

**本 科 生 毕 业 设 计（论文）**

**（2019届）**

**信息工程学院**

**题 目：在线商品用户评论分析系统设计与实现**

**学 号： 201505010323**

**姓 名： 陈怡**

**专业班级： 计算机科学与技术153班**

**指导教师： 黄雷君 职称： 讲师**

**2019年 5月 3日**



**本科生毕业设计（论文）诚信承诺书**

我谨在此承诺：本人所写的毕业设计（论文）《××××××××》均系本人独立完成，没有抄袭行为，凡涉及其他作者的观点和材料，均作了引用注释，如出现抄袭及侵犯他人知识产权的情况，后果由本人承担。

**承诺人（签名）：**

**年 月 日**

**在线商品用户评论分析系统设计与实现**

信息工程学院 计算机科学与技术153班 陈怡 指导教师：黄雷君

**摘要**:随着网络的普及，在线购物已经是我们生活中的常态，在网购时我们经常通过他人的商品评价来分析商品的优劣。但是我们人工查看耗时耗力。

本系统期望将爬虫技术和自然语言处理技术结合起来实现对在线商品评论的自动分析。系统通过使用爬虫自动抓取商品的评论数据，并通过情感倾向分析获取每条评论的好评度，通过图表向用户直观地展示商品的评论情况和基本信息。系统当前内置了多种分析方式，包括基于情感字典、基于机器学习以及基于百度Al的不同模块，通过配置文件进行更改。以期望使用者拥有良好的使用体验，系统具有操作简便、易扩展等优点。

**关键词**：网购；爬虫；自然语言处理；情感倾向分析

**DESIGN AND IMPLEMENTATION OF ONLINE PRODUCT USER COMMENT ANALYSIS SYSTEM**

**Abstract:** With the popularity of the Internet, online shopping has become the norm in our lives. In online shopping, we often analyze the merits of products through the evaluation of others' products. But we manually look at the time and effort.

The system expects to combine crawler technology with natural language processing techniques to enable automated analysis of online product reviews. The system automatically captures the review data of the product by using the crawler, and obtains the popularity of each comment through the sentiment orientation analysis, and visually displays the review status and basic information of the product through the chart. The system currently has a variety of analysis methods built in, including emotion-based dictionaries, machine-based learning, and different modules based on Baidu Al, which are changed through configuration files. In order to have a good user experience, the system has the advantages of easy operation and easy expansion.

**Keywords:**Online Shopping,Web Crawler,Natural Language Processing, Sentiment Orientation Analysis

目 录

[**本 科 生 毕 业 设 计（论文）** 1](#_Toc8074291)

[**本科生毕业设计（论文）诚信承诺书** 2](#_Toc8074292)

[1 绪论 1](#_Toc8074293)

[1.1 开发背景 1](#_Toc8074294)

[1.2 设计目的 1](#_Toc8074295)

[2 可行性研究 2](#_Toc8074296)

[2.1 技术可行性 2](#_Toc8074297)

[2.2 经济可行性 2](#_Toc8074298)

[2.3 操作可行性 2](#_Toc8074299)

[3 需求分析 3](#_Toc8074300)

[3.1 系统需求目标 3](#_Toc8074301)

[3.2 系统功能分析 3](#_Toc8074302)

[4 总体设计 4](#_Toc8074303)

[4.1 系统流程图 4](#_Toc8074304)

[4.2 数据流程图 5](#_Toc8074305)

[4.3 数据库设计 6](#_Toc8074306)

[4.3.1 商品信息表 7](#_Toc8074307)

[4.3.2 商品评论表 7](#_Toc8074308)

[4.3.3 评论分析表 7](#_Toc8074309)

[4.3.4 邮箱地址表 8](#_Toc8074310)

[4.4 系统模块介绍 8](#_Toc8074311)

[4.4.1 管理模块介绍 8](#_Toc8074312)

[4.4.2 图表显示模块介绍 9](#_Toc8074313)

[4.4.3 爬虫模块介绍 9](#_Toc8074314)

[4.4.4 分析模块介绍 9](#_Toc8074315)

[4.4.5 交互模块介绍 9](#_Toc8074316)

[4.4.6 邮件发送模介绍 10](#_Toc8074317)

[5 详细设计 11](#_Toc8074318)

[5.1 管理模块 11](#_Toc8074319)

[5.2 图表显示模块 12](#_Toc8074320)

[5.3 爬虫模块 13](#_Toc8074321)

[5.4 分析模块 16](#_Toc8074322)

[5.4.1 基于机器学习的分析模式 17](#_Toc8074323)

[5.4.2 基于情感字典的分析模式 22](#_Toc8074324)

[5.4.3 基于百度Al的分析模式 25](#_Toc8074325)

[5.5 交互模块 26](#_Toc8074326)

[5.6 邮件发送模块 28](#_Toc8074327)

[6 编码设计 30](#_Toc8074328)

[6.1 编码设计风格 30](#_Toc8074329)

[6.2 编码设计思想 30](#_Toc8074330)

[7 运行结果 31](#_Toc8074331)

[7.1 用户界面 31](#_Toc8074332)

[7.2 管理员界面 33](#_Toc8074333)

[8 总结 35](#_Toc8074334)

[9 参考文献 36](#_Toc8074335)

[10 致谢 37](#_Toc8074336)

# 绪论

## 开发背景

这是一个互联网的时代，随着互联网的持续发展，人们与网络的距离日益贴近，其中很为明显的就是网络购物。随着电子商务的急剧增长，网络购物越来越受到欢迎，占据了零售的很大一部分。而类似“双11”之类的活动的推出更是将网络购物推向了巅峰。现在的生活节奏越来越快，以往的慢节奏已经不大适应现在的人们了，显然网络购物这种快节奏的活动方式更适合现在的人们。

当我们想要购买一个不是立即需要的商品时，我们往往想到的是网络购物，例如淘宝、天猫、京东或者考拉，因为在网络上的商品常常会比实体商铺里的更加便宜也更加方便，我们通过网购平台就可以购买到心仪的商品。据调查显示，网购在购物消费总数中的占比逐年增加。但是同时由于监管不便的原因，网购也存在很多坑点。用户通过商品的评价自我分析商品的优劣，这往往耗时耗力。本系统本着方便买家的立场上进行开发，目的是降低网络购物的风险。系统主要采用爬虫技术和文本情感分析来实现，通过参考系统给出的分析结果来判断商品的优劣。

## 设计目的

本课题以设计实现对在线商品的评论进行自动抓取，分析评论的情感倾向，展示分析结果为目的。并以锻炼和检验大学四年的学习成果、程序设计能力、编码开发能力，目标实现一个简单易操作、后续扩展方便、具有一定准确度的在线商品评论分析系统。网络购物时所购买的商品良莠不齐，虽然网络平台提供了商品评论区域，但商品评论过于繁多杂乱，并且有水军混杂其中，对于我们鉴别商品好坏存在扰乱和不便，而且人力查看较慢，一般无法查看所有评论数据，往往只会浏览其中几页甚至几条评论，对于整体的评论信息没有一个更多的把握。本系统旨在爬取所有的评论数据，生成所有评论数据的情感倾向分布图，以及整体评价的分布图，使用直观的图表展示数据，给使用者简单易用的使用体验。

# 可行性研究

可行性研究的目的是用最小的代价在最短的时间内确定问题是否值得解决、是否能够使用现有的技术解决，将从以下三个方面进行本系统的可行性分析。

## 技术可行性

系统使用django作为基准开发框架。python的用途广泛，目前分别可以应用于爬虫技术、数据挖掘、数据分析、web交互等。技术比较成熟，可行性高，并且具有众多第三方开发库。对于爬虫，python具有强大的requests第三方开发库，可以使用较少的语句实现一个简单的http/https爬虫请求发送和处理。对于数据分析处理，python可以使用jieba分词进行中文文本的分词。NLTK是在自然语言处理中最常使用的一个python库。SnowNLP可以进行中文文本内容的分析和挖掘。更进一步，各大IT公司具有许多强大的Al平台，如百度Al开放平台、腾讯Al开放平台等开放平台可以使用，其中百度Al开放平台具有免费、使用方便等优点，本系统借助百度Al开放平台完成一部分的评论分析引擎的编写。系统使用Mysql作为数据库引擎，Mysql具有简单易用，对于个人开发者免费，具有轻量化，性能较强，可移植性高、运行速度较快等优点，且部署方便。系统各方面技术均成熟可用，在技术方面可行。

## 经济可行性

本系统开发工具采用pycharm、vscode、datagrip，其中数据库使用Mysql，开发语言使用python，Web框架使用django。数据库部署平台为Linux，Web应用部署平台为Windows。以上产品，对于个人为免费产品。开发成本低，经济方面可行。

## 操作可行性

本系统设计清晰，具有简单易操作的用户界面，有完善的异常处理机制和错误信息提醒和后台日志记录，用户通过操作界面的提示即可完成操作。通过系统的后台管理界面可以操作系统数据库，方便完成日常的数据维护。操作方面可行。

