|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 西华大学  毕业设计说明书     |  |  |  | | --- | --- | --- | | 题目 | : | 基于新闻的大数据分析系统  —— 数据采集与后台 | | 学院 | : | 计算机与软件工程学院 | | 年级专业 | : | 2016级计算机科学与技术 | | 姓名 | : | 欧贤 | | 学号 | : | 3120160901430 | | 指导教师 | : | 魏冬梅 | | 完成时间 | : | 2020年05月24日 | |

# 摘要

本文主要完成基于新闻的大数据分析系统的设计与测试，实现自动获取海量新浪新闻数据，构建关键词袋，自动提取正面、负面等感情特征信息，辅助用户了解舆情信息和趋势的功能。系统通过网站的形式进行与用户之间的交互，网站中的用户可分为网站管理员、网站普通用户、游客三种角色，他们各自有不同的权限，其中网站管理员的权限最大，包括对用户信息与新闻数据的管理；而网站普通用户可以享有查看新闻列表及指定新闻相关信息、查看新闻及评论的分析结果等功能；游客只享有查看新闻列表功能。网站分为前端交互页面和后台管理页面，管理员须在前端登录页面通过管理员账户进入后台管理页面，才能进行数据管理操作，普通用户可在前台展示页面调用部分后台管理页面功能，操作自身信息。

基于新闻的大数据分析系统实现了从新浪网抓取新闻数据，并识别文本情感类别，给决策者提供可靠的舆论倾向，对舆情进行有效的监控、预警及疏导。为精准识别新闻及用户所表达的情感，判断其情感趋向，本文选取了两种情感分类方法——基于依存句法分析的情感分类与基于深度学习模型的情感分类。基于依存句法分析的情感分析方法主要针对新闻报道内容；基于深度学习模型的情感分类方法主要针对新闻评论信息，其文本情感分类模型是基于LSTM（long-short-term memory）搭建的。

【关键词】 网络新闻及评论；情感分析；关键词分析；LSTM；依存句法

# Abstract

This article mainly completes the design and test of news-based big data analysis system, realizes the automatic acquisition of massive Sina news data, constructs keyword bags, and automatically extracts positive and negative emotional feature information to assist users in understanding public opinion information and trends. The system interacts with users through the form of a website. The users on the website can be divided into three roles: website administrator, website ordinary user, and tourist. They each have different permissions, of which the website administrator has the most permissions, including Management of user information and news data; ordinary users of the website can enjoy the functions of viewing the news list and designated news-related information, viewing the analysis results of news and comments, etc .; visitors only have the function of viewing the news list. The website is divided into a front-end interactive page and a background management page. Administrators must enter the background management page through the administrator account on the front-end login page to perform data management operations. Ordinary users can call some background management page functions at the front desk display page to operate their own information. .

The big data analysis system based on news realizes grabbing news data from Sina.com and identifying text sentiment categories, providing decision makers with reliable public opinion tendencies, and effectively monitoring, warning and guiding public opinion. In order to accurately identify the news and the emotions expressed by users, and to judge their emotion trends, this article selects two emotion classification methods-emotion classification based on dependency syntax analysis and emotion classification based on deep learning models. The sentiment analysis method based on dependency syntax analysis is mainly aimed at news report content; the sentiment classification method based on deep learning model is mainly aimed at news comment information, and its text sentiment classification model is based on LSTM (long-short-term memory).

【Key words】 Internet news and commentary; sentiment analysis; keyword analysis; LSTM; dependency syntax

**目录**

[1．绪论 1](#_Toc19827)

[1.1 研究背景 1](#_Toc19556)

[1.2 研究意义 1](#_Toc28295)

[1.3 研究内容 2](#_Toc18918)

[2. 系统概述 3](#_Toc22112)

[2.1 用户的特点 3](#_Toc26974)

[2.2 系统基本信息 3](#_Toc2710)

[2.3 系统开发工具 4](#_Toc19975)

[2.4 系统可行性分析 5](#_Toc11604)

[3．需求分析 6](#_Toc30145)

[3.1 系统功能需求 6](#_Toc32751)

[3.2 性能设计需求 6](#_Toc13231)

[3.3 条件与限制 7](#_Toc27524)

[3.4 系统用例图 7](#_Toc11516)

[4．概要设计 9](#_Toc18205)

[4.1 处理流程 9](#_Toc20685)

[4.1.1 情感分析处理流程 9](#_Toc4209)

[4.1.2 情感可视化处理流程 11](#_Toc19614)

[4.2 总体结构和模块外部设计 12](#_Toc13118)

[4.3 功能分配 13](#_Toc18584)

[4.4 接口设计 14](#_Toc7367)

[4.4.1 外部接口 14](#_Toc21424)

[4.4.2 内部接口 14](#_Toc1030)

[4.5 数据结构设计 15](#_Toc25986)

[4.5.1 逻辑结构设计 15](#_Toc21637)

[4.5.2 物理结构设计 16](#_Toc19528)

[4.6 运行设计 17](#_Toc22791)

[4.6.1 运行机制 17](#_Toc20327)

[4.6.2 运行控制 18](#_Toc27743)

[4.7 出错处理设计 19](#_Toc21339)

[4.7.1 出错输出信息 19](#_Toc27183)

[4.7.2 出错处理对策 20](#_Toc29195)

[4.8 维护设计 20](#_Toc9825)

[5. 详细设计 21](#_Toc21203)

[5.1 软件结构 21](#_Toc10642)

[5.2 数据采集与预处理模块 22](#_Toc10179)

[5.2.1 新闻爬取模块 22](#_Toc8660)

[5.2.2 新闻评论爬取模块 24](#_Toc8673)

[5.3 情感分析模块 25](#_Toc12905)

[5.3.1 新闻情感分类模块 26](#_Toc31858)

[5.3.2 评论情感分类模块 30](#_Toc25189)

[5.3.3 文本聚类模块 34](#_Toc32661)

[5.4 用户交互模块 36](#_Toc10587)

[5.4.1 后台用户管理模块 36](#_Toc7351)

[5.4.2 后台新闻管理模块 39](#_Toc21124)

[6. 运行测试 41](#_Toc7084)

[6.1 测试计划 41](#_Toc29180)

[6.2 测试结果分析 41](#_Toc17765)

[6.2.1 数据爬取与预处理测试 41](#_Toc17675)

[6.2.2 情感分类功能测试 42](#_Toc8617)

[7. 运行结果 44](#_Toc23956)

[7.1 数据爬取与预处理运行结果 44](#_Toc18228)

[7.2 情感分类运行结果 45](#_Toc14880)

[7.3 文本聚类运行结果 46](#_Toc31271)

[总结 47](#_Toc14113)

[致谢 48](#_Toc29366)

[参考文献 49](#_Toc27484)

# 1．绪论

## 1.1 研究背景

随着经济的发展，各种社交平台的不断兴起，截至2020年3月，我国网络用户人数已增长到9.04亿，互联网普及率已达到64.5%[1]。 随着网民规模的增大，网络上用户的生成内容越来越多，产生大量的文本信息，如新闻、微博、博客等，研究发现面对如此庞大且富有情绪表达的文本信息，如果从这些信息中挖掘有价值的隐藏信息和探索提取与之相关联的特征，对聚类分析、信息预测和推荐系统等有潜在的应用价值[2]。因此，近年来，情感分析已成为计算机语言学领域热门的研究课题。

情感分析，又被称作情感挖掘、主观分析等，是指通过自然语言处理、文本挖掘等方法来对包含主观情感的文本信息进行分析、处理，然后将结果归纳总结并推算，最终得到文本包含的观点、情感等信息的一个过程[3]。情感分析强调面向某个领域，在对评论对象进行情感分类的基础上，将情感分类的结果应用在某个具体的任务上，从而发现规律和利用规律。

## 1.2 研究意义

由于互联网中的信息传播都是实时、随机、突然和隐藏地，所以也给了一些别有用心的人可以操作并利用的空间，如果不能有效的监控、预警及疏导网民舆论的方向，那将会对社会的稳定性造成很大影响。利用计算机去分析评论信息，对信息数据进行情感分析是监控、预警及疏导网民舆论的有效手段之一。

本系统目标为在庞大的数据集中精准的区分文本的情感类型（正面、中性、负面三类）。面对浩如烟海的新闻及其评论信息，精确识别蕴藏在其中的情感倾向，对舆情的有效监控、预警及疏导，对舆情生态系统的良性发展都有着十分重要的意义[4]。

## 1.3 研究内容

本文是围绕着新闻及其评论的情感分析展开的研究，研究内容包括以下几点：

1. 数据获取：本文通过网络爬虫技术动态的从新浪网爬取新闻及其评论信息，将其进行数据清洗后保存到数据库与本地。
2. 文本情感分类：本文研究的文本分为两类：新闻正文，其文本情感分类方法是依存句法分析；新闻评论文本，其文本情感分类方法是一种深度学习模型，是一种基于长短期记忆人工神经网络建立的，应用于新闻评论文本情感分类的深度学习模型。
3. 关键词分析：本系统主要通过TextRank算法、北京大学分词工具pkuseg两种技术手段对文本进行关键词分析，并通过matplotlib库将分析结果可视化呈现；
4. 文本聚类：本系统通过AP聚类算法对同一新闻或者同一类新闻下的评论文本进行聚类操作。

# 2. 系统概述

## 2.1 用户的特点

网站中的用户按系统权限可分为三种角色——管理员、普通用户、游客，他们所拥有的权限各不相同，其中管理员拥有管理系统中的用户信息、新闻及其评论数据的权限；普通用户享有查看新闻信息、查看新闻及评论的分析结果等功能；游客只享有查看首页新闻列表功能。

表 2-1 用户描述表

|  |  |
| --- | --- |
| **用户类** | **描述** |
| 管理员 | 管理员拥有管理后台的权限，包括对系统内用户的信息的管理、新闻及评论信息的管理，同时也负责系统数据库与系统故障的维护工作。 |
| 普通用户 | 普通用户可以对新闻进行查询及查看，还可以选择新闻或评论进行情感分析以及其它基本功能。 |
| 游客 | 游客只能查看首页新闻列表，不能够选择新闻或评论进行情感分析以及其它一些基本功能的操作。 |

## 2.2 系统基本信息

本系统主要的运行环境为Windows系统，推荐使用Windows10版本；系统开发环境为JDK 8与Python 3.6；系统功能需求有如下几点：

（1）基本模块(用户注册、 登录、查询 、个人中心)；

（2）新闻及评论信息采集与预处理（如网页爬取/数据清洗等）；

（3）情感分析 （主题抽取、情感词抽取、情感分类或评分等）；

（4）后台管理(用户信息管理 、新闻及其评论数据管理)；

## 2.3 系统开发工具

本系统Python开发工具为PyCharm 2018，PyCharm 支持全面的 Python 开发。无论是数据科学还是非数据科学项目，Matplotlib 和 NumPy 等工具都可以轻松集成，这意味着在数据分析操作期间可以轻松地使用阵列查看器和交互式图表。

Python 管理平台选取的是Anaconda3， Anaconda不仅可以帮助开发人员下载、安装、更新、卸载工具包与控制项目的Python版本，还能在安装时能自动安装相应的依赖包，同时还能针对不一样的项目创建属于其自身的虚拟环境，大大简化了我们的开发工作。

Java语言开发工具为Eclipse 2018，Eclipse在java语言开发、编写、查错、编译、帮助等各方面都非常方便。

版本控制工具为Git，Git具有版本库本地化，支持脱机提交，相对独立而不影响协同开发，支持分支的快速切换，方便合并，良好的合并性能，分布式存储库，无单点故障，良好的内容完整性等特性。

