安徽财经大学字

**本科毕业设计**

|  |  |
| --- | --- |
| **题 目** | 新闻文本自动分类算法研究及系统设计 |
| **学 院** | **管理科学与工程学院** |
| **专 业** | **计算机科学与技术** |
| **班 级** | **20计科4班** |
| **学 号** | **20200771** |
| **姓 名** | **陈渊琪** |
| **指导老师** | **张晓春** |

**2023 年 3 月**

安徽财经大学管理科学与工程学院

本科生毕业论文（设计）诚信承诺书

本人承诺：

1.所呈交的毕业论文（设计）《 毕业设计过程管理系统》，是在认真学习理解《安徽财经大学学位论文作假行为处理办法》和《管理科学与工程学院本科毕业论文（设计）工作管理办法》后，保质保量独立完成的，没有弄虚作假，没有抄袭别人的内容；

2.毕业论文（设计）所使用的相关资料、数据、观点等均真实可靠，文中所有引用的他人观点、材料、数据、图表均已注释说明来源；

3.毕业论文（设计）中无抄袭、剽窃或不正当引用他人学术观点、思想和学术成果，伪造、篡改数据的情况；

4.本人已被告知并清楚：学院对毕业论文（设计）中的抄袭、剽窃、弄虚作假等违反学术规范的行为将严肃处理，并可能导致毕业论文（设计）成绩不合格，无法正常毕业、取消学士学位资格或注销并追回已发放的毕业证书、学士学位证书等严重后果；

5.若在省教育厅、学校、学院组织的毕业论文（设计）检查中，被发现有抄袭、剽窃、弄虚作假等违反学术规范的行为，本人愿意接受学院按有关规定给予的处理，并承担相应责任。

学生（签名）：

年 月 日

指导老师（签名）：

年 月 日

**新闻文本自动分类算法研究及系统实现**

**摘 要**

正文：小四 宋体 1.5倍行距 段落首行缩进2字符

**关键词：xx**；xx

**Research and system implementation of news text automatic Classification Algorithm**

**（times new roman 二号 加粗 1.5倍行间距 段前段后0.5行 居中）**

**Abstract**

正文：小四 times new roman 1.5倍行距 首行缩进2字符

**Keywords： xx; xx**

# 目 录

**一级标题三号加粗 行间距1.5倍 段前段后0.5行或6磅**

**中文宋体，英文和字符 times new roman**

# 1 引言

在如今的网络信息时代，网络数据渗透在我们日常生活、工作和学习的方方面面，人们无时无刻地利用电子设备和网络生产数据，传播数据，其中文本数据占据大量比重。在互联网信息流动的过程中，文本数据具有在内容上复杂多样，在规模上数量庞大的特性，其中最具有代表意义的网络新闻是人们获取社会信息，表述观点诉求的有效途径。然而如何从海量的网络新闻中精准地筛选出满足用户需求的新闻信息，仍然是值得关注的热点问题。因此，在当今移动互联网日益成熟的大环境下，针对用户多样化的需求，对海量新闻文本进行有效的分类和管理具有非常重要的研究意义。

得益于人工智能与机器学习的快速发展，深度学习算法在文本分类领域广泛应用。文本分类方法从大类上可以分为单标签文本分类和多标签文本分类，目前单标签文本分类技术已相对成熟，然而新闻文本所包含的信息复杂多元，以单个标签对其进行归类往往是不全面的，一则新闻可提取出多个主题，每个主题对应一个标签，从而一则新闻可包含多个分类，即多标签文本分类。本文拟利用深度学习理论知识，结合新闻文本分类具体实例，对多种文本分类算法进行对比分析，尝试选取一种算法具体分析，并利用python语言基于该算法实现多标签新闻文本自动分类系统。

# 2 可行性研究

该章节简要阐述了新闻文本分类的应用场景与开发价值，从经济可行性、技术可行性、操作可行性和法律可行性三个角度对新闻文本自动分类系统的开发可行性进行分析，最后对系统开发周期进行预估和计划。