# 需求分析

## 系统需求目标

本次设计目标开发基于python3、django的在线商品用户评论分析系统，采用b/s模式。以下是本设计要开发完成的五个需求目标：

1. 实用性：设计界面美观实用简洁，软件与用户交互性较强。
2. 稳定性：软件目标是实现用户通过网页界面可实现在线商品的评论分析。采用成熟框架编写，稳定性较高。
3. 友好性原则：操作简单是任何软件的开发的目标，操作简单并不意味着功能简单，将复杂的功能简单化会给用户留下很深的印象。软件在各个阶段给用户以良好的提示。能保证用户的使用体验。
4. 安全性原则：每个使用者无需经过任何的身份验证。各项连接出错要有良好的出错提示。关闭服务端的信息泄露，保证服务器的安全。
5. 可扩展性原则：设计友好的接口，方便系统后续的叠代更新。对业务代码服务化，交互和逻辑分离，为了以后更好拓展，更好维护，重要功能实现独立化，可以脱离整体系统使用。

## 系统功能分析

软件要求不断完善，功能也是逐步添加的，以下为该在线商品用户评论分析系统的基本功能需求：

* 用户通过搜索栏输入需要进行分析的商品地址，系统将商品地址返回后台处理。
* 读取商品链接，如果该商品已经进行过评论分析，则读取该商品的分析总结发送给用户，否则通知爬虫引擎进行爬取。
* 爬虫引擎读入商品链接，爬取商品评论和相关信息，存入数据库，等待分析引擎处理。
* 编写多个分析引擎，可以通过配置文件更改分析引擎的使用，并实现单独的接口化。
* 分析引擎读取商品评论，分析评论信息，生成商品分析总结，存入数据库。
* 用户在输入新的商品链接同时可以添加自己的邮箱，在分析完成后，如果该商品有绑定的需要提醒的邮箱，则发送相应邮件提醒用户分析已完成，并附带相关分析结果的链接。
* 具有后台管理页面，使管理员可以通过网页端操作数据库和系统。