表 2-2 系统开发工具表

|  |  |  |
| --- | --- | --- |
| **系统名称** | | 基于新闻的大数据分析系统 |
| **系统功能** | | 新闻及新闻评论情感分析 |
| **系统负责人** | | 欧贤；丁一钉 |
| **系统开发工具** | **a.编程语言工具** | PyCharm 2018；Eclipse 2018 |
| **b.数据库及管理工具** | MySQL 8.0；Navicat 12 |
| **c.文档编写工具** | WPS 2019 |
| **d.版本控制** | Git |

## 2.4 系统可行性分析

对系统进行可行性研究是系统分析阶段第二项活动，此项活动的主要目的是：进一步明确系统的目标、规模和功能，调查分析系统开发的背景、必要性和意义，根据需要和可能提出初步计划和系统开发计划。调查与分析的结果显示，本系统的设计包括以下三个可行性：

1. 技术可行性:本系统是在Windows 10操作平台上运行的，数据库管理系统选用的是mysql的8.0版本，本系统的编程语言开发平台选用PyCharm 2018+Eclipse 2018，其中PyCharm 在人工智能和机器学习方面有着显著的优势；Eclipse 是当下主流的Java IDE工具之一。
2. 经济可行性:本系统对网络舆情分析与监控有辅助作用，精准识别新闻及用户所表达的情感，判断其情感趋向，给决策者提供可靠的舆论倾向。
3. 操作可行性:本系统界面美观、功能适用、使用简洁，运行环境要求低。

# 3．需求分析

## 3.1 系统功能需求

经过详细分析，确定了该项目有如下功能需求：自动获取海量新浪新闻数据，构建关键词袋，自动提取正面、负面等感情特征信息，辅助用户了解舆情信息和趋势。具体包括：

1. 新闻及其评论信息采集与预处理：动态爬取新浪网新闻及其评论信息，并对爬取的数据进行清洗，然后保存到数据库；
2. 关键词分析：提取文本关键词、构建词袋等；
3. 文本情感分析：对文本进行情感分类或评分等；
4. 用户管理：管理员管理用户信息，用户管理个人信息；
5. 新闻数据管理：管理员对新闻数据增删改。

## 3.2 性能设计需求

(1)系统的稳定性需求：系统应具备稳定性，在系统运行期间不出现自动终止，程序bug等；

(2)系统的可维护性需求：系统的可维护性需求与系统的稳定性需求是分不开的，当系统运行不稳定，出现数据丢失错误、运行中断时，要保证能够恢复或者还原[5]；

(3)系统的易用性需求：系统要界面美观、功能适用、使用简洁；

(4)系统的完整性需求：系统除了有必要的功能界面，还应该有数据管理界面，如：用户信息界面，后台管理界面等。

(5)系统的可扩展性需求：后期有必要改进系统的开发和建设，在系统设计和开发之前应该充分考虑到系统的可扩展性需求。

## 3.3 条件与限制

只有注册用户才能选择新闻或其评论进行情感分析，游客通过注册之后才能使用这些功能。只有系统的管理员拥有数据库中的用户信息与新闻及评论数据增删查改操作的权限，注册用户只能查看、修改部分个人信息。

## 3.4 系统用例图

本系统共有三种参与者——管理员、游客与注册用户。管理员拥有用户信息的增删查权限与新闻及评论数据增删查改的权限；游客只有通过注册之后，才能拥有查看分析结果等权限，注册用户拥有查看、修改个人信息与选择新闻及评论进行数据分析并查看分析结果等权限。

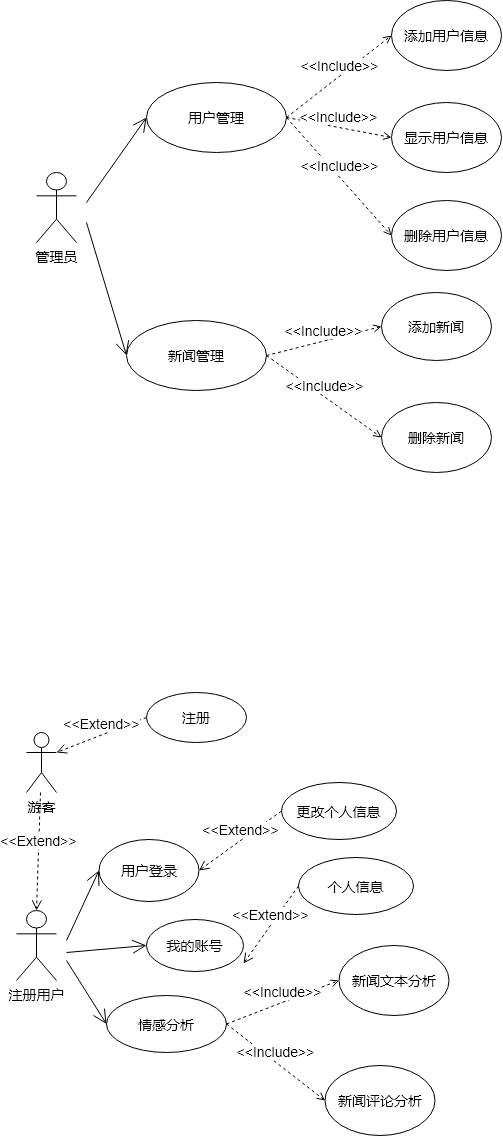


图3-1 管理员用例图

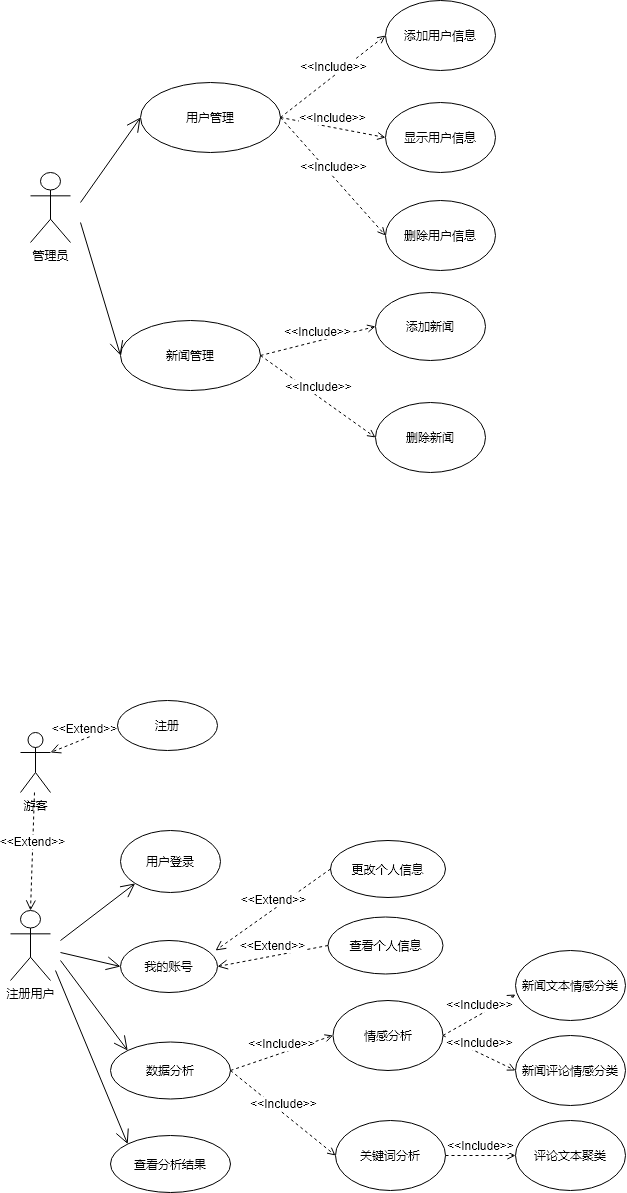


图3-2 游客、注册用户用例图

# 4．概要设计

## 4.1 处理流程

本部分主要介绍系统后台文本处理流程，主要包括新闻及评论的情感分析处理流程，新闻及评论分析结果可视化处理流程。

### 4.1.1 情感分析处理流程

本系统是通过基于依存句法分析方法对新闻文本进行情感分类，通过基于深度学习模型的方法对新闻的评论进行情感分类。

处理流程如图4-1所示，用户登录系统，选择新闻，然后选择分析对象是新闻正文还是新闻评论，若选择新闻正文分析，则从数据库读取新闻正文，通过依存句法分析方法计算文本情感得分；若选择新闻评论分析，则从数据库读取评论数据存储位置，读取评论数据，通过深度学习模型判断每条评论文本的情感趋向；然后展示情感得分或情感类型。

依存句法分析的基本思想是：首先判断每个词是否是情感词，如果该词为情感词，然后根据依存句法的分析，获取关联情感词的程度副词，最后再根据情感词和其关联的程度副词来计算整个句子的情感得分。

本系统是基于LSTM神经网络算法搭建的用于新闻评论文本情感分类的深度学习模型，其主要处理流程包括以下两个步骤：

1.将处理好的评论语料转换为语义词向量;