## 2.1 需求概述

当今社会，互联网应用在人类日常生活中无处不在，每天都有亿万条数据从互联网应用中产生，对这些数据进行人工分类是一项耗费成本巨大并且执行效率很低的任务。针对大数据中的海量文本信息，自然语言处理领域的研究实现了自动分类文本数据，高效分析文本信息。然而大多数研究都是基于英文文本数据，中文文本分类的技术研究仍然是我们关注的热点。

大数据时代，人们每天都在各式各样的互联网设备中生产数据、分析数据、利用数据，在这亿万条数据中，文本数据占据了大量的比重。在互联网信息流动的过程中，文本数据具有在内容上复杂多样，在规模上数量庞大的特性，其中最具有代表意义的网络新闻是人们获取社会信息，表述观点诉求的有效途径。截至 2021 年 6 月，我国网络新闻用户规模达 7.60亿，占网民整体的 75.2%，微博、微信、抖音等互联网热点应用无时无刻的向庞大的用户群体推送着各式各样的网络新闻，然而如何从海量的网络新闻中精准地筛选出满足用户需求的新闻信息，仍然是值得关注的热点问题。因此，在当今移动互联网日益成熟的大环境下，针对用户多样化的需求，对海量新闻文本进行有效的分类和管理具有非常重要的研究意义。

## 2.2 系统可行性

在进行软件项目开发之前，需对项目的开发价值进行评测分析，以下分别从经济可行性、技术可行性、操作可行性和法律可行性四个角度进行阐述。

### **2.2.1经济可行性**

**（1）成本估算**

|  |  |
| --- | --- |
| 项目名称 | 成本估计(单位/元) |
| 软件开发 | 5000 |
| 服务器租赁 | 500 |
| 维护工资 | 3000 |
| 人力开销 | 2000 |
| 合计 | 10050 |

**（2）效益**

|  |  |
| --- | --- |
| 项目 | 收益（元/年） |
| 一次性收益 | 无 |
| 经常性收益（下载文献） | 80000 |
| 不可定量收益 | 无 |
| 企业定制服务 | 20000 |
| 总计 | 100000 |

**（4）收益/投资比**

一次性支出：26000元

经常性支出：24000元/年

收益：100000元/年

收益/投资比：100000\*5/(26000+24000\*5)=3.424657

**（5）货币的时间价值**

五年预计收益：100000\*5-（26000+24000\*5）=354000（元）

文献管理系统的前期软件开发阶段，需要耗费巨大的人力，投资的成本和未来获得的效益会随着软件每年的货币时间价值有所改变。文献管理系统五年收益为354000元。实际收益需要通过货币时间价值来进行核算。用利率的形式表示货币的时间价值。假设年利率为i，如果现在存入P元，则n年后可以得到的钱数为：

H:\temp\ksohtml14132\wps1.jpg (1)

这也就是P元钱在n年后的价值。反之，如果n年后能收入F元钱，那么这些钱的现在价值是：

H:\temp\ksohtml14132\wps2.jpg (2)

假定年利率为12%，利用上面计算货币现在价值的公式可以算出系统5年预计收益的现在价值，如下表2-5所示。

表2-5 将来收入折算成现在值

|  |  |  |
| --- | --- | --- |
| 将来值（元） | （1+i)^n | 现在值（元） |
| 354000 | 1.7623 | 200873 |

**（6）投资回收期估算**

文献管理系统基础建设成本约为26000元，第一年内收入为100000元，软件投资的回收期约为

 （3)

因此软件的投资回收期约为0.26年，本系统开发成本较低，维修成本适中，软件投资回报期较短可以较快获得利润，用户群体丰富且需求量大，经济利益客观值得投资。

### **2.2.2技术可行性**

新闻文本自动分类系统的实现需要采用机器学习算法，主要包括文本特征提取、特征向量化、分类器训练等技术。在Python语言中，有众多的机器学习库和自然语言处理库可供选择，如scikit-learn、Tensorflow、NLTK等。这些库提供了强大的功能和易于使用的API，使得机器学习算法的实现变得更加简单和高效。此外，Python还具有丰富的第三方包和库，可用于数据处理、可视化和模型部署等方面。因此，从技术上来说，新闻文本自动分类系统的实现具有可行性。