# 总体设计

## 系统流程图

该软件为在线商品用户评论分析系统，总的系统规划基于传统的b/s模式。可以实现多用户同时在线，如下图所示：

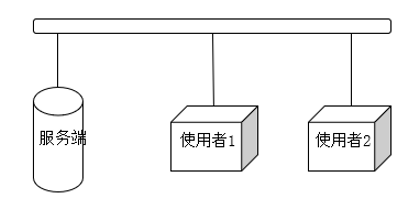


图 4‑1系统模式

系统对于操作成功和失败，异常错误都应存在相应的提示，以便用来帮助用户使用操作系统，同时也可以帮助开发者和管理员更好的开发、维护系统。

系统的流程图如下：

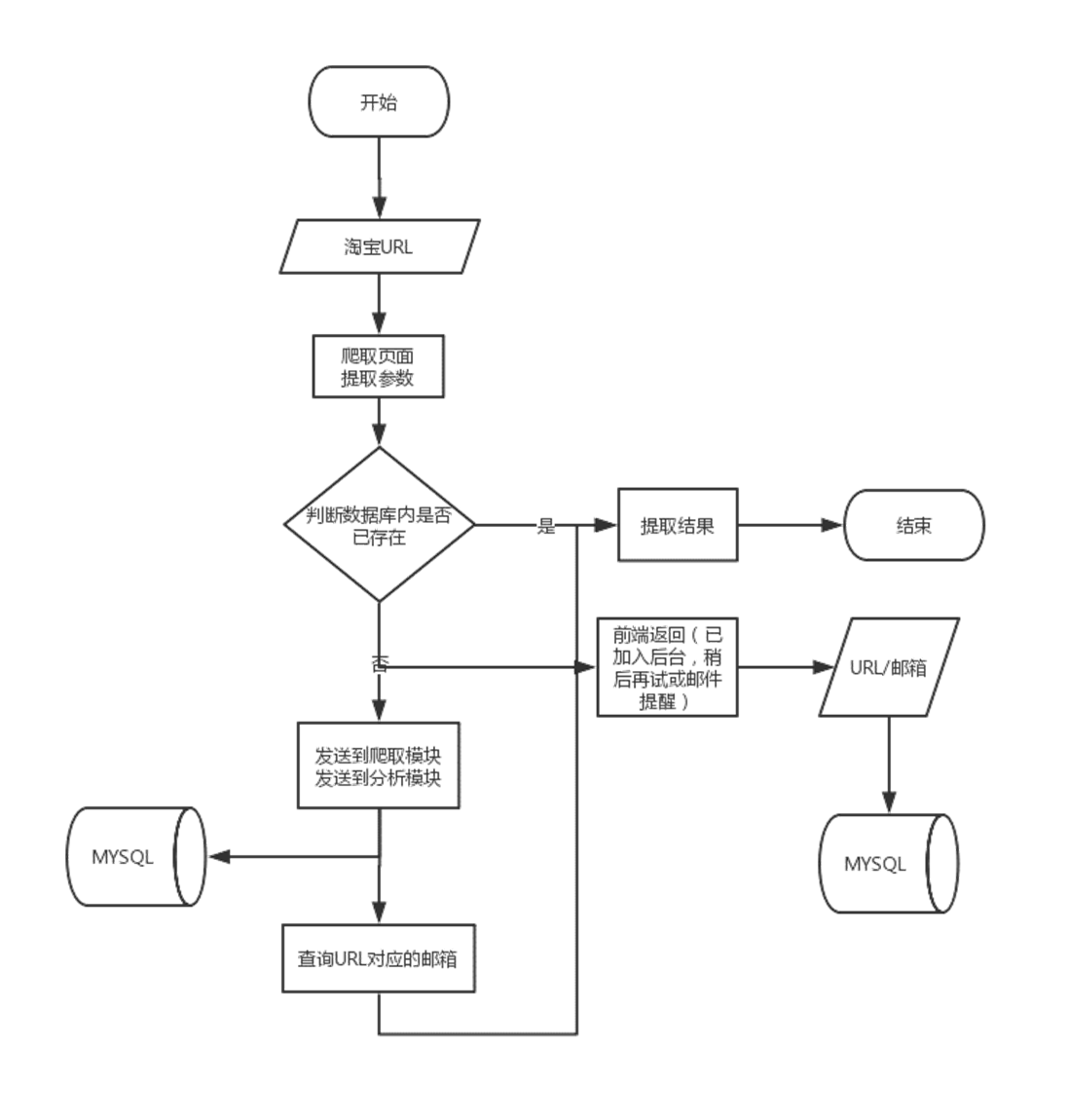


图 4‑2系统流程图

## 数据流程图

数据流程图显示了通过系统的数据流的逻辑，以下是本系统的数据流程图：

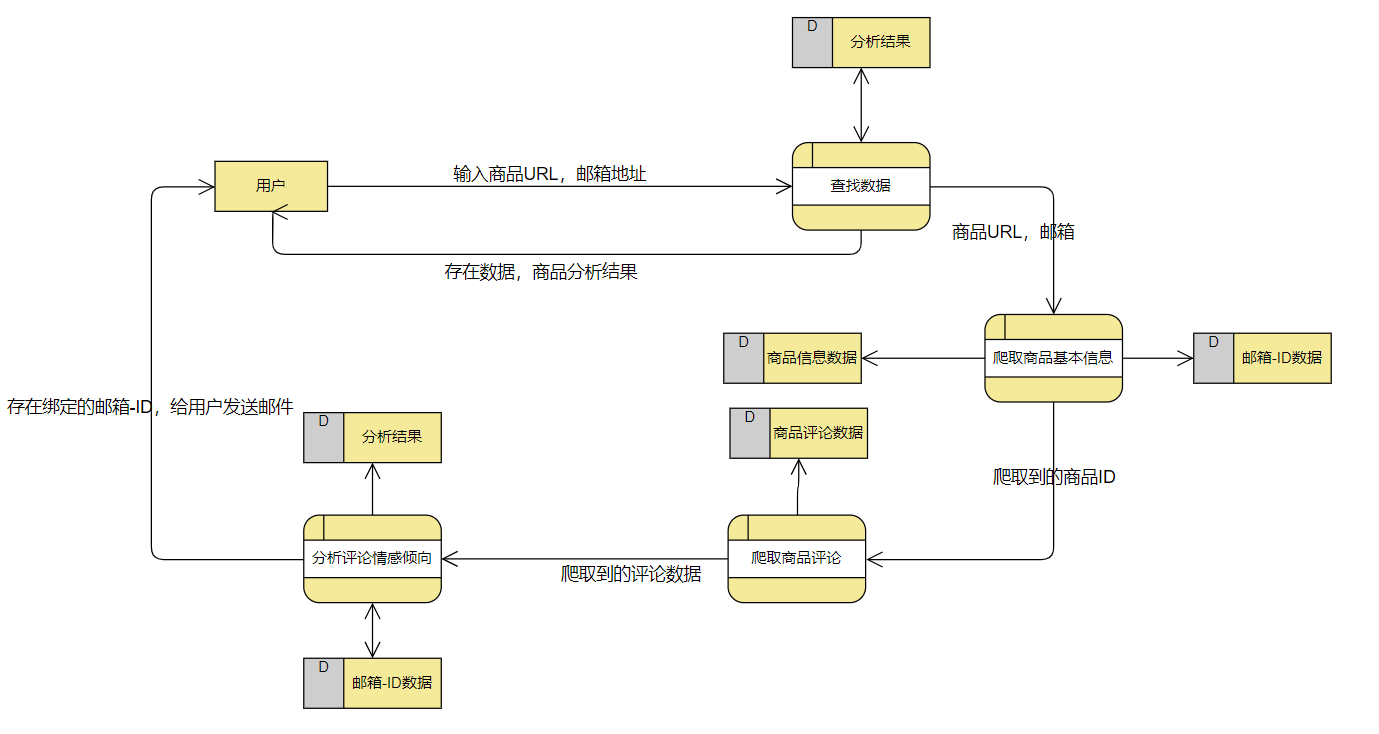


图 4‑3数据流程图

## 数据库设计

本系统共设计了商品信息、商品评论、评论分析、邮箱地址，共四个数据库表，以下为数据库的E-R图：

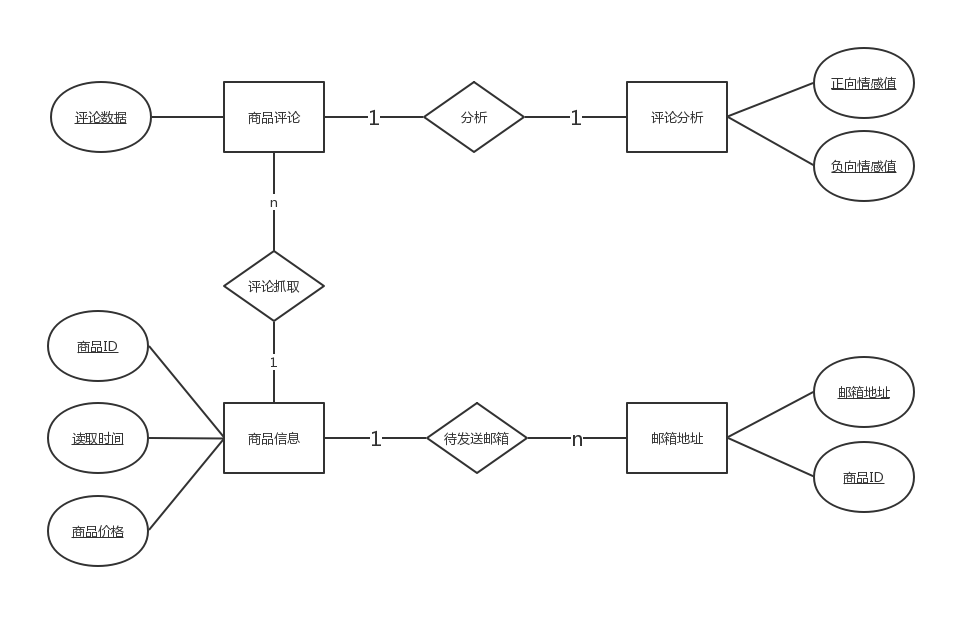


图 4‑4数据库E-R图

### 商品信息表

商品信息表以保存商品的有关信息，包括商品ID，商品URL地址，商品价格，保存时间等字段，表名为：MyModel\_taobao，商品信息表的表结构设计如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 字段意义 | 字段类型 | 是否主键 | 允许为空 | 备注 |
| 1 | id | 序号 | int | 是 | 是 | 自动生成 |
| 2 | taobao\_name | 商品名称 | varchar(100) |  | 否 |  |
| 3 | taobao\_id | 商品ID | varchar(50) |  | 否 |  |
| 4 | taobao\_url | 商品URl | varchar(512) |  | 否 |  |
| 5 | taobao\_shop\_name | 商品店铺名称 | varchar(100) |  | 否 |  |
| 6 | taobao\_price\_now | 商品价格 | varchar(50) |  | 否 |  |
| 7 | taobao\_time | 保存时间 | date |  | 否 | 自动生成 |

### 商品评论表

商品评论表以保存商品评论的有关信息，包括商品ID，商品总体评论，商品详细评价，保存时间等字段，表名为：MyModel\_spider，商品评论表的表结构设计如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 字段意义 | 字段类型 | 是否主键 | 允许为空 | 备注 |
| 1 | id | 序号 | int | 是 | 是 | 自动生成 |
| 2 | spider\_id | 商品ID | varchar(50) |  | 否 |  |
| 3 | spider\_detail\_Common | 商品总体评价JSON | varchar(3000) |  | 否 |  |
| 4 | spider\_detail\_All | 商品详细评价JSON | longtext |  | 否 |  |
| 5 | spider\_time | 保存时间 | date |  | 否 | 自动生成 |

### 评论分析表

评论分析表以保存评论的分析结果的有关数据，包括商品ID，商品评价数据的正、负向情感倾向值，保存时间等字段，表名为：MyModel\_analyse，评论分析表的表结构设计如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 字段意义 | 字段类型 | 是否主键 | 允许为空 | 备注 |
| 1 | id | 序号 | int | 是 | 是 | 自动生成 |
| 2 | analyse\_id | 商品ID | varchar(50) |  | 否 |  |
| 3 | analyse\_positive\_prob | 正向情感倾向值JSON | longtext |  | 否 |  |
| 4 | analyse\_negative\_prob | 负向情感倾向值JSON | longtext |  | 否 |  |
| 5 | analyse\_time | 保存时间 | date |  | 否 | 自动生成 |

### 邮箱地址表

邮箱地址表以保存待发送的邮箱的有关数据，在评论分析完成后可以通过该表查询是否有需要发送的邮件，该表包括商品ID，邮箱，保存时间等字段，表名为：MyModel\_mail，邮箱地址表的表结构设计如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 序号 | 字段名 | 字段意义 | 字段类型 | 是否主键 | 允许为空 | 备注 |
| 1 | id | 序号 | int | 是 | 是 | 自动生成 |
| 2 | taobao\_id | 商品ID | varchar(50) |  | 否 |  |
| 3 | mail | 待发送邮件的邮箱地址 | varchar(100) |  | 否 |  |
| 4 | add\_time | 保存时间 | date |  | 否 | 自动生成 |

## 系统模块介绍

系统主要模块包括管理模块、图表显示模块、爬虫模块、分析模块、交互模块、邮件发送模块这六个主要模块，下图为系统功能模块图：

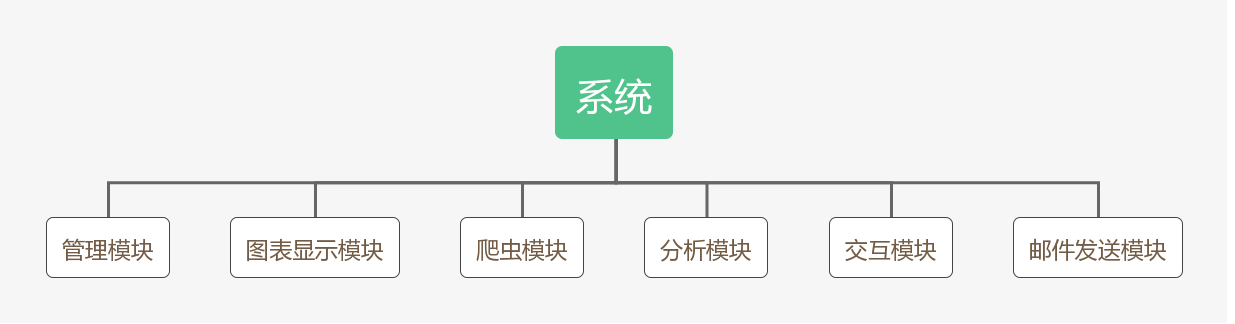


图 4‑5系统模块

### 管理模块介绍

管理员可以登录管理员界面查看系统数据库信息，进行系统常规设置等，下图为管理模块的设计概要：

|  |  |  |  |
| --- | --- | --- | --- |
| 功能编号 | 4.4.1 | 功能名称 | 管理模块 |
| 功能描述 | | 管理员通过该模块进行操作数据库 | |
| 输入项 | | 操作数据库的数据 | |
| 处理描述 | | 查询、增加、删除、更新数据库的数据 | |
| 输出项 | | 数据库详情 | |
| 界面要求 | | 图形化用户界面 | |

### 图表显示模块介绍

读取分析结果，生成简单易懂的图表，为用户直观的呈现分析结果，下图为图表显示模块的设计概要：

|  |  |  |  |
| --- | --- | --- | --- |
| 功能编号 | 4.4.2 | 功能名称 | 图表显示模块 |
| 功能描述 | | 商品ID对应的分析结果图表显示 | |
| 输入项 | | 商品ID | |
| 处理描述 | | 通过商品ID查询商品分析数据表中的分析结果，通过HTML页面的图表向用户显示分析结果 | |
| 输出项 | | 无有效数据信息提示  分析结果数据、图表 | |
| 界面要求 | | 显示有图形化用户界面，通过url接口调用 | |

### 爬虫模块介绍

将商品链接发送到爬虫引擎，爬取商品链接下的评论，存入数据库等待分析模块运行。适当使用反爬技术，设置合理的爬取速度防止IP、账户被封，下图为爬虫模块的设计概要：

|  |  |  |  |
| --- | --- | --- | --- |
| 功能编号 | 4.4.3 | 功能名称 | 爬虫模块 |
| 功能描述 | | 商品ID对应的评论数据爬取 | |
| 输入项 | | 商品ID | |
| 处理描述 | | 通过请求相关的信息接口，爬取商品评论数据，并进行数据清理和分类 | |
| 输出项 | | 商品评论数据 | |
| 界面要求 | | 无图形化用户界面，通过函数接口调用 | |

