(A∣B)是后验概率：在事件B发生的情况下，事件A发生的概率。

P(B∣A)是似然概率：在事件A发生的情况下，事件B发生的概率。

P(A) 是先验概率：在不知道B是否发生时，事件A的初始概率。

P(B) 是证据概率：事件B发生的总概率，通常通过全概率公式计算。

2.2 朴素贝叶斯

现在我们假设有多个特征，且每个特征之间相互独立。则我们可以将公式简化为：



其中



又因为我们最后要比较A和A¬，所以对于

可以忽略分母P(B)

所以最后可以认为



2.3 数据集

数据集主要使用trec06c[]，trec06c是一个由国际文本检索会议（TREC）发布的公开垃圾邮件语料库，分为英文（trec06p）和中文（trec06c）两个子集。数据源均为真实邮件，原始结构和内容均被完整保留。在获取数据后，我们能够直接查看邮件内容，但并非所有信息都符合训练需求，因此仅提取邮件中的中文部分作为训练语料。同时，我们还需对原始数据进行一系列预处理操作，主要包括以下几点：

1. 字符过滤，需要清除所有非中文内容，例如标点符号、英文字母、数字以及链接等特殊字符；
2. 停用词处理，对邮件文本进行分词后进一步清洗无效词项。

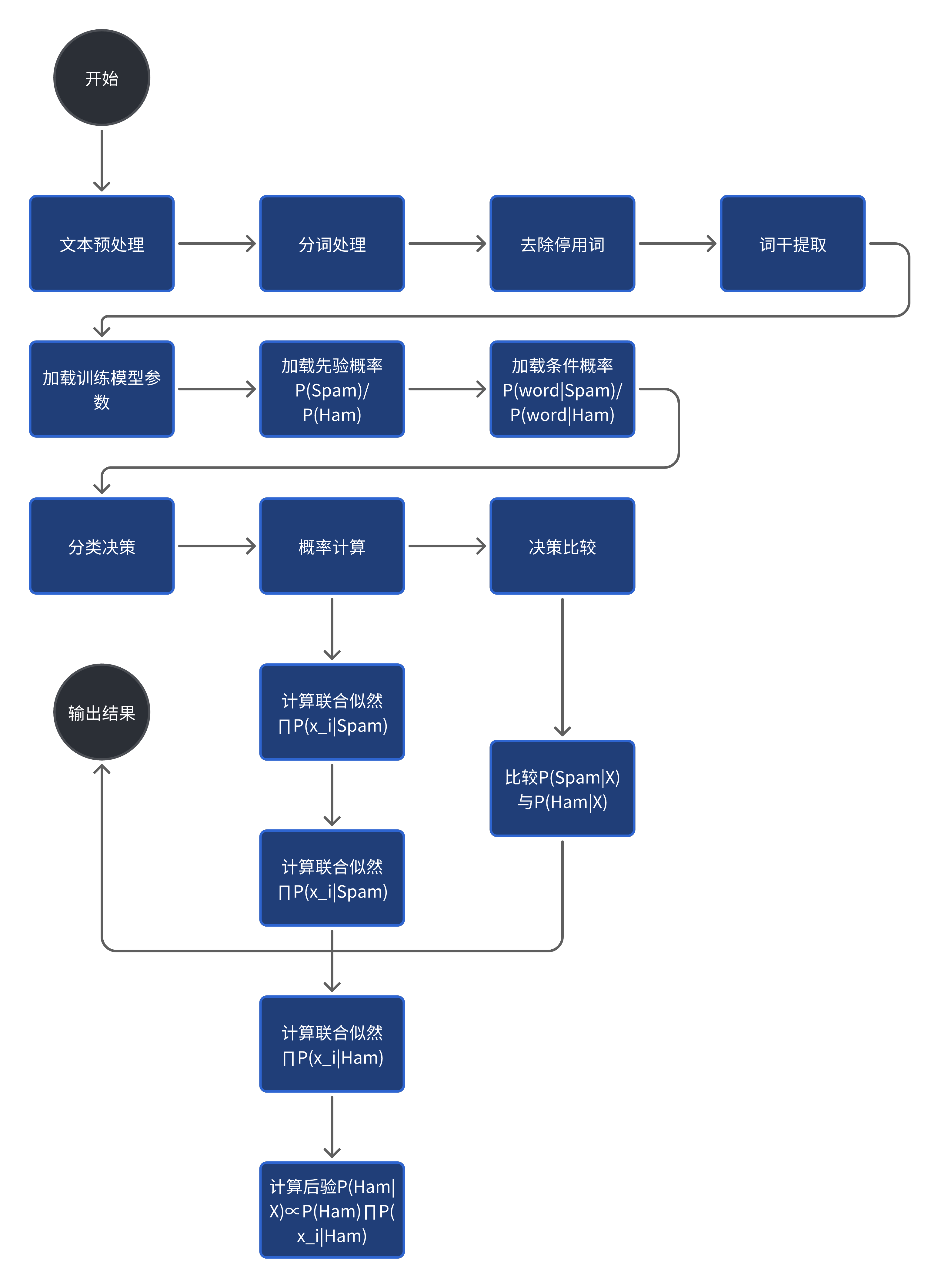
2.4 实现步骤

在垃圾邮件分类任务中，文本特征通常被定义为邮件中出现的词汇集合。基于朴素贝叶斯分类器的建模方法，需对每个词汇项在两类标签（垃圾邮件/正常邮件）下的条件概率进行参数估计。

模型通过比较联合似然与先验的乘积 ，依据贝叶斯决策理论将邮件判属后验概率更高的类别。此过程依赖于特征条件独立性假设，即假设各词汇在给定类别标签下统计独立，从而将高维联合概率分解为单变量条件概率的乘积。尽管该假设在自然语言中通常不成立，但因其显著降低计算复杂度且在实践中表现出较强的鲁棒性，仍被广泛应用于文本分类任务。

对于垃圾邮件分类任务来说，不同的特征就是在邮件中出现的词汇。我们需要计算不同词汇分别在垃圾邮件和正常邮件下的概率。

之后将所有的词汇出现的条件下为垃圾邮件的概率相乘，得出垃圾邮件概率A。同理可得正常邮件的概率A¬，比较A和A¬，即可判断邮件是否为垃圾邮件



3．关键技术介绍

3.1 Python

Python是一种广泛应用于算法开发与数据处理的高级编程语言，其语法简洁易懂，代码可读性强，能够显著降低开发复杂度。得益于丰富的第三方库支持（如NumPy、Pandas、scikit-learn等），Python在机器学习、自然语言处理等领域展现出强大的工具链优势，开发者可通过模块化调用快速实现算法原型设计与迭代优化[10]。此外，Python活跃的开源社区为技术问题的解决提供了持续支持。