2.将转换的语义词向量通过不同的语义合成方法合成所对应句子或文档的特征表达式，最后通过LSTM神经网络进行分类[6]。

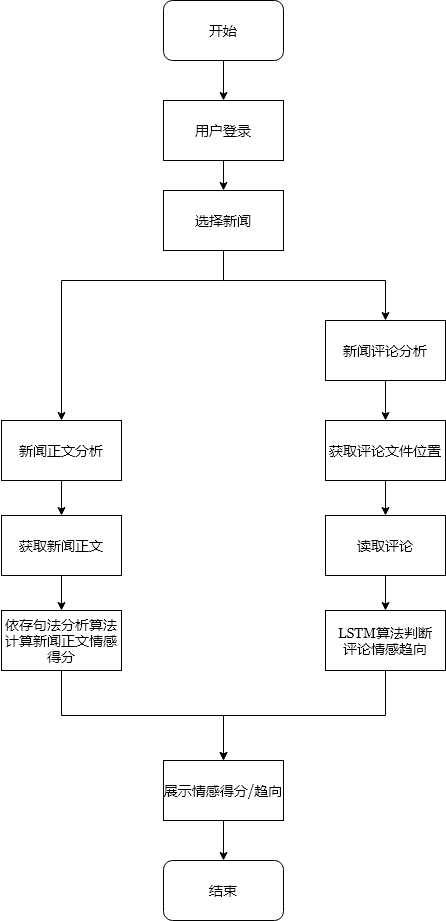
****

图4-1 情感分析流程图

### 4.1.2 情感可视化处理流程

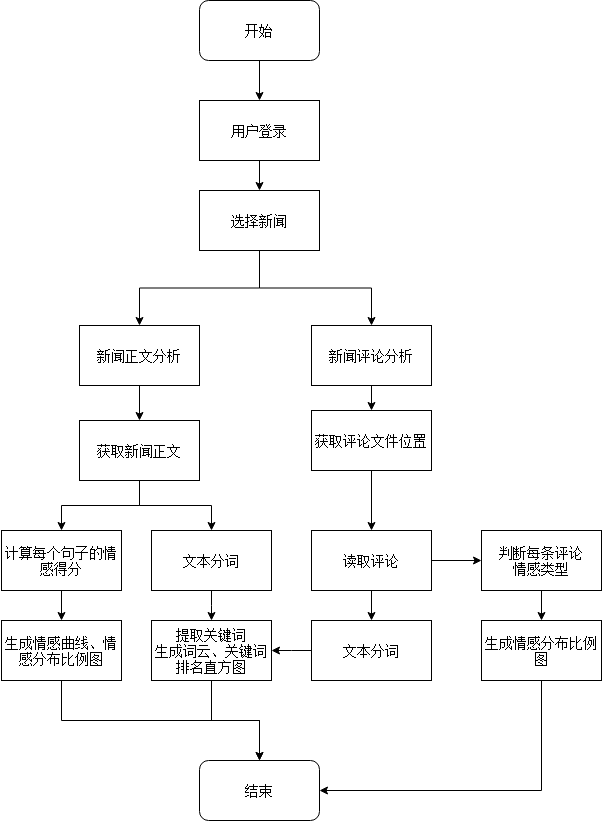


图4-2 情感可视化流程图

## 4.2 总体结构和模块外部设计

考虑到系统的简单易操作性，本文将系统的总体功能结构拆分为：用户注册、用户登录、用户自助服务、新闻查询、情感分类、分类结果展示、后台管理等模块，详见表4-1。

表 4-1 总体结构及模块表

|  |  |  |
| --- | --- | --- |
| **编号** | **功能名称** | **功能描述** |
| **1** | **用户注册** | 检验和完成用户注册并保存用户注册信息 |
| **2** | **用户登录** | 完成用户登录和检验用户登录信息 |
| **3** | **用户自助服务** |  |
| **3.1** | 用户修改个人信息 | 用户登录以后可以对自己的个人信息进行维护、修改 |
| **4** | **新闻查询** | 对数据库内新闻的多种查询,方便用户查找需要的新闻 |
| **5** | **关键词分析** | 中文分词；词云制作；词频统计等 |
| **6** | **情感分类** |  |
| **6.1** | 新闻正文情感分类 | 使用依存句法分析算法计算情感得分 |
| **6.2** | 新闻评论情感分类 | 使用神经网络进行情感趋向判断 |
| **6.3** | 新闻评论文本聚类 | AP聚类算法进行文本聚类 |
| **7** | **后台管理** |  |
| 7.1 | 用户管理 | 添加、显示、删除用户信息 |
| 7.2 | 新闻数据管理 | 添加、删除新闻数据 |
| **8** | **分析结果展示** | **网页展示分析结果** |
| **9** | **数据采集** | **从新浪网爬取新闻及其评论** |
| **10** | **数据预处理** | **对爬取数据清洗并保存** |

## 4.3 功能分配

根据上一节的系统功能需求设计，现把基于新闻的大数据分析系统的功能需求具体划分为用户注册、登录、自助服务、新闻查询、新闻或评论分析、后台管理六大模块，其与程序结构的关系如下表4-2所示。

表 4-2 各项功能与程序结构的关系表

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **程**  **功 序**  **能 结**  **需 构**  **求** | **用户注册** | **用户登录** | **用户自助服务** | **新闻查询** | **新闻或评论选择** | **关键词分析** | **情感分析** | **新闻详情** | **分类显示** | **用户管理** | **新闻数据管理** |
| **用户注册** | √ |  |  |  |  |  |  |  |  |  |  |
| **用户登录** |  | √ |  |  |  |  |  |  |  |  |  |
| **用户自助服务** |  |  | √ |  |  |  |  |  |  |  |  |
| **新闻查询** |  |  |  | √ |  |  |  | √ |  |  |  |
| **新闻或评论分析** |  |  |  |  | √ | √ | √ | √ | √ |  |  |
| **后台管理** |  |  |  |  |  |  |  |  |  | √ | √ |

## 4.4 接口设计

本部分主要介绍系统的接口设计，包括外部接口与内部接口两大部分，其中外部接口又划分为用户接口、软件接口、硬件接口与硬件接口四个部分。

### 4.4.1 外部接口

(1) 用户接口

本软件属于网上数据分析系统，所以在新闻或评论分析时提示用户进行注册和登录验证。

(2) 软件接口

服务端：本软件需要JDK 8与python 3.6语言环境、MySQL 8.0数据库系统的支持。Python开发环境中通过pymysql模块连接数据库，Java开发环境中通过JDBC驱动连接数据库，实现与数据库数据之间的交互。

客户端：无特殊的需要支持的软件，只要安装有现在主流的浏览器即可。

(3) 硬件接口

因为软件是网上数据分析系统，服务端与客户端都需要网卡驱动以支持上网功能。

### 4.4.2 内部接口

由于本软件的性质是网上数据分析系统，主要的内部接口是通过各网页面之间的跳转关系和对象调用方法及所需的传递参数，来实现各模块间的数据交换。

## 4.5 数据结构设计

该部分主要给出系统所使用的各数据结构的设计及其关系，包括逻辑结构设计、物理结构设计。

### 4.5.1 逻辑结构设计

本系统主要包括管理员、一般用户、新闻数据三个实体，其中管理员与用户和新闻数据之间存在着管理关系，一般用户与新闻数据存在查看关系，具体设计如图4-3所示。

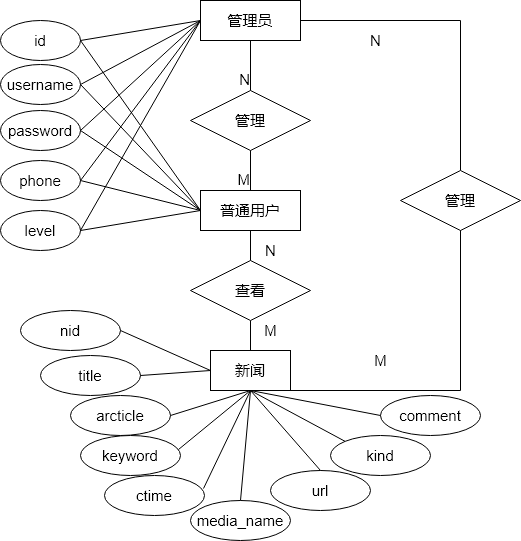


图4-3 E-R关系图

### 4.5.2 物理结构设计

本系统主要设计了两张数据库表——用户表（User）、新闻表（News），用户表用于存储用户信息，新闻表用于存储新闻相关信息，具体结构设计如表4-4、4-5所示。

表 4-3 数据库表单

|  |  |
| --- | --- |
| **数据表名** | **注释** |
| User | 用户表 |
| News | 新闻表 |

数据库名：Graduation\_Project

表4-4 User（用户表）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **序号** | **列名** | **数据类型** | **长度** | **主键** | **允许空** | **注释** |
| 1 | id | int | 11 | 是 | -1 | 用户id |
| 2 | username | varchar | 16 | 否 | 0 | 用户名 |
| 3 | password | varchar | 16 | 否 | 0 | 登陆密码 |
| 4 | phone | varchar | 11 | 否 | 0 | 手机号 |
| 5 | level | bit | 1 | 否 | 0 | 权限 |

表4-5 News（新闻表）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **序号** | **列名** | **数据类型** | **长度** | **主键** | **允许空** | **注释** |
| 1 | n\_id | varchar | 60 | 是 | -1 | 新闻id |
| 2 | title | varchar | 100 | 否 | 0 | 新闻标题 |
| 3 | article | longtext | 0 | 否 | 0 | 新闻内容 |
| 4 | keyword | varchar | 50 | 否 | 0 | 新闻关键字 |
| 5 | ctime | datetime | 0 | 否 | 0 | 发布时间 |
| 6 | media\_name | varchar | 20 | 否 | 0 | 发布媒体 |
| 7 | url | varchar | 100 | 否 | 0 | 新闻链接 |
| 8 | kind | bigint | 6 | 否 | 0 | 新闻类型 |
| 9 | comment | varchar | 40 | 否 | 0 | 新闻评论  文件位置 |

## 4.6 运行设计

该部分描述在一个系统中施加不同的外部操作时所引起的各种不同的运行设计。包括运行机制、运行控制两个部分。

### 4.6.1 运行机制

系统运行后数据采集模块爬取新浪网新闻及评论数据，数据预处理模块清洗数据，然后将数据保存到数据库，用户通过浏览器访问 web 服务器，web 服务器调用情感分析模块，情感分析模块从数据库读取待分析内容并分析，然后将分析结果返回，以网页的形式呈现给用户。

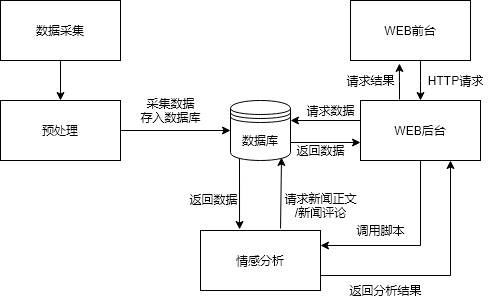


图4-4 系统运行机制图

### 4.6.2 运行控制

本文的系统运行控制是按照事务所属的模块来划分的，包括登录、注册、新闻显示、新闻或评论分析结果展示等，详情如表4-6所示。

表 4-6 运行控制表

|  |  |
| --- | --- |
| **事务名** | **事件流** |
| **登录模块** | |
| 登录 | 1. 用户登录：通过输入用户名，密码，连接数据库查询是否存在此用户。若有，则记住登录用户的类型并转到前台页面，若无，则提示出错信息 2. 管理员登录：通过输入管理员名，密码，验证码，连接数据库查询此登录名是否为管理员。若是，则转到后台管理页面 |
| **注册模块** | |
| 注册 | 游客按要求填写注册信息，系统判断输入的用户名是否已被注册，若未被注册则判断用户的注册信息是否符合要求，如果符合则注册成功，否则注册失败 |
| **普通游客模块** | |
| 新闻显示 | 用户进入系统以后，系统自动从数据库中读取信息，并且在首页新闻显示部分按新闻类别显示新闻标题信息 |
| **用户模块** | |
| 新闻分析选择及结果展示 | 用户可选择新闻分析其正文并查看分析结果 |
| 评论分析及结果展示 | 用户可选择新闻分析其评论并查看分析结果 |
| 用户自助服务管理 | 1、用户对个人信息进行修改  2、用户对个人登录密码进行修改 |
| **管理员模块** | |
| 1. 用户管理 | 分配给操作系统的用户不同的角色，包括管理员及用户，不同的角色可进行不同的操作 |
| 1. 新闻管理 | 管理员修改、添加、删除新闻信息 |

## 4.7 出错处理设计

### 4.7.1 出错输出信息

表 4-7 出错信息表

| **类别** | **发生情况** | **系统输出信息** | **处理方法** |
| --- | --- | --- | --- |
| 用户注册 | 用户输入的用户名已经存在 | 页面提示“该用户名已经存在！” | 跳转到注册页面，用户更改用户名 |
| 用户未将必须的注册相关内容填写完整 | 页面提示“……不能为空，请重新输入！” | 跳转到注册页面，需要用户将个人注册信息填写完整 |
| 两次输入的密码不相同 | 页面提示“两次输入的密码不一致，请重新输入！” | 跳转到注册页面，用户重新输入一致的登录密码 |
| 用户登录 | 验证码错误 | 页面提示“验证码错误，请重新输入！” | 跳转到登录页面 |
| 用户输入的用户名或密码不正确 | 页面提示“用户名或密码不正确，请重新输入！” | 跳转到登录页面，用户重新输入正确的用户名和密码 |
| 用户信息修改 | 两次输入的密码不相同 | 页面提示“两次输入的密码不一致，请重新输入！” | 跳转到用户修改密码页面 |
| 用户没有将必须需要的注册信息填写完整 | 页面提示“……不能为空，请重新输入！” | 停留在修改页面，需要用户将个人信息填写完整 |
| 用户未登录 | 未登录用户进分析结果展示页面 | 页面提示“您还未登录，不能查看分析结果！” | 跳转到首页 |
| 用户或管理员密码修改 | 提供错误的当前密码 | 页面提示“当前密码不匹配，请重输入！” | 跳转到用户或管理员密码修改页面 |
| 两次输入的密码不相同 | 页面提示“两次输入的密码不一致，请重新输入！” | 跳转到用户或管理员密码修改页面 |
| 未提供新设定的当前密码 | 页面提示“新密码不能为空！” | 跳转到用户/管理员修改密码页面 |
| 管理员登录 | 验证码错误 | 页面提示“验证码错误，请重新输入！” | 跳转到登录页面 |
| 管理员输入的账号或密码不正确 | 页面提示“用户名或密码不正确，请重新输入！” | 跳转到登录页面，重新输入正确的管理员账号和密码 |
| 管理员添加新闻信息 | 没有填写所有必填的新闻相关内容 | 页面提示“待添加新闻……内容不能为空！” | 跳转到添加新闻页面 |
| 管理员删除新闻信息 | 没有填写所有必须需要的新闻信息 | 页面提示“待删除新闻……不能为空” | 跳转到删除新闻页面 |
| 管理员添加用户基本信息 | 没有填写所有必须需要的用户信息 | 页面提示“待添加用户……不能为空” | 跳转到添加用户页面 |
| 管理员删除用户基本信息 | 没有填写所有必须需要的用户信息 | 页面提示“待删除用户……不能为空” | 跳转到删除用户页面 |
| 用户查看分析结果 | 没有选择待分析文本 | 页面提示“待分析文本不能为空” |  |