### 分析模块介绍

将爬取到的商品评论发送到分析引擎，进行评论处理和分析，去除无用信息，合并相似信息，分析评论信息，最后生成分析结论，下图为分析模块的设计概要：

|  |  |  |  |
| --- | --- | --- | --- |
| 功能编号 | 4.4.4 | 功能名称 | 分析模块 |
| 功能描述 | | 评论数据情感倾向值分析 | |
| 输入项 | | 评论数据 | |
| 处理描述 | | 分析评论数据的情感倾向值，分别保存正向情感倾向值和负向情感倾向值 | |
| 输出项 | | 情感倾向值数据 | |
| 界面要求 | | 无图形化用户界面，通过url接口调用 | |

### 交互模块介绍

用户通过网页来查看分析的结果，也可以通过邮件接收分析的结果。由用户输入商品链接，传输到服务器，服务器进行评论爬取、分析，并生成分析结果发送给用户。商品链接可一次性输入多条，使用分号（“;”）或回车分割。用户通过填写邮箱，可以接收分析完成的提醒，减少不必要的等待时间，下图为交互模块的设计概要：

|  |  |  |  |
| --- | --- | --- | --- |
| 功能编号 | 4.4.5 | 功能名称 | 交互模块 |
| 功能描述 | | 用户使用的交互界面 | |
| 输入项 | | 商品URL地址 | |
| 处理描述 | | 接收用户输入的URL地址，调用其他模块进行交互，完成评论抓取、分析模块的衔接，存在错误时及时提醒用户 | |
| 输出项 | | 开始分析  分析结果  错误提示 | |
| 界面要求 | | 图形化用户界面 | |

### 邮件发送模介绍

读取设置好的邮箱地址和密码，以此作为发件人，向需要发送提醒的用户发送邮件，提示用户分析已完成，下图为邮件发送模块的设计概要：

|  |  |  |  |
| --- | --- | --- | --- |
| 功能编号 | 4.4.6 | 功能名称 | 邮件发送模块 |
| 功能描述 | | 发送需要进行提醒的邮件 | |
| 输入项 | | 商品ID | |
| 处理描述 | | 查找邮箱地址数据库表内，该商品ID是否有对应需要发送的邮件，如果存在邮件则给用户发送提醒邮件 | |
| 输出项 | | 发送成功  发送失败 | |
| 界面要求 | | 无图形化用户界面，通过函数接口调用 | |

# 详细设计

系统主要模块包括管理模块、图表显示模块、爬虫模块、分析模块、交互模块、邮件发送模块，接下来将从这六个主要模块对系统的设计进行描述。

## 管理模块

管理员可以登录管理员界面查看系统数据库信息，进行系统常规设置等。

django预设了后台管理的模板，可以通过django的后台管理模板快速地生成一个管理页面。通过激活管理工具和创建超级用户即可使用：

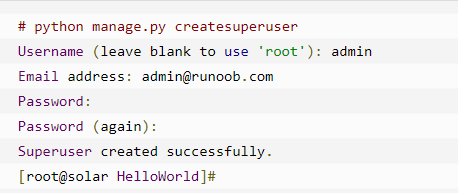


图 5‑1创建管理员账户

在admin.py文件中绑定模型，即可方便地进行数据库的管理，同时设置数据库显示的字段，配置条件筛选，字段查找等功能，如下所示，进行模型注册和相关的配置：



图 5‑2注册模型

通过登陆账户，可以看到管理模块可以进行管理员账户的认证和授权以及数据库的维护相应的功能：



图 5‑3后台页面

## 图表显示模块

读取分析结果，生成简单易懂的图表，为用户直观的呈现分析结果。

图表显示模块需要接收GET方法传入的商品ID，查询数据相应的数据，再通过django的模板和数据进行绑定，生成相应的图表。

这里通过Chart.js的图表显示插件，方便地进行柱状图、饼状图、折线图等的生成。

通过商品ID从数据库中查询相应的数据。

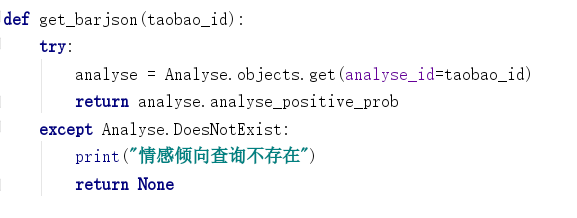


图 5‑4查询数据

之后进行相应的处理，将数据调整成Chart.js所支持的数据格式，并设置到list中，方便之后传递到前端。



图 5‑5处理查询的数据

最后将数据返回到前端界面，通过Chart.js插件完成相应的图表生成。

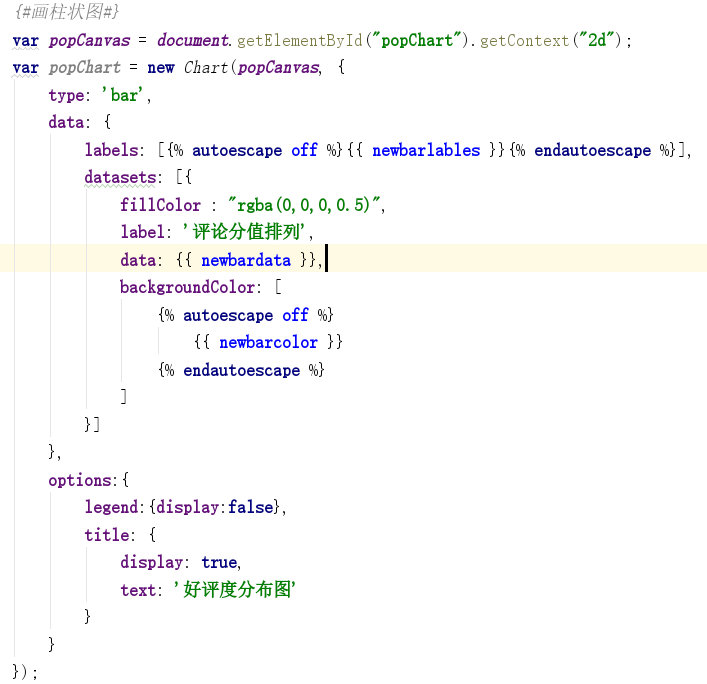


图 5‑6生成图表

## 爬虫模块

将商品链接发送到爬虫引擎，爬取商品链接下的评论，存入数据库等待分析模块运行。适当使用IP代理池和反爬技术，设置合理的爬取速度防止IP、账户被封。

爬虫模块需要将商品的评论和总体评价数据进行爬取，常规做法是通过scrapy爬取页面数据，但是这样需要爬取整个淘宝页面的数据，而且对于AJAX请求的数据无法很好地进行处理，这里选用分析淘宝商品详情页面载入过程中的请求，选取其中返回评论JSON的请求，分析该请求的参数和认证授权方式，通过使用该请求接口来获得淘宝商品的评论。这样的好处是减少代码量，减轻系统的体量，同时也可以提供爬虫模块的运行速度，爬虫极度依赖网络的良好，若是爬取的请求过多，必然会增大系统运行不流畅，而使用特点请求接口可以大大地减少请求数，也可以减轻主机的CPU占用和网络消耗。

这里通过使用burp suite和火狐浏览器进行配合查看在浏览淘宝商品的过程中，会产生哪些请求，首先要对火狐浏览器设置代理和导入证书，之后开启burp suite的监听功能，浏览一个淘宝商品的评论后，再查看burp suite抓取到的请求，这里我们可以看到https://rate.taobao.com/feedRateList.htm请求是淘宝返回商品评价的JSONP请求