在本系统中，Python作为核心开发语言贯穿整个技术实现流程。利用Pandas完成邮件文本的清洗与特征提取任务，例如去除HTML标签、统计关键词频率等预处理操作。系统后端通过groid库搭建简易交互界面。同时，Python的JSON模块被用于模型参数的轻量化存储，确保分类器能够灵活部署至不同运行环境。针对公开数据集的不足，系统还借助Python脚本调用第三方大模型API，通过自动化流程扩充训练数据，提升模型泛化能力。

3.2 jieba

jieba 是一款专为中文自然语言处理任务开发的高效分词工具，其核心依赖前缀词典和动态规划算法，实现了高精度的中文分词。与传统方法相比，jieba 支持三种分词模式：精确模式、全模式以及搜索引擎模式，可灵活应对不同的应用需求。比如，精确模式兼顾效率与准确性，适合文本分析及特征提取场景；搜索引擎模式则通过更细致的词语切分，有助于提高检索系统的命中率。另一个优势是它支持自定义词典，方便开发者根据特定领域（如垃圾邮件中的术语）灵活调整分词粒度，从而增强语义识别的针对性。

在本系统中，jieba 被用于处理中文邮件内容。由于垃圾邮件常包含广告词、伪装用语或不规则拼写，仅依赖空格或标点进行切词效果有限。借助 jieba 的精确模式，系统能更有效地将正文拆分成独立词汇，并配合停用词表（如“的”“和”等常见虚词）过滤干扰信息。接着，系统对分词结果进行词频统计，构建特征向量供朴素贝叶斯分类模型使用。针对“薇亻言”“佰付美”等混淆手法，通过加载自定义词典，增强识别这些特殊关键词的能力，从而提升整体模型在识别语义伪装时的鲁棒性。jieba 的灵活性与效率，为中文垃圾邮件的文本建模提供了强有力的支持。

3.3 pandas

pandas 是 Python 生态中广泛应用的数据处理库，核心依托于 DataFrame 和 Series 两种数据结构，支持高效完成数据清洗、转换与分析任务。它可以从 CSV、Excel、数据库等多种数据源加载结构化信息，并提供灵活的索引、切片、合并与聚合功能，尤其在处理缺失值和异常数据方面表现出色。相比传统工具，pandas 依靠矢量化计算显著提升了大规模数据的处理性能，其清晰直观的链式操作方式也极大简化了复杂逻辑的实现流程，因此已成为数据科学工作流中的常用工具之一。