### 4.7.2 出错处理对策

数据采集模块采用的是网络爬虫技术采集新浪网数据，在请求网页数据时会出现请求失败，数据获取不全等情况，管理员应在数据采集后查看错误记录文件，对出问题新闻手动重新爬取信息。

## 4.8 维护设计

必须按照软件产品设计规范的步骤进行开发，并充分考虑系统的可维护性，在各个阶段编写详细的文档，以方便系统的移植或二次开发。定期维护数据库，根据信息更新修改数据库信息，保持数据库最近最新。通常使用改正性维护、适应性维护、完善性维护、预防性维护这4种维护活动来完成系统维护。

# 5. 详细设计

## 5.1 软件结构

本系统大致分为数据采集与预处理、情感分析、用户交互三大模块。

数据采集与预处理模块包括新闻正文采集、新闻评论采集、数据清洗、数据存储四个功能，新闻正文采集、新闻评论采集的功能是通过网络爬虫对新浪网新闻及其评论网页内容进行抓取，数据清洗的功能是提取出抓取内容里的新闻id、新闻标题、新闻内容、新闻评论文本、评论时间、评论点赞数等信息，数据存储的功能是将清洗过的新闻数据信息保存到数据库，评论信息以JSON文件的形式保存。情感分析模块包括新闻正文情感分析与新闻评论情感分析，是本系统的主要模块。用户交互模块包括新闻数据展示、数据分析结果展示、用户管理界面、新闻数据管理界面等功能。

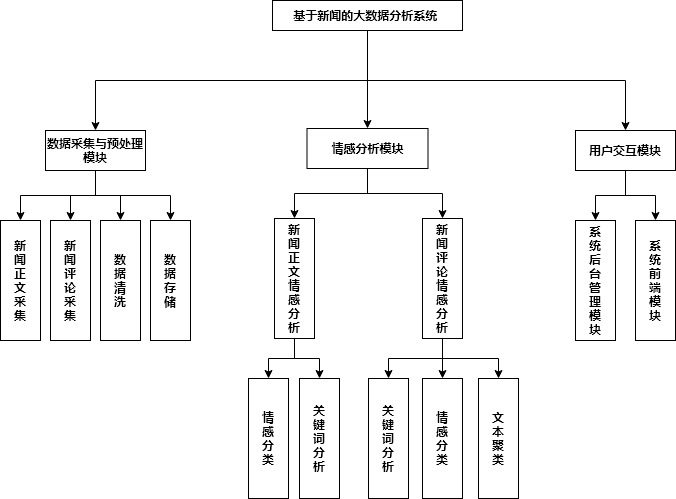


图5-1 软件结构图

## 5.2 数据采集与预处理模块

该模块的功能是从网络中采集新闻信息及其评论信息并进行数据清洗与保存，其数据来源是新浪网新闻中心全部类型的新闻及其评论。

### 5.2.1 新闻爬取模块

(1) 程序描述

该模块的功能是采集新闻信息，包括新闻的标题、正文内容、关键字等。

新闻正文获取：通过requests模块获取网页信息，然后使用Beautiful Soup库的select()方法定位并获取新闻正文。

新闻其他信息获取：通过Request模块获取网页js文件脚本信息然后将获取的json数据转换为字典，提取出所需信息。

(2) 输入项

表5-1 输入项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **数据来源** | **数据处理** |
| 新闻类别 | news\_kind | String | 新浪网 | 添加到url |
| 新闻页面URL | news\_url | String | 新浪网 | 请求网页信息 |
| 新闻其他信息URL | parse\_url | String | 新浪网 | 请求网页信息 |

(3) 输出项

表5-2 输出项表

| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| --- | --- | --- | --- | --- |
| 新闻id | n\_id | String | 写入数据库 | 数据库 |
| 新闻标题 | title | String | 写入数据库 | 数据库 |
| 新闻内容 | article | String | 写入数据库 | 数据库 |
| 新闻关键字 | keyword | String | 写入数据库 | 数据库 |
| 发布时间 | ctime | datetime | 写入数据库 | 数据库 |
| 发布媒体 | media\_name | String | 写入数据库 | 数据库 |
| 新闻链接 | url | String | 写入数据库 | 数据库 |
| 新闻类型 | kind | int | 写入数据库 | 数据库 |
| 评论文件地址 | comment | String | 写入数据库 | 数据库 |

(4) 逻辑流程

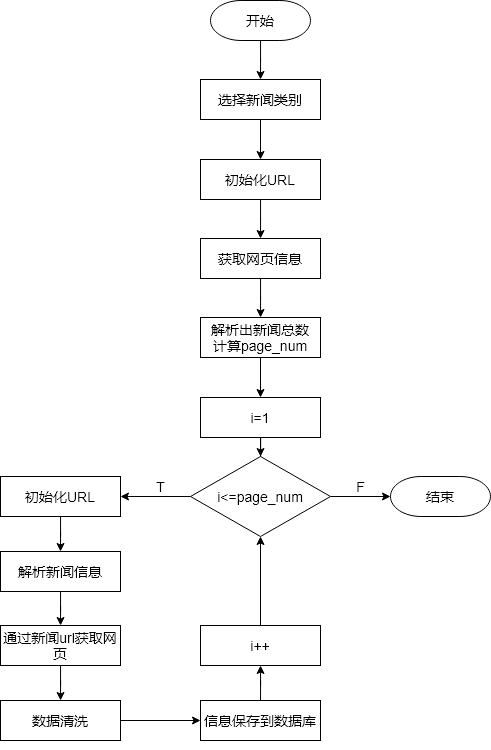


图5-2 新闻爬取流程图

(5) 测试计划

点击新闻分类选择模块按钮，确保新闻类别列表能正确显示，点击某个分类，确保脚本能正确爬取新闻信息及新闻内容。

### 5.2.2 新闻评论爬取模块

(1) 程序描述

通过Request模块获取新闻详情网页的评论js文件脚本信息，将获取的json数据转换为字典，解析出该新闻的总评论数，若小于200条则一次性请求数据，若大于200则按200条/页再次请求数据，然后将获取的json数据转换为字典，提取出所需信息。当数据量大于200时，由于要多次请求网页数据，为避免频繁访问，设置每次请求后将休息5秒。

(2) 输入项

表5-3 输入项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **数据来源** | **数据处理** |
| 新闻id | news\_id | String | 新浪网 | 添加到url |
| 新闻频道 | channel | String | 新浪网 | 添加到url |
| 解析地址 | parse\_url | String | 新浪网 | 请求网页信息 |

(3) 输出项

表5-4 输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 评论id | comment\_id | String | 写入文件 | json |
| 评论内容 | content | String | 写入文件 | json |
| 评论设计 | time | String | 写入文件 | json |
| 分布位置 | area | String | 写入文件 | json |
| 点赞数 | agree | String | 写入文件 | json |

(4) 逻辑流程

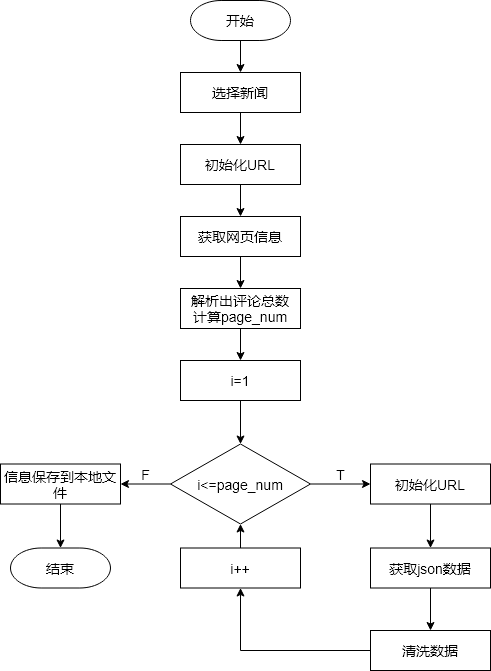


图5-3 新闻评论爬取流程图

(5) 测试计划

指定一新闻URL，确保脚本能正确爬取新闻评论信息及新闻评论内容。

## 5.3 情感分析模块

该模块主要是研究对新闻及其评论文本的情感分析，其包括情感分类与关键词分析。

情感分类方面，本系统采用了基于依存句法分析的方法与基于深度学习的方法。

关键词分析方面，本系统采用的是北京大学开源分词工具pkuseg、TextRank算法与AP聚类算法。pkuseg主要用在分词阶段，pkuseg与常用的jieba分词对比，在中文新闻文本分词中有着更高的准确率；TextRank算法主要用于文本关键词提取，它是一个基于图排序的、用于文本处理的算法，考虑了相邻词的语义关系，可以更好的表达文本，更准确的提取文本中关键词[7]； AP聚类算法主要用于某一个新闻或者某一类新闻的评论文本的聚类，AP聚类算法是一种基于数据点间的"信息传递"的聚类算法[8]， 每个数据点都被视为图的一个节点，并且图的有向边被视为节点之间消息的传递，目的是发现到所有数据点的总相似度最大的最优类代表点集合。

### 5.3.1 新闻情感分类模块

(1) 程序描述

新闻文本的情感分类是通过依存句法分析的方法计算文本情感得分来实现的，得分为0代表中性，得分在[-1,0)之间为负面，得分在（0，1]之间为正面，其处理流程如下：

用户登录系统，进行新闻选择界面，选择想要分析的新闻，新闻分析脚本接收待分析新闻的id，脚本从数据库读取新闻信息进行情感分类、关键词分析，将分析结果返回到前台展示。

(2) 输入项

表5-5输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **数据处理** |
| 新闻id | n\_id | String | 传参 | 网页爬取 | 查询数据库 |
| 新闻内容 | content | String | 传参 | 数据库查询 | 分句分词 |

(3) 输出项

表5-6 输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 情感得分 | score\_result | int | 写入txt文件 | 文本 |
| 词云 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 情感曲线 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 关键词排名 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 情感比例图 | n\_id(.HTML) | 文件 | HTML文件 | 网页 |