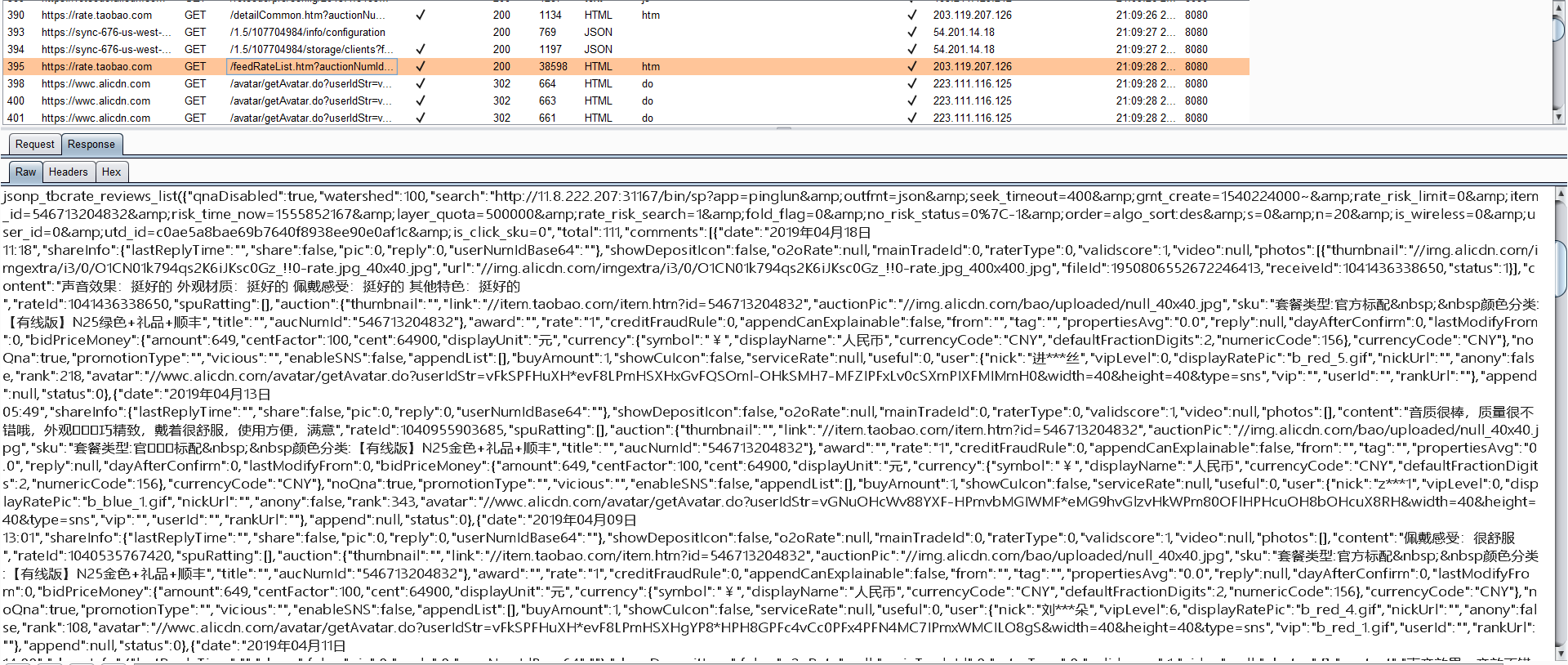


图 5‑7获取https请求

通过逐步删除请求中的相关参数，最后分析得到GET参数中的auctionNumId控制商品的ID，控制JSONP返回的评论数据是哪一个商品的，currentPageNum控制商品评论的页数，控制返回的评论数据是第几页的数据，pageSize控制商品评论的页面大小，控制返回的评论数据的多少，Cookie参数中的x5sec控制身份认证，在多次请求后，如果这个参数不存在或为无效参数，将会跳转到验证码机制的页面，但是通过分析发现，这个x5sec参数的数值不会过期，长时间或退出登陆任然有效，所以可以通过获取一个x5sec参数的数值并使用这个数值长期的获取商品评论的JSON数据。

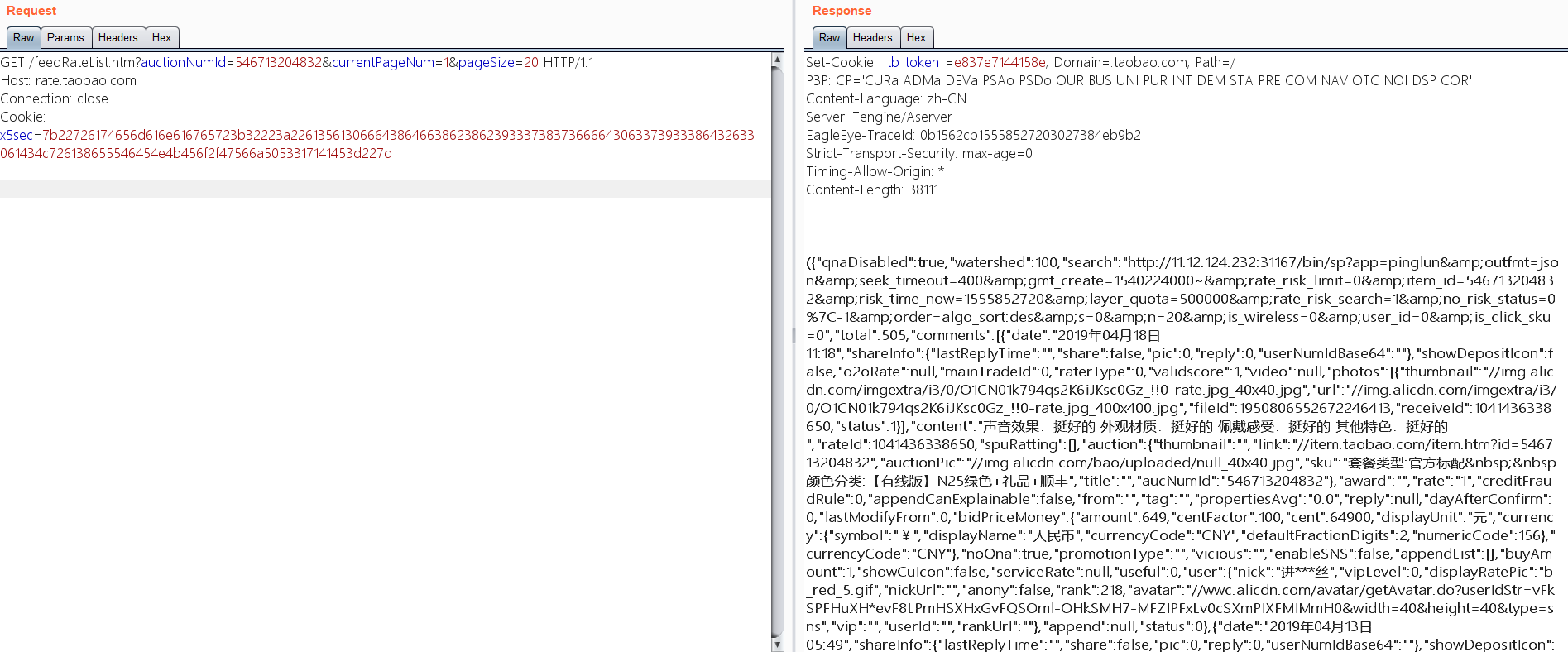


图 5‑8分析请求参数

当爬虫模块被调用并传入商品ID时，系统会创建一个新线程并开始评论数据的爬取



图 5‑9开启爬虫线程

通过请求之前分析获得请求接口，一页一页地发送请求，并将获得的返回JSON的数据进行清洗，整合，并存入数据库，在数据爬取完成后，通过调用分析模块的接口进行分析文本的情感倾向值。



图 5‑10爬取评论数据

## 分析模块

将爬取到的商品评论发送到分析引擎，进行评论处理和分析，去除无用信息，合并相似信息，分析评论信息，最后生成分析结论。

分析模块需要将爬虫模块获取到的商品评论进行分析，这里设置了三种分析方式，并可以独立使用。

自然语言处理的一个研究方向是分析文本的情感倾向值，分析一个文本情感倾向值，可以知道这个文本的情感是善意的还是恶意的，正向的情感倾向值表示文本的观点是积极的、友好的，负向的情感倾向值表示文本的观点是悲观的、恶意的。在在线商品用户评论分析系统中，若是一个评论的情感倾向值是正向的，则代表这个评论是一个好评，这个值越大代表好评度越高，若是一个评论的情感倾向值是负向的，则代表这个评论是一个差评，这个值越大代表差评度越高。目前的情感倾向值分析主流使用两种分析方式，一种是基于情感字典的分析方式，这种分析方式依赖于情感字典，情感字典越丰富，准确率越高，优点是方便移植，情感字典较为通用，缺点是一个好的情感字典是需要一个长时间的构建，较为耗时耗力。另一种较为流行的分析方式是基于机器学习，通过大量的训练集来构建分类器，并通过一定数量的测试集进行校准和筛选不同的分类方法，优点是准确度较高，缺点时依赖分类器，一个领域构建的分类器给另一个领域必然是准确度大大降低的，甚至同一个领域的不同分支构建的分类器也会有很大的差别。

这个模块设置了三种不同的分析模式，可以通过配置文件进行更改，分别为基于SnowNLP的分析模式、基于情感字典的分析模式和基于百度Al的分析模式，接下来将分别介绍这三种模式的详细设计：