在本系统中，pandas 主要用于邮件数据的预处理和特征构建。系统在从原始邮件集中提取文本内容、发件人地址等关键字段后，借助 pandas 实现结构化存储与数据清洗操作，为后续分析与建模打下基础。

3.4 Grdio

Gradio 是一个开源的 Python 库，旨在帮助开发者快速搭建机器学习模型的交互式演示界面。其最大优势在于，只需少量代码即可生成功能完整的可视化 Web 应用，支持灵活配置输入（如文本、文件）与输出（如标签、图表）组件，无需前端开发经验也能实现模型功能的直观呈现。同时，Gradio 内置自动部署机制，允许用户一键将本地模型分享为可公开访问的链接，显著降低了算法原型展示与协作的技术门槛。

在本系统中，Gradio 被用于构建一个用于垃圾邮件分类的交互测试平台。用户可通过网页上传邮件内容或直接输入检测文本，系统则实时调用朴素贝叶斯模型进行分类，并以标签高亮和置信度百分比的形式展示结果。Gradio 的轻量化设计与强交互性，使系统在可用性和演示效率上获得了显著提升。

4. 算法实现

4.1 模型格式

在本系统中，模型的表现形式为各个词汇在垃圾邮件（spam）和正常邮件（ham）的出现次数。为了方便开发，我们使用json格式来存储相关数据。

JSON格式如下：



各字段含义：

“word\_counts”：记录每个单词在正常邮件（ham）和垃圾邮件（spam）中的出现次数。

"total\_ham": 正常邮件总数

"total\_spam": 垃圾邮件总数

"ham\_word\_num": 正常邮件总词数（去重）

"spam\_word\_num": 垃圾邮件总词数（去重）

"vocab\_size": 词汇表大小

4.2 数据加载模块

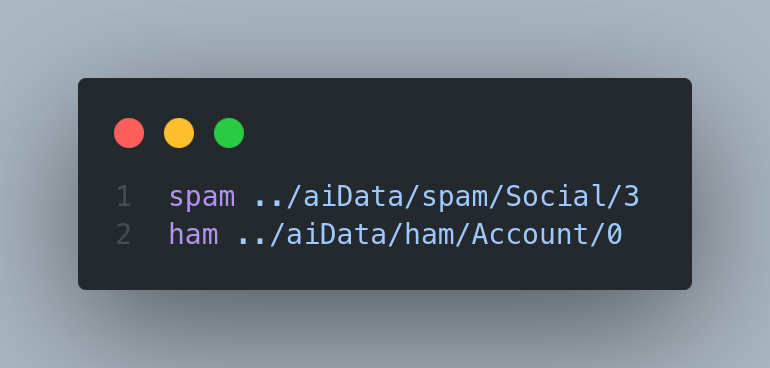
4.2.1 index索引

对于数据集，有一个index文件用来标记各个邮件的位置和类型。其格式如下，

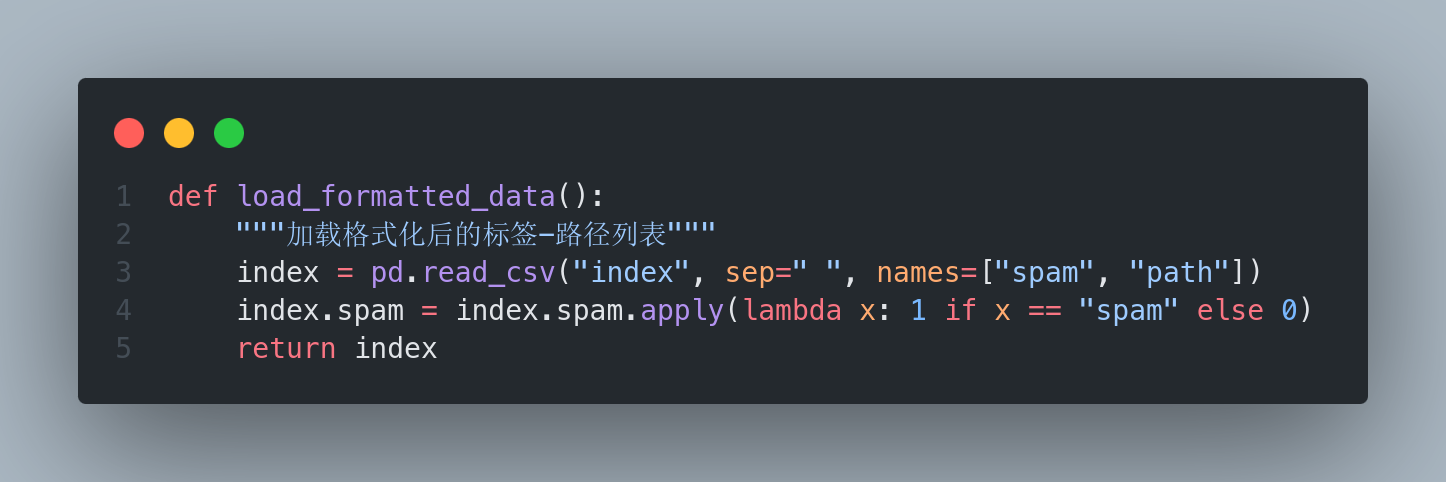
<type> <path>

其中type的值为spam或ham

path则是邮件相对于index文件的位置

例子：

之后使用pandas对index进行读取，在内存中建立对各个邮件的索引：



4.3 核心分类器

4.4 程序运行

参考文献

1. 魏智文,李双梅. 大学生信息交流平台系统的设计与实现[J]. 江苏科技信息,2021,38(22):50-52.
2. 赵志耘.深刻理解和诠释“高端交流平台”的内涵与战略意义[J].数字图书馆论坛,2021(03):1-2.
3. 余俊伟.“企业党建”搭建了交流学习的平台[J].上海企业,2022(01):57-58.
4. 刘小英,刘强. 在线学习交流平台的设计与实现[J]. 攀枝花学院学报,2021,38(02):62-67.
5. 李浩冉,拉巴顿珠. 校园信息交流平台的设计与实现[J]. 西藏科技,2020,(05):73-75+80.