(4) 情感词典设计

如图5-4所示，本文的情感词典由知网情感词典、清华大学李军中文褒贬义词典、BosonNLP情感词典、程度副词词典四部分组成，词典内容包含两个部分——词语，权重。

如图5-5所示，本文所使用的情感词典中积极词占比约为71%，消极词占比约为29%。

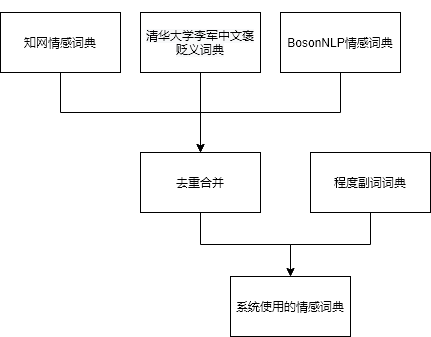
****

图5-4 情感词典组成图

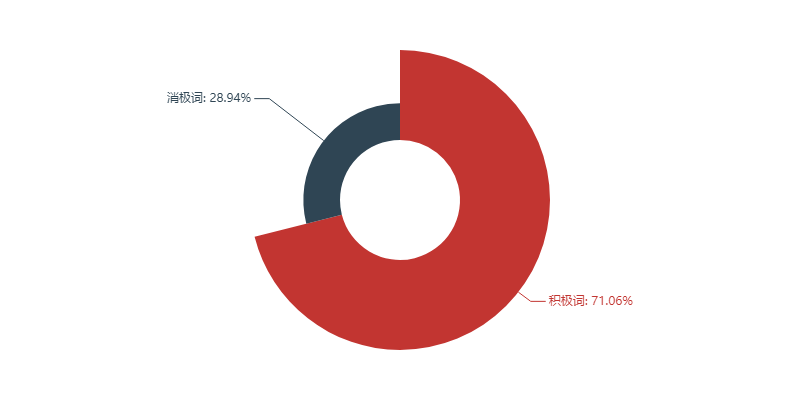
****

图5-5 情感词比例图

(5) 算法设计

依存句法分析经过剖析句子的成分之间的依存关系来展示其句法结构[9]，与将句子分为主语、形容词和语法元素的传统句法分析相比，依存句法认为句子的不同组成部分（词语）之间存在依存关系，其首先专注于句子的动词部分，然后分析动词与句子中其他成分之间的关系。依存关系是指句子中词与词之间支配与被支配的关系，其既可以是词与词之间的句法关系，也可以是词与词之间的语义关系。

基于依存句法分析的情感分类方法基本流程如下：

1）通过pkuseg工具对文本进行分句分词操作，并借助情感词典完成情感句过滤；

2）通过哈工大 LTP 语言云平台对过滤后的情感句进行词性标注、依存句法分析，最终从待处理文本提取出[词，词性，依存弧的父节点词，父节点词词性，关系]列表；

3）借助情感词典与程度副词词典计算情感句情感得分，然后通过加权整合，作为文本的情感分数。

(6) 逻辑流程

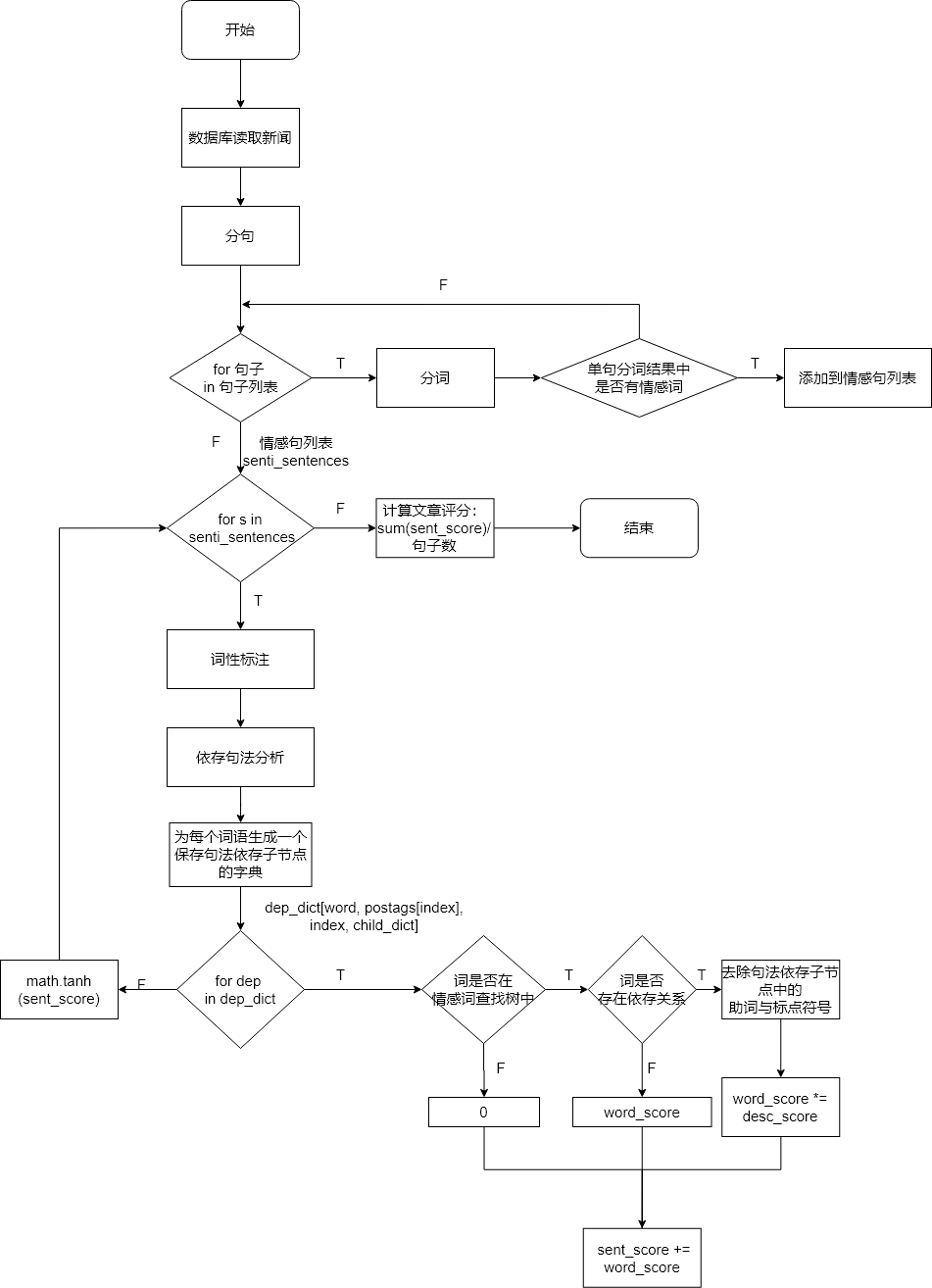


图5-6 新闻情感分类流程图

(7) 测试计划

挑选5到10篇情感趋向明显的新闻报道进行情感分类，判断结果与人工分类结果是否一致，计算算法准确率。

### 5.3.2 评论情感分类模块

(1) 程序描述

本文的评论情感分类使用的是基于深度学习模型的情感分析方法，是基于LSTM神经网络搭建的新闻评论文本情感分类的深度学习模型。

处理流程：用户登录系统，进行新闻选择界面，选择想要分析的新闻，评论分析脚本接收待分析的新闻id，脚本从数据库读取新闻对应评论信息文件存储位置，读取评论信息文件信息进行情感分类、关键词分析，将分析结果返回到前台展示。

(2) 输入项

表5-7 输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **数据处理** |
| 新闻id | n\_id | String | 传参 | 网页爬取 | 查询数据库 |
| 评论文件位置 | comment\_file | String | 传参 | 数据库查询 | 文件读取 |

(3) 输出项

表5-8 输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 情感趋向 | predict\_result | int | 写入txt文件 | 文本 |
| 词云 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 情感曲线 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |
| 关键词排名 | n\_id(.PNG) | 文件 | PNG文件 | 图片 |

(4) 训练词向量

本文借助Word2Vec方法训练词向量，以网络抓取的新闻评论作为训练语料，使用jieba分词对语料中的每个句子分词并组成词列表，然后使用gensim工具包的word2vec开始训练词向量。鉴于训练语料较大，故本文采用的是更为成熟的jieba分词工具进行分词，可大大节省时间。本文训练忽略词频低于10的词的统计，得到的是100维的词向量。训练步骤如下：

1.建立一个空的模型对象；

2.遍历一次语料库建立词典；

3.第二次遍历语料库建立word2vec模型。

(5) 训练分类模型

本文使用的分类模型是基于LSTM的，基于神经网络的分类模型需要给模型提供已人工分类的标注语料做训练，本文采用的训练语料是网络中开源的评论标注语料，包括谭松波老师整理的酒店评论语料集以及其他开源的评论语料。其中正面评论共8000条，负面评论共8700条，中性评论共4300条。本文的语料使用划分为70%作为模型的训练数据集，30%作为模型测试数据集。

LSTM是一种特殊的RNN网络，该网络设计出来是为了解决长依赖问题。如图5-7为LSMT的一个单元。如图5-8所示，LSTM的细胞状态由三个门进行控制——遗忘门、输入门和输出门[10]；

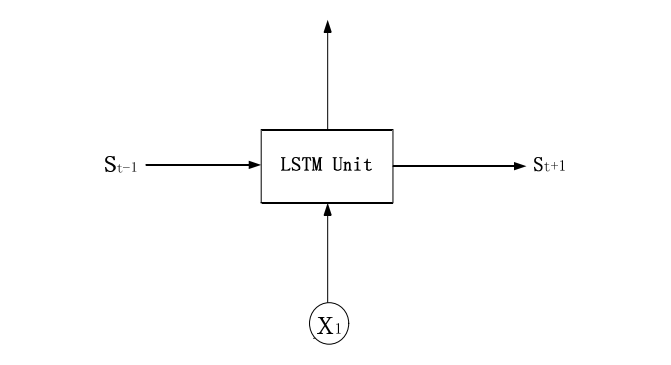


图5-7 LSTM网络单元

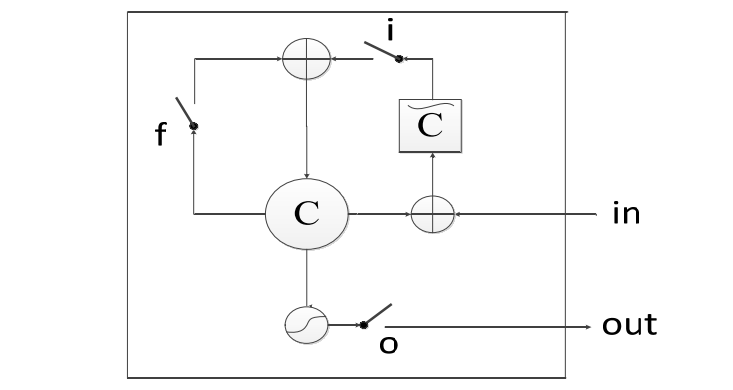


图5-8 LSTM网络结构

遗忘门决定cell需要丢弃哪些信息，其计算公式为：



输入门决定需要更新哪些信息，更新的规则就是通过遗忘门选择遗忘旧细胞的一部分信息，通过输入门选择添加候选细胞信息的一部分得到新的细胞信息，其计算公式为:



输出门判断输出细胞的哪些状态特征，其计算公式为



表示t-1 时刻LSTM 单元的输出，W 表示系数矩阵，b 为偏置向量，σ为标准 sigmoid 函数[11]。

本文使用的LSTM模型结构如图5-8所示，其输入的向量维度为100维，输出的为50维。本文使用的分类模型损失函数选取的是categorical\_crossentropy，优化方法选取的是adam，其创建流程如下：