### **2.2.3 操作可行性**

新闻文本自动分类系统的操作主要包括文本数据的预处理、模型训练和模型应用等环节。其中，文本数据的预处理需要进行分词、去停用词、词形还原等处理；模型训练需要选择合适的算法和参数，进行训练和调优；模型应用则需要将待分类文本进行特征提取、特征向量化、分类器预测等处理。这些操作需要熟悉Python编程语言和机器学习算法的相关知识和技能。同时，对于大规模的文本数据集，还需要具备数据处理和存储的能力。因此，该系统的操作可行性需要基于相关专业知识和实践经验。

此外，新闻文本自动分类系统的操作还需要考虑用户体验和易用性。系统应该提供易于操作和友好的用户界面，帮助用户快速和准确地完成分类操作。同时，系统应该具备良好的稳定性和可扩展性，以适应不断变化的用户需求和技术环境。

### **2.2.4 法律可行性**

新闻文本自动分类系统需要收集和使用大量的新闻文本数据，因此需要考虑数据隐私保护和知识产权等法律问题。在数据隐私保护方面，需要遵守相关法律法规，如个人信息保护法等，保护用户的个人信息和隐私。在知识产权方面，需要确保数据采集和使用的合法性，避免侵犯他人的知识产权。同时，还需要考虑到文本数据的版权问题，以避免侵犯新闻媒体的版权。因此，在实现新闻文本自动分类系统时，需要仔细考虑相关法律法规，遵守合法规定。同时，也需要保障用户的知情权和选择权，明确告知用户数据采集和使用的目的和方式，避免侵犯用户隐私和权益。

除此之外，还需要考虑系统的合规性和安全性。合规性方面，需要遵循相关法律法规和政策规定，确保系统的合法性和合规性。安全性方面，需要采取必要的技术措施，保障系统和数据的安全性和可靠性，防范数据泄露和黑客攻击等安全风险。

因此，在实现新闻文本自动分类系统时，需要综合考虑技术、操作和法律等多个方面的因素，确保系统的可行性和合规性。同时，还需要不断更新和改进系统，以满足不断变化的用户需求和技术环境。

## 2.3 项目进度计划

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 项目进度名称 | 开始时间 | 所需时长（天） | 结束时间 |
| 1 | 问题定义 | 2023/2/27 | 5 | 2023/3/3 |
| 2 | 可行性分析 | 2023/3/4 | 7 | 2023/3/10 |
| 3 | 需求分析 | 2023/3/11 | 8 | 2023/3/19 |
| 4 | 总体设计 | 2023/3/20 | 16 | 2023/4/4 |
| 5 | 详细设计 | 2023/4/5 | 19 | 2023/4/23 |
| 6 | 编码和单元测试 | 2023/4/24 | 11 | 2023/5/4 |
| 7 | 综合测试 | 2023/5/5 | 21 | 2023/5/25 |
| 8 | 软件维护 | 2023/5/26 |  |  |

1.问题定义：确定选题、问题分析。分析新闻文本自动分类系统将要解决的问题；确定工作目标和工程规模。

2.可行性分析：从经济、技术、操作和法律四个层面分析新闻文本自动分类系统对于问题的解决方案，是否值得对该项目进行资源投入。

3.需求分析：从用户角度对该管理系统提出需求，分别从系统功能需求、性能需求、安全性需求、可靠性需求和系统兼容性需求几个角度分析。

4.总体设计：梳理新闻文本自动分类系统的体系结构，将程序模块化。

5.详细设计：对新闻文本自动分类系统的各个小模块进行详细设计，确定模块功能所需要的算法和数据结构。

6.编码和单元测试：将设计好的各个模块以python语言实现，并对各个模块的功能进行测试。

7.综合测试：将经过单元测试检验的各个小模块装配起来，根据系统功能进行综合测试。

8.软件维护：系统正式上线后，诊断后修复使用过程中出现的软件错误，并修改软件以适应环境的变化，满足用户要求

# 3 需求分析

第二章对新闻文本自动分类系统的开发可行性进行了多方面的分析，本章主要从用户需求角度对软件的功能和性能等方面进行概述，并设计系统的顶层数据流图和各个功能模块的数据流图，根据各个功能模块设计系统数据库。