### 基于机器学习的分析模式

机器学习是当前进行情感倾向分析的有力途径，通过一定数量的训练数据对分类器进行训练校准，可以得到一个较为准确的分类模型。使用机器学习进行分析的基本流程为：

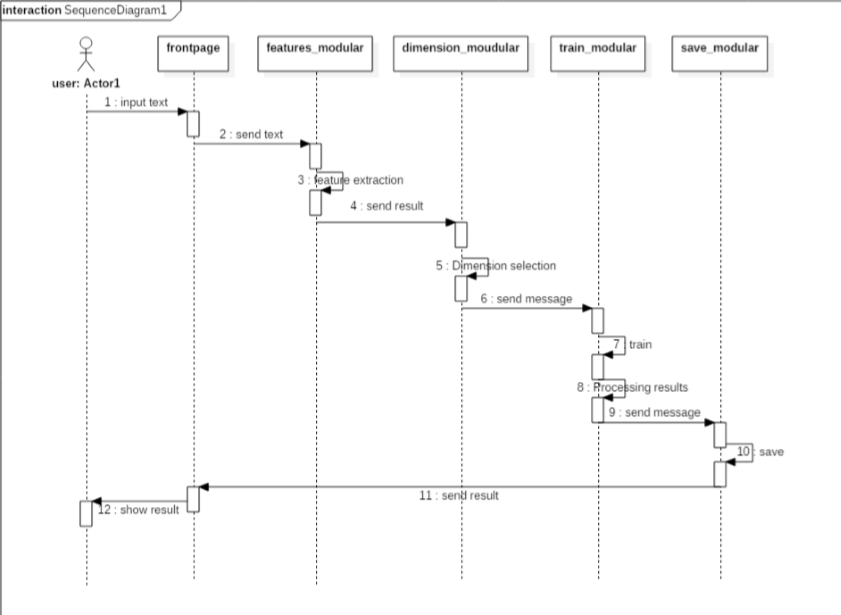


图 5‑11机器学习流程

1. 提取特征：

将文本提取为特征形式表示

把所有词作为特征进行提取

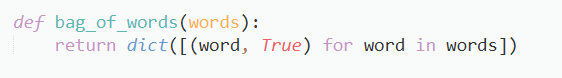


图 5‑12所有词特征提取

把双词作为特征进行提取

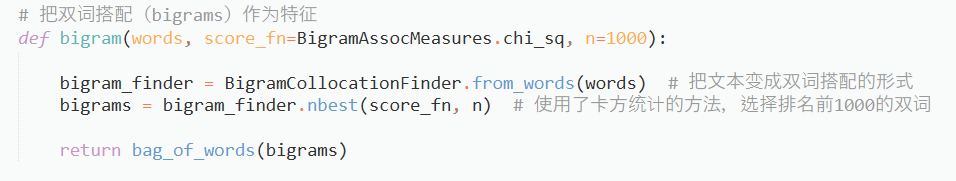


图 5‑13双词特征提取

把所有词搭配双词作为特征提取

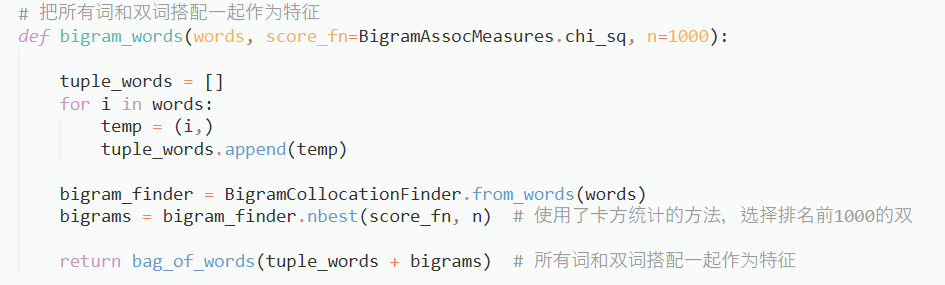


图 5‑14搭配特征提取

1. 选择特征

选择信息量较大的特性，抛除对分类训练影响较大的特性

计算每个词的信息量



图 5‑15计算每个词信息量

计算每个词和双词搭配的信息量



图 5‑16计算搭配的信息量

对信息量进行排序，选出信息量较大的词



图 5‑17选择特征值

经过以上步骤完成对文本的特性值的提取

1. 构建分类器，检验分类的准确度，并保存分类器

载入数据

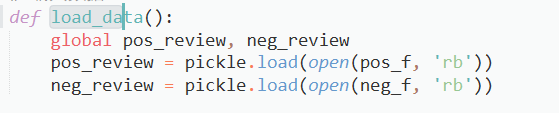


图 5‑18载入训练集

赋予类标签



图 5‑19赋类标签

将测试集分割成训练集和开发测试集，并且分割标注的数据和标签



图 5‑20分割数据

训练分类器，并使用开发测试集校验准确度

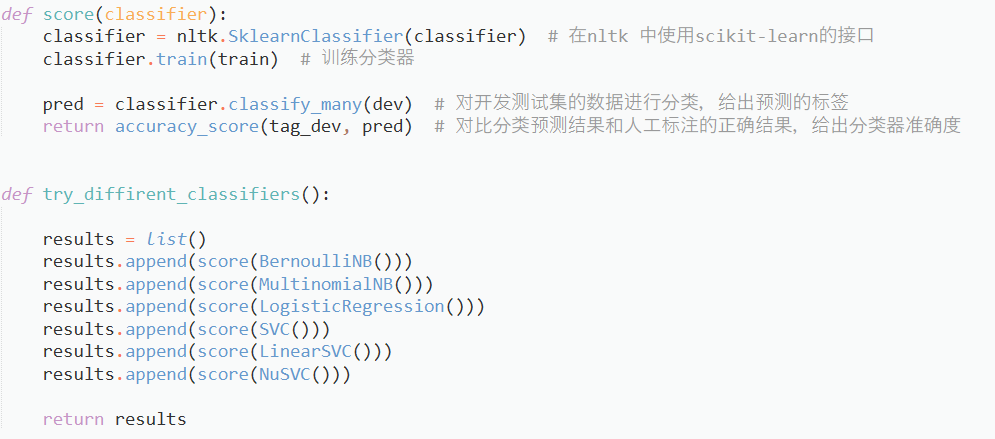


图 5‑21计算正确率

选择其中准确率较高的分类方法，这里选择了朴素贝叶斯分类器以及词和双词搭配的特征提取方式，并存储分类器，将训练好的分类器保存到文件中

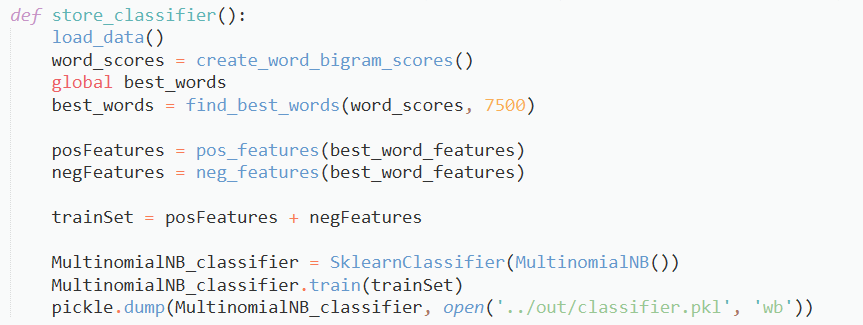


图 5‑22保存分类器

1. 最后使用训练好的分类器进行文本的情感分析

读入URL传递的文本数据，清理读入的数据，删除不必要的特殊符号

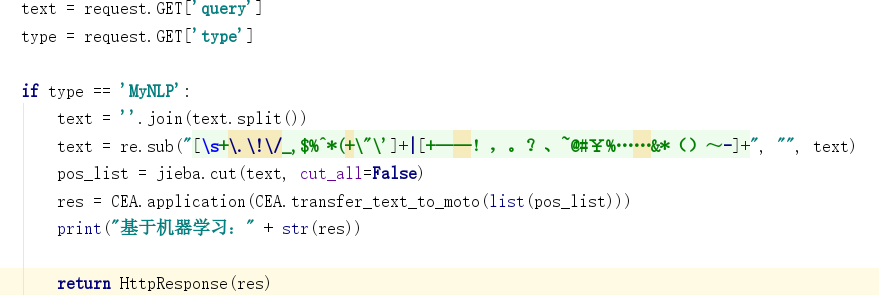


图 5‑23清理文本

将文本转换为特征表示

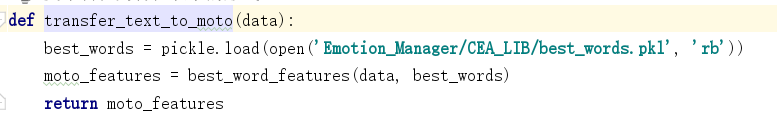


图 5‑24特征转换

载入分类器计算情感倾向的概率值



图 5‑25计算情感值

### 基于情感字典的分析模式

情感字典是另一种较为常用的分析模式，在不同领域经常使用差不多相近情感字典，较为通用但准确率较低。运用情感字典进行分析的基本流程为：