1)将所有词的向量集传入 embedding层，由于训练集数量较大，使用嵌入式词向量来表示单词更有效率。

2)通过embedding 层, 将新的词表示传入 LSTM cells。

3)LSTM cells连接一个softmax output layer 。

模型训练中，本文训练轮数选取的是5，批次大小选取的是32，经过5轮训练之后，语料数据在验证集上的准确率达到了88%。

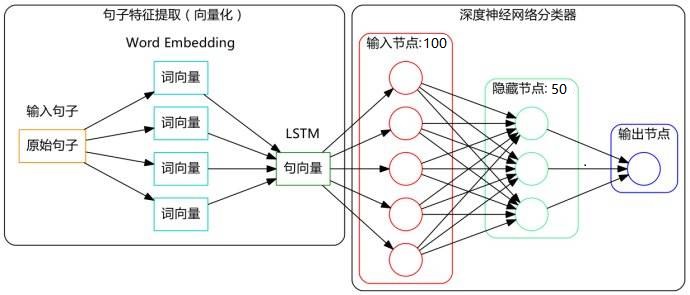
****

图5-8 LSTM模型结构图

(6) 逻辑流程

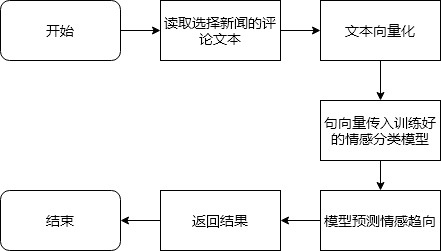
****

图5-9 评论情感分类流程图

(7) 测试计划

选择新闻，点击评论分析按钮，确保能跳转到数据分析结果页面，并能完整展示分析结果。

### 5.3.3 文本聚类模块

本文通过AP聚类算法实现某一个新闻或者某一类新闻的评论文本聚类，其基本思想是将所有的数据点都视为潜在的聚类中心，然后将数据点两两相连以形成网络(相似性矩阵)，再通过网络中各条边的消息的吸引度与归属度来计算各样本的聚类中心[12]。

1. 程序描述

本文通过词向量和词向量对应的tf-idf 值构建文本向量，通过计算两个文本向量的余弦相似性得到文本之间的相似性，最后通过AP聚类算法得到观点数量，并从观点中筛选典型评论[13]。

AP聚类算法基本流程如图5-10所示：

1）读取评论文本，并根据已训练好的word2vec模型生成文本向量；

2）计算相似度矩阵；

3）调用sklearn包中的AP算法在相似度矩阵上执行聚类并返回聚类标签。

典型评论筛选基本流程如图5-11所示：

1. 从已近分好类的评论中读取评论文本；
2. 分别提取每一个分类下的文本关键词并向量化；
3. 将对应分类中的每句评论分别分词并向量化；

4）计算每句评论与其对应的类标签关键词的相似度；

5）挑选出每个分类下与其关键词最大相似的评论作为典型评论。

(2) 输入项

表5-9 评论文本聚类输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **处理数据** |
| 评论文本 | comment | String | 文件读取 | 数据爬取 | 添加到实体 |
| word2vec模型 | word\_model | 模型 | 模型加载 | 模型训练 | 模型加载 |

(3) 输出项

表5-10 评论文本聚类输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 评论标签 | label | int | 写入文件 | 文件 |
| 典型评论 | typical\_comments | String | 写入文件 | 文件 |

(4) 逻辑流程

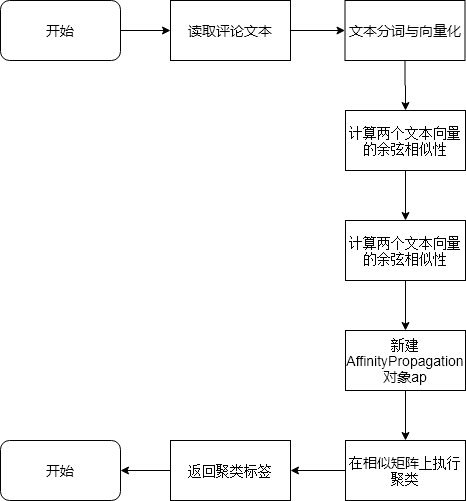
****

图5-10 AP聚类流程图

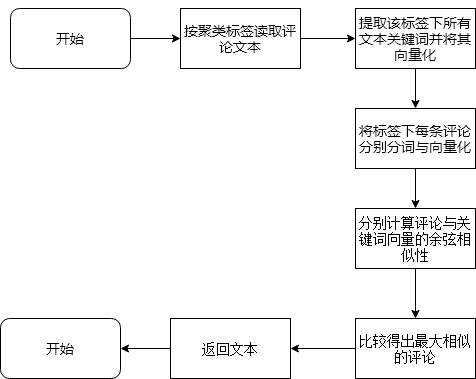
****

图5-11 典型评论筛选流程图

(5) 测试计划

人工分类好一批评论文本作为测试数据，将测试数据传入到程序，比对程序的分类结果，计算准确度。

## 5.4 用户交互模块

### 5.4.1 后台用户管理模块

(1) 程序描述

该模块实现系统管理员对用户信息的管理的设计，包括添加用户信息、修改用户信息和删除用户信息三个功能点。管理员登录系统，进入后台管理首页，点击菜单栏中的用户管理，弹出下拉选项，点击下拉选项中的添加用户信息，进入用户信息添加页面，输入用户信息，进行用户信息的添加，添加成功后页面跳转到用户列表页面；点击下拉选项中的用户信息修改，跳转到用户列表页面，点击列表中的修改按钮，跳转到修改页面，对用户信息进行修改。

(2) 输入项

表5-11 用户信息添加输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **处理数据** |
| 用户id | id | int | 键盘 | 用户 | 添加到实体 |
| 用户名 | username | varchar | 键盘 | 用户 | 添加到实体 |
| 登陆密码 | password | varchar | 键盘 | 用户 | 添加到实体 |
| 手机号 | phone | varchar | 键盘 | 用户 | 添加到实体 |
| 添加 | add\_bt | -- | 点击按钮 | -- | 添加到数据库 |
| 取消 | Cancel\_bt | -- | 点击按钮 | -- | 清空数据 |

表5-12 用户信息修改输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **处理数据** |
| 用户id | id | int | 键盘 | 用户 | 添加到实体 |
| 用户名 | username | varchar | 键盘 | 用户 | 添加到实体 |
| 登陆密码 | password | varchar | 键盘 | 用户 | 添加到实体 |
| 手机号 | phone | varchar | 键盘 | 用户 | 添加到实体 |
| 权限 | level | int | 键盘 | 用户 | 添加到实体 |
| 确定 | update\_bt | -- | 点击按钮 | -- | 替换库中数据 |
| 取消 | Cancel\_bt | -- | 点击按钮 | -- | 清空数据 |

(3) 输出项

表5-13 用户信息修改输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 用户id | id | int | 写入数据库 | 数据库 |
| 用户名 | username | varchar | 写入数据库 | 数据库 |
| 登陆密码 | password | varchar | 写入数据库 | 数据库 |
| 手机号 | phone | varchar | 写入数据库 | 数据库 |
| 权限 | level | int | 写入数据库 | 数据库 |

(4) 逻辑流程

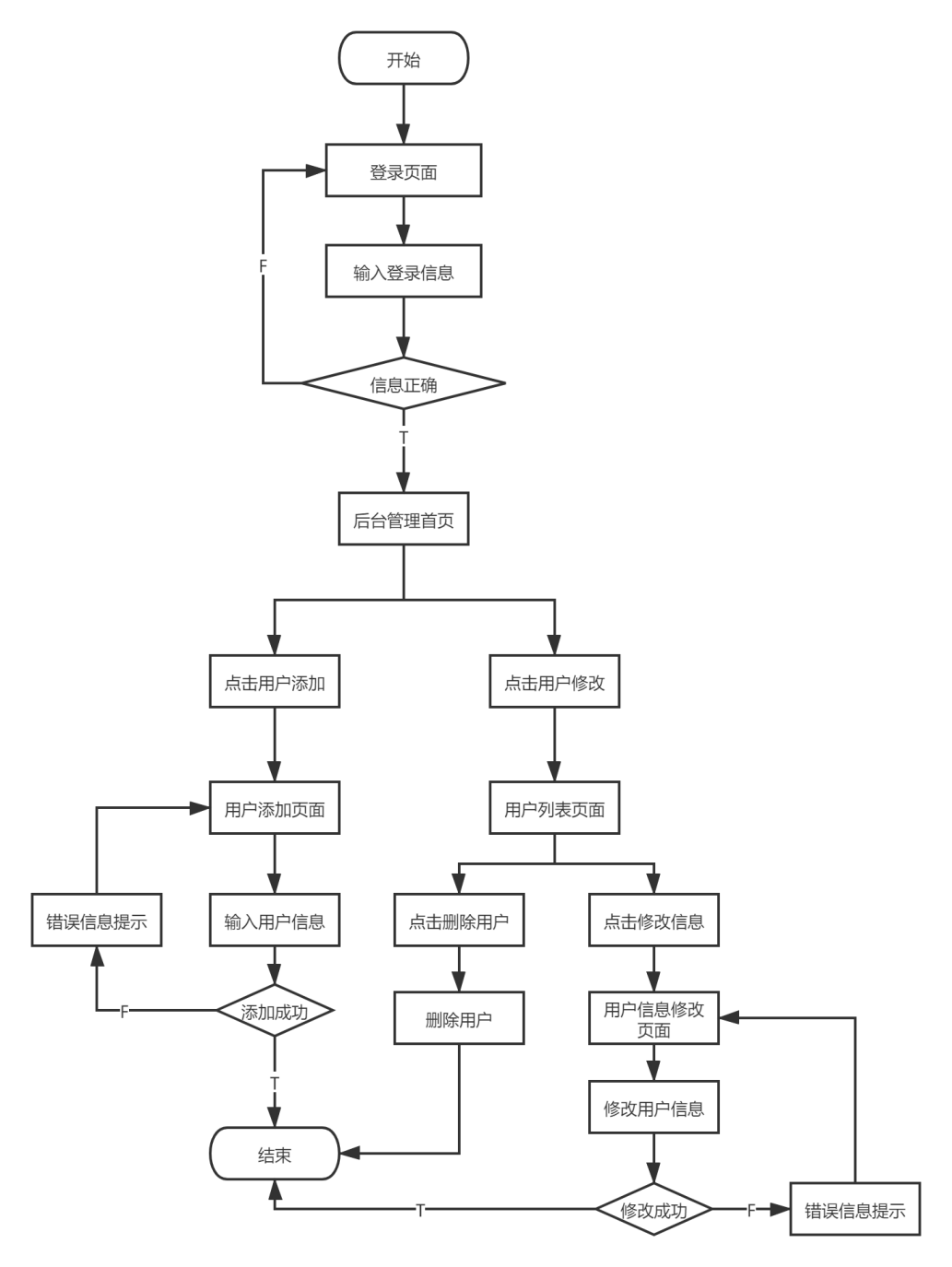


图5-12 后台用户管理模块流程图

### 5.4.2 后台新闻管理模块

(1) 程序描述

该程序完成管理员对新闻的管理，包括添加、修改和删除新闻信息。管理员登录系统，进入后台管理首页，点击新闻管理，弹出下拉列表。点击新闻添加，进入新闻添加页面，输入新闻URL，进行新闻的添加，添加成功后系统跳转到新闻列表页面，或点击新闻列表进入列表页面，在新闻列表中可点击删除进行新闻的删除，也可点击编辑，进入新闻修改页面，进行新闻的修改。