## 3.1 系统综合需求

新闻文本自动分类系统分为两大部分，分别面向用户和管理员，管理员通过该系统从网络中爬取新闻数据，由系统核心算法将这些新闻按标签分类，并呈现给用户；用户可根据分类标签选择对应的新闻。

### **3.1.1功能需求概述**

新闻文本自动分类系统通过爬虫程序从网络爬取新闻数据，利用深度学习算法对新闻数据进行分类，最后呈现给用户。主要功能可以分为以下四个部分：

1.数据爬取：系统所使用的数据均来源于网络，使用爬虫程序高效快速的爬取新闻数据。

2.文本归类：利用所爬取的新闻数据，基于深度学习算法对其进行归类。

3.用户管理：保存并管理普通用户和管理员的账号注册信息、个人信息。

4.新闻数据存储：爬取的新闻数据以及分类之后的数据保存至对应数据库存储。

### **3.1.2 系统性能需求概述**

1.系统的稳定性：本系统要求在没有人为干预的情况下可以稳定的运行，并且系统自身可以处理一些异常和错误，要求系统平均每天出错的次数不超过 3 次。

2.响应时间：每条新闻的平均爬取时间不超过5秒，在10000条数据以内文本分类时间不超过10分钟。

### **3.1.3 其他需求概述**

1.系统的可靠性：首先本系统需要要保证能够向用户展示新闻文本正确的分类结果，其次系统还需要精准捕捉用户的异常操作，以及系统内部出现的异常和错误，及时向用户以及管理员反馈错误信息并且记录到管理日志中。

2.系统的易用性：要求本系统界面简洁明了，前端交互界面能够让用户一目了然，提高用互体验好感。

3.系统可恢复性：要求本系统在出现异常错误是能够通过人工操作，甚至不需要人工操作可以恢复到正常运行的状态，管理员对数据库进行操作出现错误时，必须保证数据回滚，保护数据安全。

4.系统的安全性：首先本系统必须保证用户的传输的隐私数据的保护，其次要保证系统数据库的安全性，避免数据泄露。

## 3.2 数据流图分析

### **3.2.1 顶层数据流图**

### **3.2.2 功能级数据流图**

### **3.2.3 数据流图细化与分解**

## 3.3 数据字典

（由于数据字典定义内容非常多，可针对每一小类选择代表性的示例展示）

### **3.3.1数据流条目定义**

实例：

表3-X 管理员信息数据字典定义

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据流 | | | | |
| 系统名：AUFE学生选课系统 | 编号：D-001 | | | |
| 条目名：管理员信息 | 别名： | | | |
| 来源：管理员 | 去处：管理员 | | | |
| 数据流结构：  管理员：{管理员账号+密码+邮箱+手机号码}所有管理员 | | | | |
| 简要说明：管理员数据流记录着管理员的基本信息，可用于管理员的信息注册、查看、修改，登录验证等处理 | | | | |
| 修改记录 | 编写 | 喻杰 | 日期 | 2023.10.7 |
| 审核 | 刘士强 | 日期 | 2023.10.8 |

根据3.2节数据流图分析情况可知，系统主要涉及的数据项包括：xx, xx, xx, xx和xx分别定义如图3-X~3-Y所示。

（每个类型的数据字典定义前应有一段总结性的开场白。）

### **3.3.2数据存储条目定义**

实例：

表3-X 学生记录数据字典定义

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据存储 | | | | |
| 系统名：AUFE学生选课系统 | 编号：D-006 | | | |
| 条目名：学生记录 | 别名： | | | |
| 存储组织：每个学生一条记录 | 记录数：10^8 | | 主关键字：学生学号 | |
| 记录组成：  项名：学号 密码 手机号码 邮箱 学院 专业 班级 姓名  长度： 8 20 11 30 15 15 10 10 | | | | |
| 简要说明：存储每个学生的基本信息 | | | | |
| 修改记录 | 编写 | 喻杰 | 日期 | 2023.10.7 |
| 审核 | 刘士强 | 日期 | 2023.10.8 |