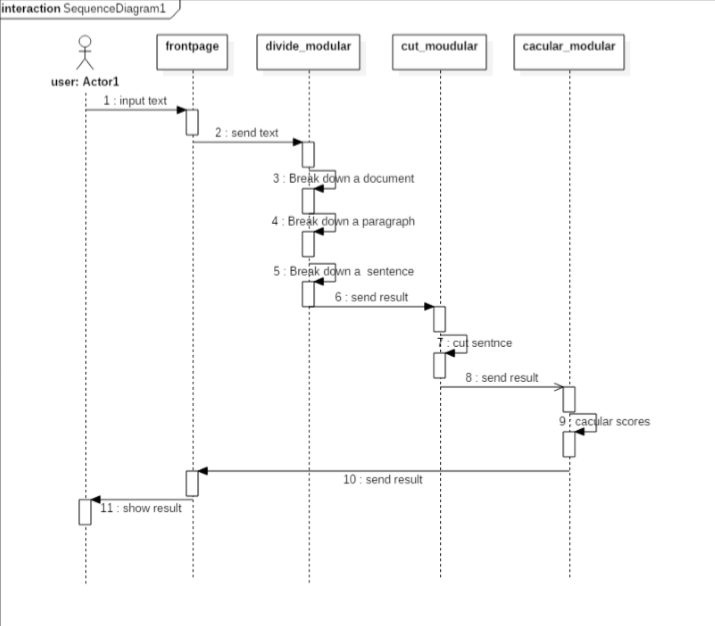


图 5‑26情感字典流程

1. 加载情感字典：

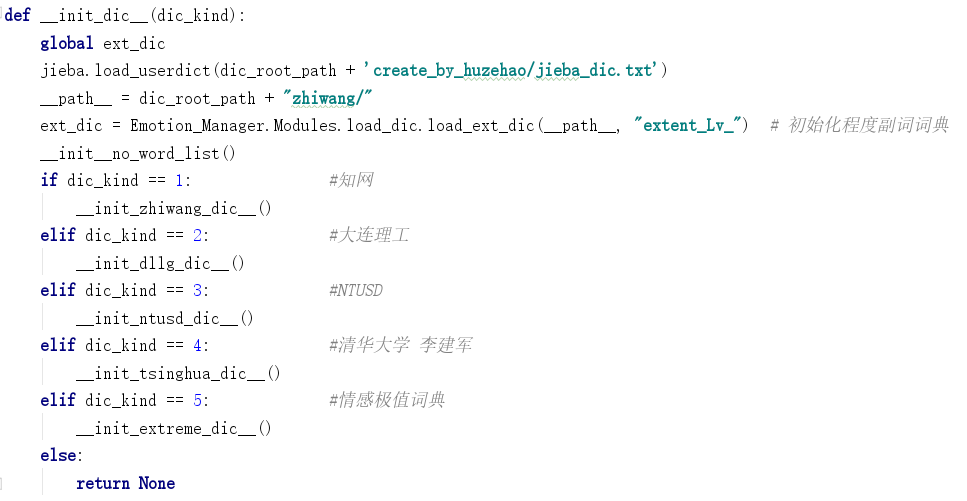


图 5‑27载入情感字典

1. 得出词语详细信息：

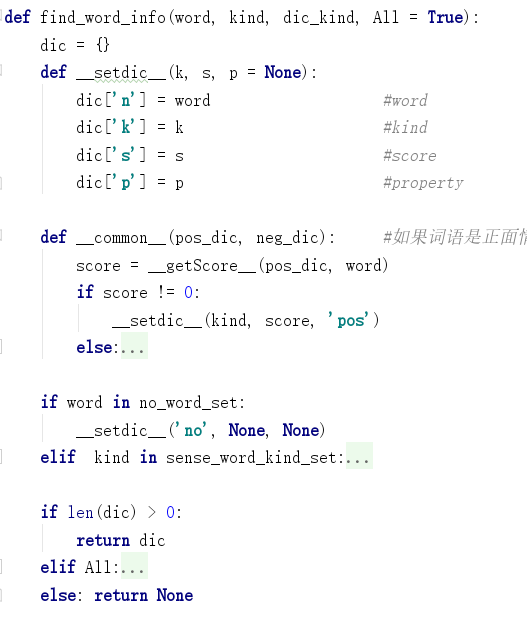


图 5‑28获取词语信息

1. 计算短句子情感值：

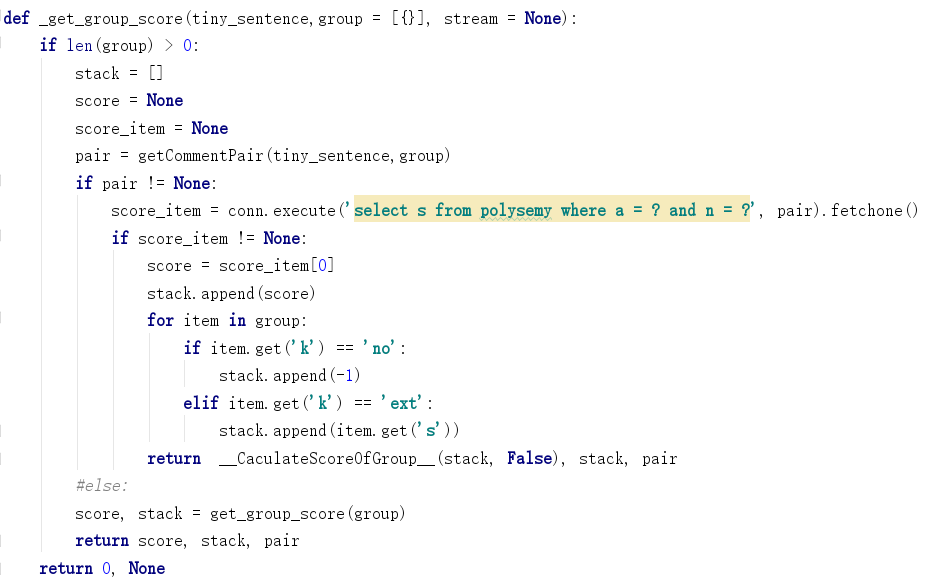


图 5‑29计算句子情感值

1. 获取整个中文文本的情感值：

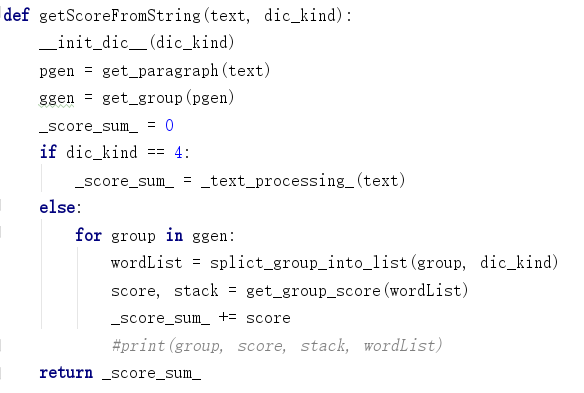


图 5‑30计算文本情感值

### 基于百度Al的分析模式

百度Al开放平台提供了文本情感倾向分析的相关接口，通过使用开放平台可以快速地构建系统，缩短情感倾向分析模块的开发时间。在百度Al平台获得相关授权后，通过API文档的说明进行构建情感倾向分析模块。

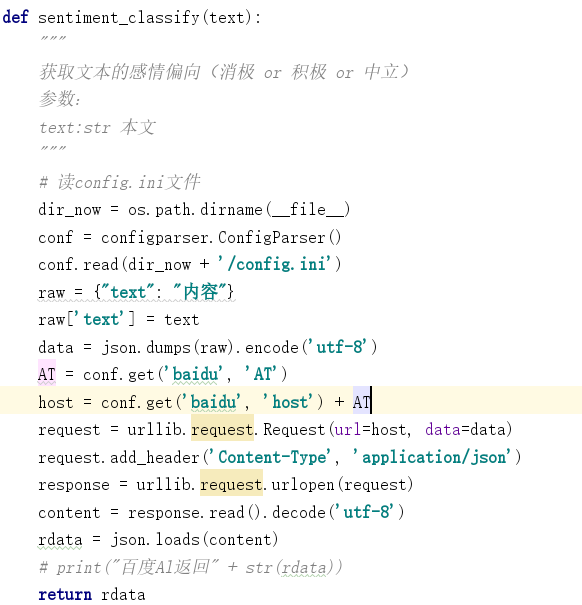


图 5‑31获取文本情感值

目前百度Al开放平台以支持训练集数据定制的情感倾向分析，可以使分析获得的数据更加准确。

## 交互模块

用户通过网页来查看分析的结果，也可以通过邮件接收分析的结果。由用户输入商品链接，传输到服务器，服务器进行评论爬取、分析，并生成分析结果发送给用户。商品链接可一次性输入多条，使用分号（“;”）或回车分割。用户通过填写邮箱获取分析结果完成的提醒，减少不必要的等待时间。

交互模块需要将用户传入的商品URL地址进行接收，并通过调度爬虫模块和分析模块进行数据的爬取和分析。

为了用户良好的感官体验，这里运用AJAX技术进行数据的传递，用户可以传入多个URL地址，并通过回车、逗号或者分号进行分隔，用户在输入URL地址时也可以同时输入自己的邮箱地址，前端会先进行一个校验，排除错误的、不规范的URL地址和邮箱地址



图 5‑32发送AJAX请求

后端接收到数据后，会再进行一次校验，排除有人绕过前端和校验或者网络出现问题的情况，通过查询URL地址所对应的数据，若该地址对应的商品ID存在相应的数据，已经进行过评论的爬取和分析，将会返回商品ID，由前端再向图表显示模块发送获取分析结果的请求。若是不存在相应数据，将会向调用爬虫模块进行相关评论数据的爬取，并向前端返回数据开始爬取分析的提示消息。若是同时前端返回了邮箱地址，则会在开始爬取时将商品ID和邮箱地址存入相应的数据库，以便在分析结束后向用户发送提醒邮件