(2) 输入项

表5-14 新闻添加/修改输入项表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输入方式** | **数据来源** | **数据处理** |
| 新闻标题 | title | String | 网页爬取 | 新浪网 | 添加到实体 |
| 新闻内容 | article | String | 网页爬取 | 新浪网 | 添加到实体 |
| 新闻关键字 | keyword | String | 网页爬取 | 新浪网 | 添加到实体 |
| 发布时间 | ctime | String | 网页爬取 | 新浪网 | 添加到实体 |
| 新闻类型 | kind | String | 网页爬取 | 新浪网 | 添加到实体 |
| 新闻评论  文件位置 | comment | String | 网页爬取 | 新浪网 | 添加到实体 |
| 提交 | BtnAdd | -- | 点击按钮 | -- | 添加到数据库 |

(3) 输出项

表5-15 新闻添加/修改输出项表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **名称** | **标志** | **数据类型** | **输出方式** | **输出媒介** |
| 新闻ID | NewsId | int | 写入数据库 | 数据库 |
| 信息标题 | NewsTitle | varchar | 写入数据库 | 数据库 |
| 发布人 | ReleaseMan | varchar | 写入数据库 | 数据库 |
| 发布时间 | ReleaseDate | varchar | 写入数据库 | 数据库 |

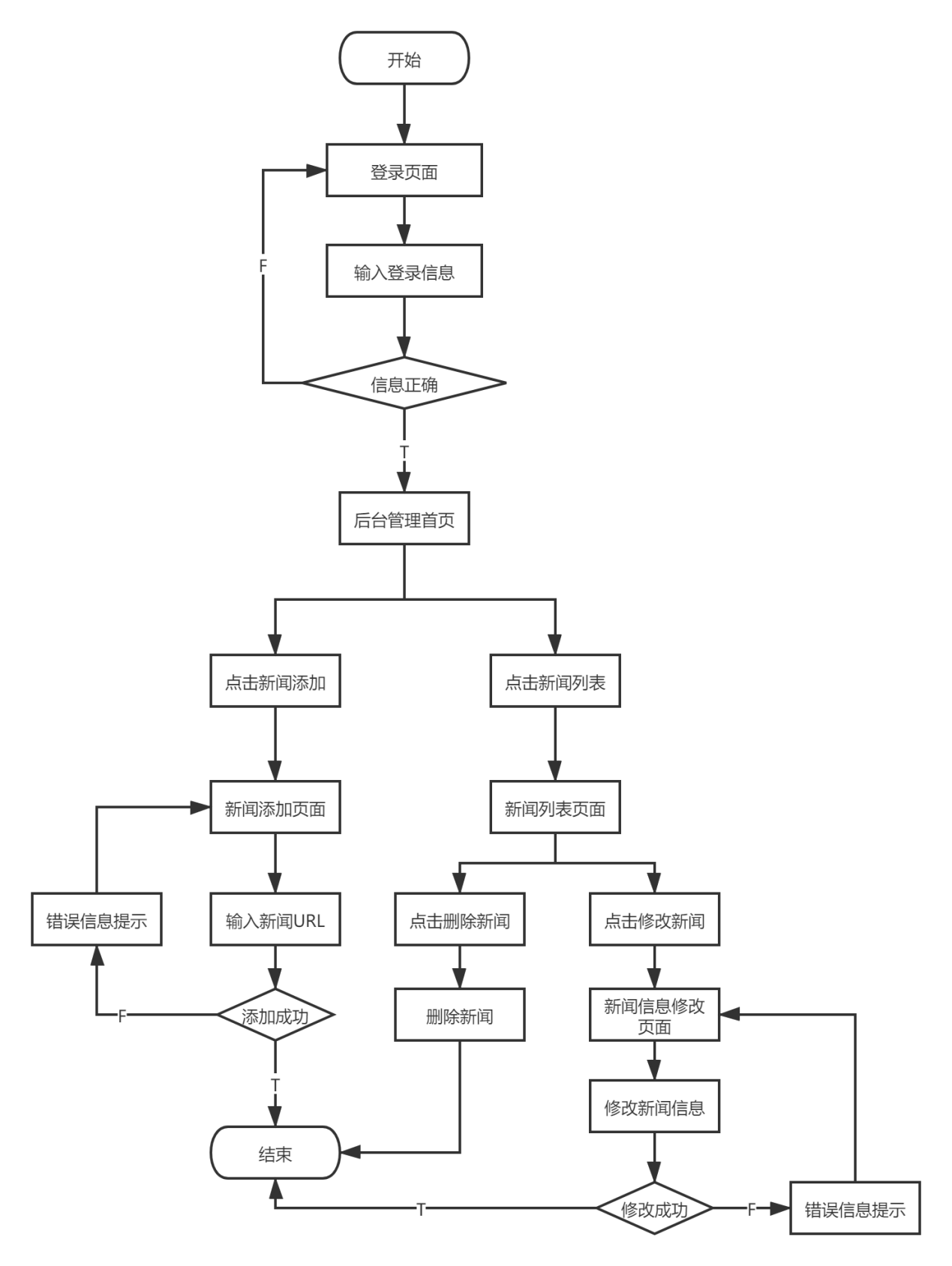
(4) 逻辑流程

图5-13 新闻管理流程图

# 6. 运行测试

## 6.1 测试计划

本文现已完成系统的设计与实现，为了确定系统的可行性和可靠性，需要进一步对其性能与功能进行测试然后对结果进行分析，本文主要就系统的数据爬取与预处理、情感分析两大模块进行功能测试与准确性测试。测试计划如下：

1. 为验证数据爬取与预处理的准确性，本文选取了多篇新闻对其进行数据抓取与预处理，判断是否能准确的抓取网页与提取出所需内容。
2. 为验证情感分类结果的准确性，本文选取五篇新闻及其评论，先对其情感趋向进行人工标注，然后与系统的分析结果一一比对，计算准确度。

## 6.2 测试结果分析

### 6.2.1 数据爬取与预处理测试

数据爬取与预处理功能测试主要是测试系统能否准确的抓取所需数据。本次测试选取的数据是2020年5月17日当天的所有国内新闻。测试结果如图6-1、6-2所示，新闻信息抓取与预处理无任何报错，评论信息共抓取1000条，抓取出现两次报错，两次都是网页请求失败，原因为请求过于频繁。

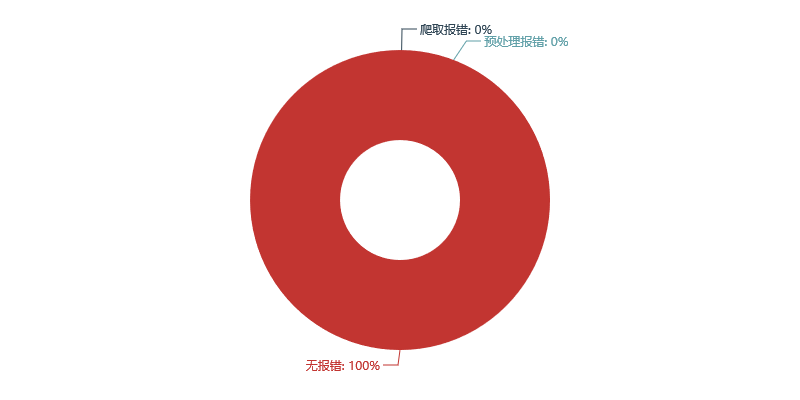


图6-1新闻爬取与预处理测试结果

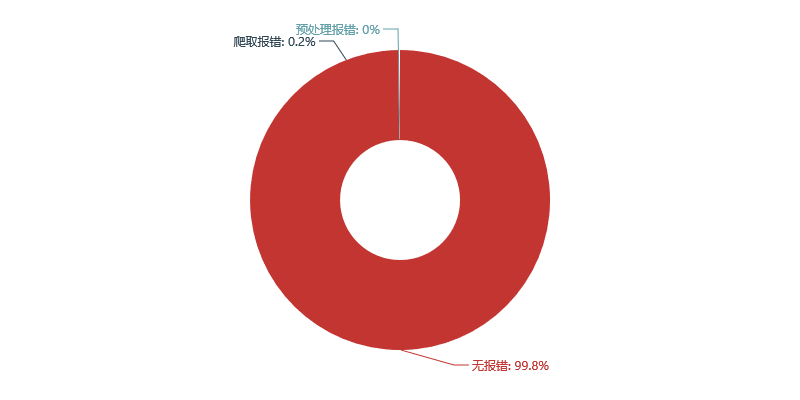


图6-2新闻评论爬取与预处理测试结果

### 6.2.2 情感分类功能测试

情感分类功能测试主要是测试系统情感分析的准确度。本次测试选用的数据是2020\05\17当天所有新闻与新浪网“嫦娥四号和“玉兔二号”结束休眠 已奋战500天”、“河底“抹水泥”，大理苍山五溪生态治理惹争议”、“武汉十答全民核酸检测：消除社会恐惧和偏见，个人免费”、“3月份以来，中国已出口口罩509亿只”、“中共中央政治局常务委员会召开会议 习近平主持”五篇新闻的评论。测试结果如图6-3、6-4所示，新闻文本准确度在90%以上，评论文本准确度在86%以上，从结果图中还可以分析出，评论分析中的中性文本判断失误的频率明显减低，说明针对中性评论的优化，起到了预期效果。

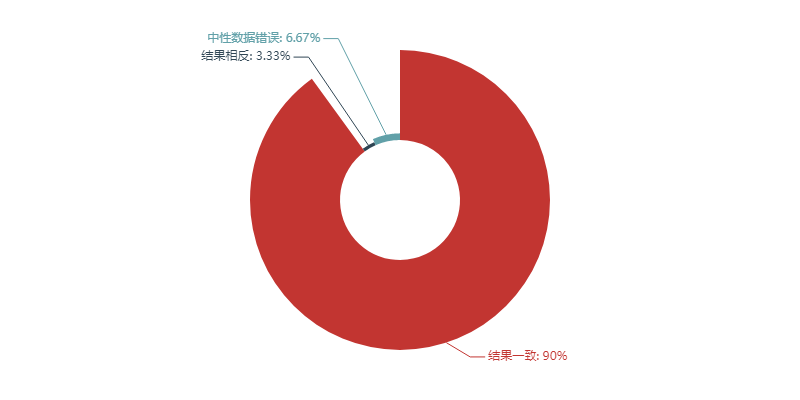


图6-3新闻文本情感分类测试结果

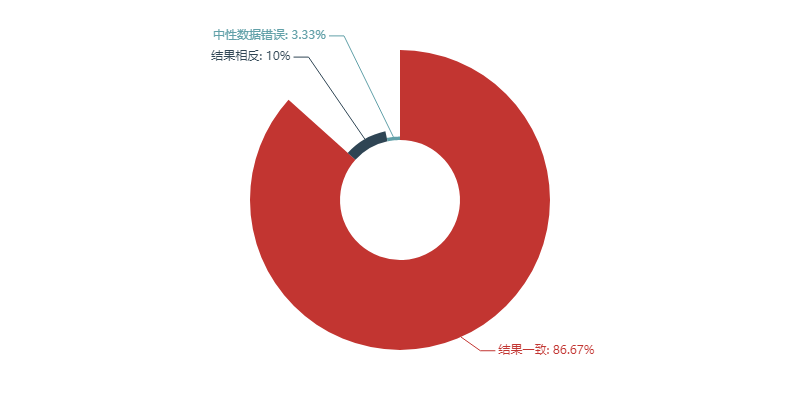


图6-4新闻评论情感分类测试结果

# 7. 运行结果

## 7.1 数据爬取与预处理运行结果

数据爬取过程如图7-1所示。管理员选择爬取新闻种类，开始爬取，按每页50条新闻爬取信息，默认爬取当前种类的200条新闻及其评论。数据预处理结果如图7-2所示，清洗爬取的网页信息，获取所需要的新闻及评论相关数据，然后插入数据库。