### **3.3.3数据处理储条目定义**

实例：

表3-X 用户登录管理数据字典定义

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据处理 | | | | |
| 系统名：AUFE学生选课系统 | 编号：D-010 | | | |
| 条目名：用户登录管理 | 别名： | | | |
| 输入数据流：学生信息 | 输出数据流：学生信息及消息反馈 | | | |
| 加工逻辑：   1. 账号信息合法性检查； 2. 账号匹配； 3. 根据用户类别禁用或开放相应功能。 | | | | |
| 简要说明：对学生信息进行管理 | | | | |
| 修改记录 | 编写 | 喻杰 | 日期 | 2023.10.7 |
| 审核 | 刘士强 | 日期 | 2023.10.8 |

### **3.3.4数据项条目定义**

## 3.4实体联系分析

### **3.4.1 实体提取及实体图**

表3-18 实体提取结果

|  |  |  |
| --- | --- | --- |
| 实体 | 主码 | 其他属性 |
| 用户 | 用户ID | 用户密码，姓名，性别，用户权限等级 |
| 文献 | 文献编号 | 文献编号，名称，作者，文献来源，文献类别，文献发表时间，文献关键词 |
| 管理员 | 管理员编号 | 姓名，性别，级别，密码 |
| 检索记录 | 检索序号 | 检索日期 |

图 3-2 文献管理系统-文献属性图



图 3-3 文献管理系统-用户属性图



图 3-4 文献管理系统-管理员属性图

### **3.4.2 实体联系图**

# 4系统设计

概述段落

## 4.1总体设计

概述段落

### **4.1.1 系统功能结构**



图4-1 变换型数据流图示例



图4-2 变换型数据流图对应的软件结构图



图4-3 事务型数据流图示例



图4-4 事务型数据流图对应的软件结构图

按照变化型数据流图和事务型数据流图分析软件结构

### **4.1.2 系统层次图**

## 4.2 详细设计

概述段落

### **4.2.1 代表性模块设计**

#### 4.2.1.1 xx模块设计简介

（四级标题： 中文黑体 英文times new roman 1.5倍行间距 段前段后0.5行）

**（1）xx模块简介**

**（2）xx模块流程图（2，3, 4）可选择一种**

**（3）xx模块NS图**

**（4）xx模块PAD图**

**（5）针对复杂组合逻辑的判定表或判定树分析**

#### 4.2.1.2 YY模块设计简介

**（1）yy模块简介**

**（2）yy模块流程图（2，3, 4）可选择一种**

**（3）yy模块NS图**

**（4）yy模块PAD图**

**（5）针对复杂组合逻辑的判定表或判定树分析**

### **4.2.2 系统数据库设计**

#### 4.2.2.1 数据库的逻辑设计

#### 4.2.2.2 数据库的表设计及表与表之间的关联

# 5 编码与测试

概述段落

## 5.1 编码

### **5.1.1 编码规则简介**

### **5.1.2代表性模块示例**

## 5.2测试

概述段落

### **5.2.1 白盒测试**

#### 5.2.1.1 xx模块独立路径分析

#### 5.2.1.2 xx模块测试用例设计

#### 5.2.1.3 xx模块测试情况分析

### **5.2.2 黑盒测试**

#### 5.2.2.1 xx模块的等价类分析

#### 5.2.2.2 xx模块测试用例设计

#### 5.2.2.3 xx模块测试情况分析

# 6 系统使用说明

概述段落

## 6.1 系统运行环境和配置

## 6.2 系统操作说明（按照结构图或层次图的框架依次介绍）

### **6.2.1 XX1模块说明**

### **6.2.2 XX2模块说明**

### **6.2.3 XX3模块说明**

### **6.2.4 XX4模块说明**

# 7 总结

# 参考文献

按国家标准GB3469执行。

# 致谢