6. 付吉菊.校园信息交流平台的设计与实现[J].吉林工程技术师范学院学报,2021,37(05):95-98.
7. 王胜勇,王羚伊.校园生活论坛交流平台的设计与实现[J].数码世界,2018(11):231.
8. 王志亮,纪松波. 基于SpringBoot的Web前端与数据库的接口设计[J]. 工业控制计算机,2023,36(03):51-53.
9. 王琪. SpringBoot在线学习系统的开发[J]. 互联网周刊,2023,(05):60-62.
10. 郭艳华.基于Vue框架的海量数据处理系统设计[J].信息与电脑(理论版),2022,34(23):16-18.
11. 王以伍,舒晖. 基于SpringBoot+Vue前后端分离的高校实验室预约管理系统的设计与实现[J]. 现代计算机,2023,29(01):114-117.
12. 银莉,杜啸楠. 基于SpringBoot和Vue的社区服务平台设计与实现[J]. 电子技术,2022,51(12):182-183.
13. Paul Krill. Redis Stack outfits Redis for real-time apps[J]. InfoWorld.com,2022.
14. Guy Harrison. Redis and Amazon's MemoryDB[J]. Database Trends and Applications,2021,35(5).
15. 杨卓凡. 基于MySQL不同存储引擎下数据存储效率研究[J]. 电脑知识与技术,2022,18(21):18-20.

致 谢

当我的论文写到这里时，意味着我的大学生活也已经进入尾声，四个春夏秋冬里，有春日里的樱花树，夏日里的冰淇淋，秋日里的银杏叶，还有冬日里的暖阳。一草一木都是那么的四季分明。

在这四年里，我学到了很多，收获了很多，不仅有专业知识上的收获，更多的收获到了信息学院每一位老师带给我的温暖，很感谢我的指导老师，柳老师，她就像一枚定海神针，陪伴着我完成毕业设计，让我在迷茫的时候找到自己的方向；还有指导我数学建模的潘老师，她对我的每一次教导，都让我醍醐灌顶，在她身上我学到了太多太多东西；另外还有和蔼可亲的杨老师，他让我明白作为一名学生不仅是学好专业知识，还要有良好的学习态度，温柔耐心的冯老师，还有人美心善的张老师和赵老师，能受到这里每一位老师的教导，对于我而言都是最幸运的事。

感谢我的父母，他们都是面朝黄土背朝天的庄稼人，他们通过每一滴汗水，每一粒庄稼，养育我这二十多年，而即将毕业离校去工作的我，也将通过自己的努力，去回报他们的恩情，用行动告诉他们，你们的女儿长大了，可以成为你们的依靠了！

感谢我的室友们，与她们同窗四年的生活充满了快乐，聪明热心的小王，漂亮的小冯，还有可爱的小张，是她们的存在，让我的大学生活更加丰富多彩。

最后，愿西安财经大学越来越好，愿老师们工作顺利，万事胜意，也愿我的同学们前途似锦！