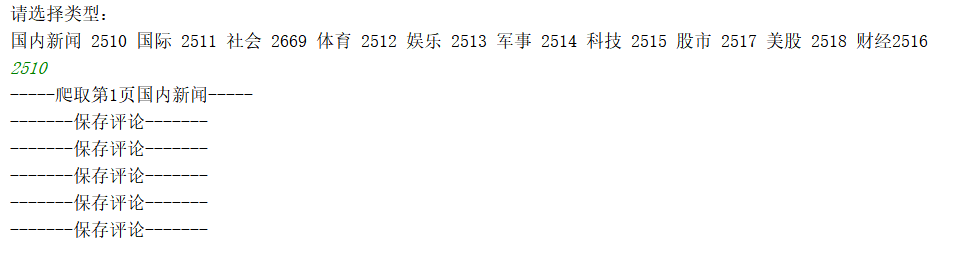


图7-1 数据爬取过程图



图7-2 预处理结果图

## 7.2 情感分类运行结果

用户传入新闻id，系统自动分析文本，得出文本每句情感分数，计算文本情感分数，并判断文本情感趋向。结果如图7-3、7-4所示，该新闻文本在8-14句、16-21句之间情感波动较小，情感得分在0.5~1之间的句子占33.33%，情感得分在-1~-0.5之间的句子占66.67%，文章总体偏负面。

用户选择新闻评论文本，系统自动分析文本，得出每句评论情感趋向，结果如图7-5所示，正面评论占39.47%，中性评论占2.63%，负面评论占57.9%

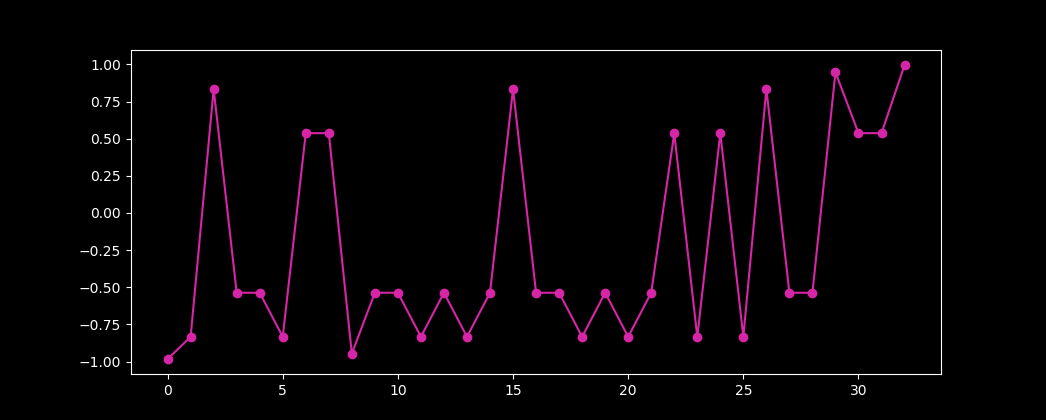


图7-3 新闻文本情感曲线图

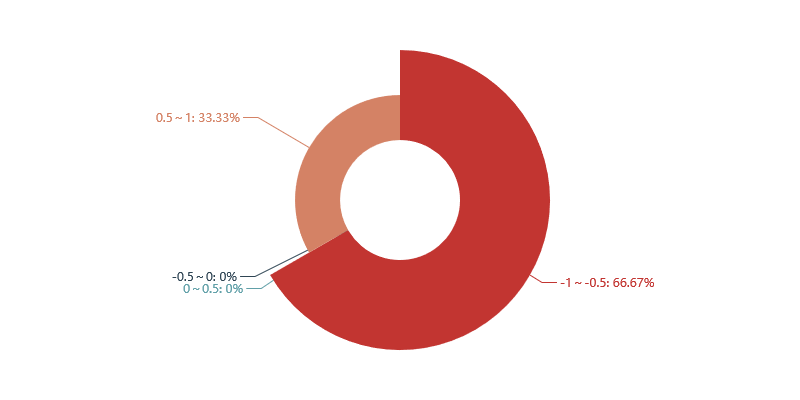


图7-4 新闻文本情感比例图

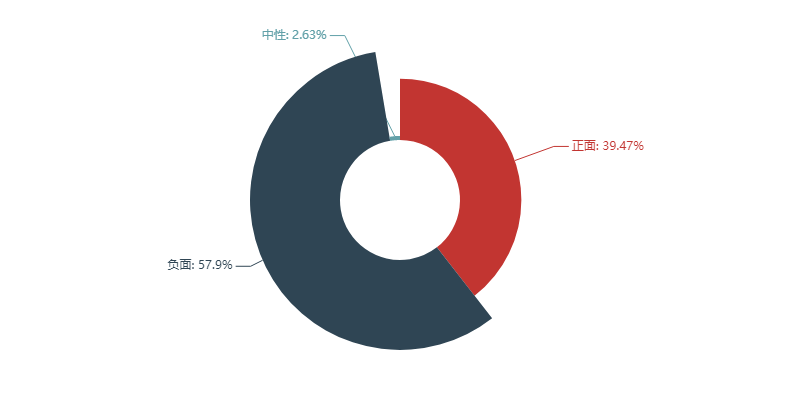


图7-5 新闻评论情感比例图

## 7.3 文本聚类运行结果

用户选择新闻，系统自动将新闻评论文本分类，并得出每个分类里面的典型评论。评论文本聚类结果如图7-6所示，关键评论提取如图7-7所示。

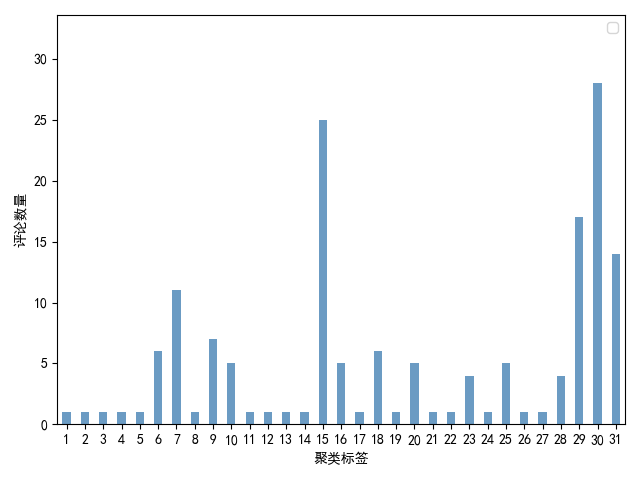


图7-6 评论聚类结果图

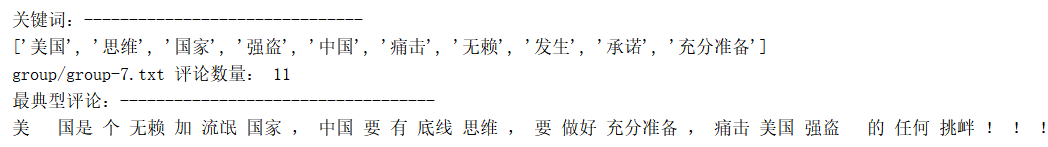


图7-7 关键评论提取结果图

总结

本文主要是对网络新闻文本包括新闻正文及其评论的情感分析的设计与实现。相较于其他研究主要有以下几点不同与优势：

1. 由于本文研究对象为新闻正文与评论文本，两者篇幅差距较大，故在情感分析方法上面，本文选择了两种方法——依存句法分析与基于LSTM的深度学习模型，从分析速率与准确度考虑，最终选择了在新闻正文的情感分析研究上使用依存句法分析，新闻评论的情感分析研究上使用深度学习模型。
2. 在深度学习模型的研究中，与传统的二分类模型差别的是，本文采用的是三分类模型，三分类除了涉及到二分类中的正面和负面情感外，还有一种中性情感，从原始数据中，提取具有语义转换的句子，从而获得3个具有不同情感的数据。
3. 在新闻正文中文分词中，本文选取了更具领域性的北京大学开源分词工具pkuseg，pkuseg提供了在MSRA（新闻语料）上训练得到的模型，在新闻文本分词中更为准确。但考虑到其分词耗时长在大数量的评论分词中本文选择的是较稳定、快速的jieba分词工具。
4. 本系统的模块符合高内聚、低耦合的原则。

基于以上分析和算法研究，结合python网络爬虫技术和Java web技术，设计并完成了一个基于新闻的大数据分析系统，并对系统进行功能测试，测试结果表明，系统符合预期的要求。

但是本系统还是存在不足，比如本系统的评论情感分类准确度不高，主要原因是，网络新闻的标记语料的缺乏；中性评价出现的概率不高，主要原因是中性数据集的数量要比其他两个数据集少一半多，并且通过简单规则“然而”，“但”提取出来的中性数据集质量也不是很高，所以才会出现偏差。

致谢

毕业设计是四年大学生活的结束标志，经过五个多月的努力，此次毕业设计终于完成了，这也标志着我美好而充实的大学生活即将结束。四年的大学生活中，老师、同学与身边的朋友们给予了我许多真诚的帮助与关怀。在此，我表示万分感谢。

这次毕业设计的完成，首先要感谢魏冬梅老师的精心指导。疫情居家期间，魏老师仍坚持每周的指导，在每周的指导时，魏老师督促与检测我们毕业设计的完成度，并对我们已完成部分提出意见与建议，让我及时了解到自己毕业设计中的不足与下一步目标。在本次毕业设计中，如果没有魏老师的悉心指导，我无法轻松地完成程序的编写与相关文档的准备。

除了感谢魏老师外，我还要感谢那些帮助与关怀我的同学、朋友们。我要感谢陈晓玮同学在相关文档的编写过程中的提醒与建议；还有我的朋友周伍，在神经网络学习方面，他给我提供了许多学习资料与指导，得益于他的帮助，我才能在几周之内完成神经网络的构建。感谢我的大学同学以及舍友们，我很荣幸能跟大家一同度过难忘而又可贵的四年大学生活。

参考文献

[1] 中国互联网络信息中心（CNNIC）.第45次中国互联网络发展现状统计报告[R].2020年4月28日.

[2] 陈敬,刘欢欢,李寿山,周国栋.基于主动学习的新闻读者情绪分类方法 [J]. 山西大学学报(自然科学版), 2017,40(155), 01 26-31.

[3] 赵妍妍.句子级情感信息抽取与极性识别的方法研究[D].哈尔滨工业大学, 2011.

[4] 唐立婷.关于网络舆情生态系统的构建与运行机理研究[J].现代经济信息,2019,15 411+413.

[5] 李雪.面向网络新闻的舆情检测与分析系统设计与实现[D].山东师范大学,2014.(4)

[6] 段传明.传统情感分类方法与基于深度学习的情感分类方法对比分析[J].软件导刊,2018,17(01):22-24.

[7] 韦修喜,黄华娟,周永权. 基于AP聚类的约简孪生支持向量机快速分类算法 [J]. 计算机工程与科学, 2019,41(298), 10 189-194.

[8] Frey B J, Duech D.Clustering by passing messages between data points[J].Science, 2007, 315 (5814):972-976.

[9] 陶杨明. 基于深度学习的细粒度文本情感分析研究[D].浙江工商大学,2020.

[10] 张原溥.基于LSTM长短期记忆网络的情感分析[J].电子世界,2018(17):16-17+20.

[11] 张英. 基于深度神经网络的微博短文本情感分析研究[D].中原工学院,2017.

[12] 韦修喜,黄华娟,周永权. 基于AP聚类的约简孪生支持向量机快速分类算法 [J]. 计算机工程与科学, 2019,41(298), 10 189-194.

[13] 季蕾娜. 基于用户权威度的中文微博话题检测研究[D].昆明理工大学, 2014.