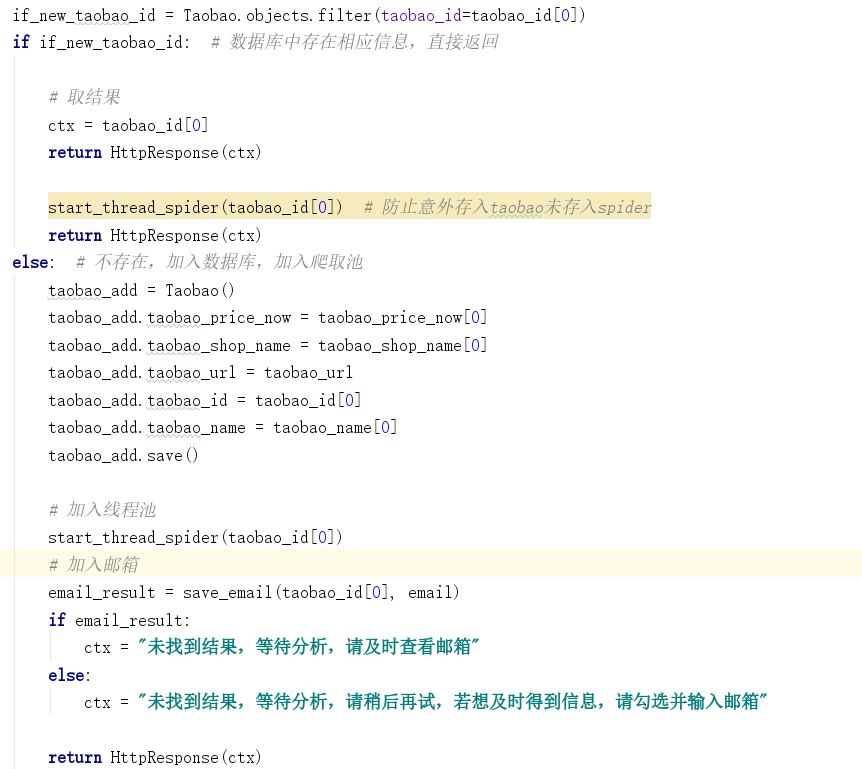


图 5‑33处理请求参数

## 邮件发送模块

读取设置好的邮箱地址和密码，以此作为发件人，向需要发送提醒的用户发送邮件进行提醒，告知用户评论分析已完成。

邮箱发送模块需要在当评论数据分析完成后调用时，通过传递的商品ID查询邮箱地址数据表，查看是否存在相应需要发送提醒邮件的邮箱地址，若不存在，直接返回。若存在，则首先读取需要发送邮件的邮箱地址并使用传递的商品ID产生拼接后的可以查看分析结果的URL地址，地址经过邮件发送给用户，在提醒用户分析完成的情况下同时发送相应的URL地址，方便用户的查看。这里使用python内置的mail库进行邮件的发送。

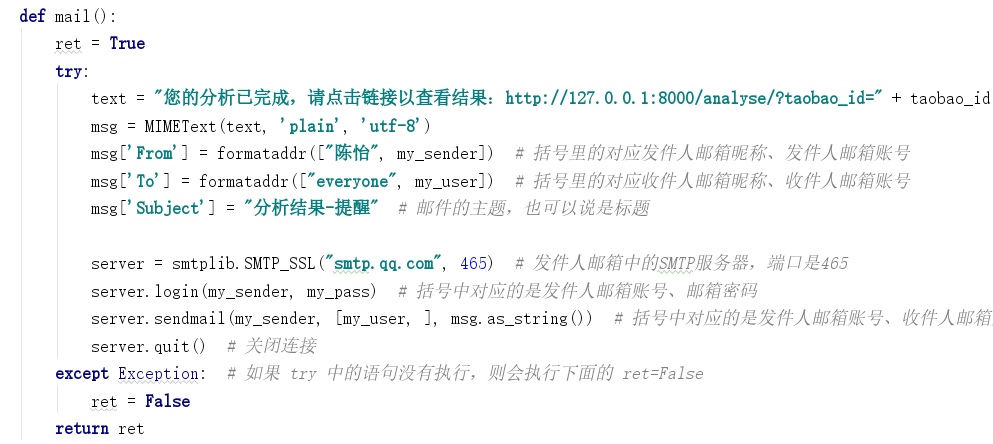


图 5‑34发送邮件

# 编码设计

## 编码设计风格

本系统编码设计风格应遵循以下基本规范：

1. 关于系统中出现的变量名、函数名、数据库表名、字段名等应采用能体现其功能或特点的英文单词组合或缩写命名。这样编码的好处是方便理解和记忆，提高代码的可读性。方便一段时间后再次对系统进行迭代时不需要耗费时间去理解变量名、函数名、表名、字段名等，可以快速地投入编码设计中。
2. 程序中复杂过程及关键功能的实现都应在代码中有详细注释，相关功能函数的实现都应在函数定义时有功能说明注释和调用说明注释。这样的好处是便于进行代码的编写、调试和后续的版本迭代。
3. 系统的交互要趋于简单清晰，减少用户的学习成本，方便用户使用的同时要注意有相关的提示，增强用户的体验感受。
4. 系统的各个模块应尽量趋向于独立化，实现模块可以脱离系统整体使用。这样设计的好处是代码编写实现时可以分块进行，而不用时时都考虑到系统的整体。在调试代码时也可以分模块进行调试，减少代码调试、bug查找、修复时候的工作量。并且同时也可以增强系统的可迭代性，当某个模块不完善需要重构或进一步开发迭代时，可以不触及系统的整体，将模块升级即可。

## 编码设计思想

本系统实现对在线商品评价数据的爬取、清理、分析都应部署在服务器后台实现，前台只实现商品URL地址的传递和评论分析结果的图表显示。在数据发送和接收后，都应及时验证数据的有效性和合法性，减少系统不必要的性能消耗。

对于系统的实现应在编码开始前做好足够的准备，如熟悉Django框架的各个部件，熟悉Python代码编写语法和规范，了解文本情感倾向分析的原理和编码过程，防止系统编写过程中，编码的凌乱和不顺畅，增大无谓的代码量。

要善于利用开源社区中的资源，在功能完善的情况下使代码尽量的简洁明了。要善于使用网络上的资源，不应浪费时间去编写轮子。

# 运行结果

程序目前运行稳定，无明显bug，接下来进行运行结果的展示。系统分为用户界面和管理员界面。用户界面进行在线商品的URL地址输入和展示，管理员界面进行数据库的查看和相关系统维护。

## 用户界面

用户输入淘宝商品的URL并点击提交后如果系统之前进行过相应的分析，则会对该商品的分析结果以图表的形式展示，用户如果输入多个URL并以回车进行分割，则会并列展示多个商品的分析图表。分别展示商品的好评度分布图、整体评价分布图、好、中、差评分布图以及历史价格折线图。

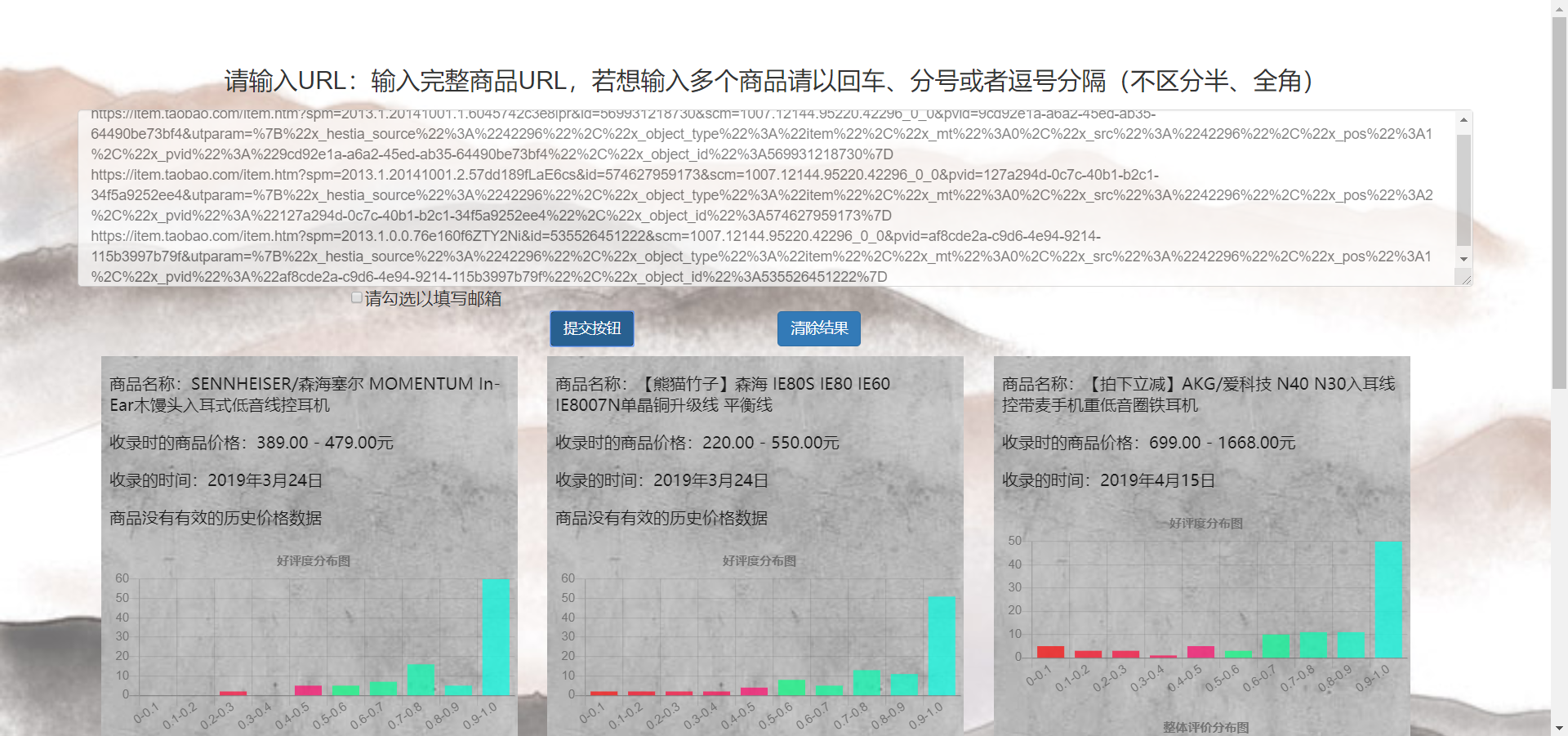


图 7‑1结果查看1

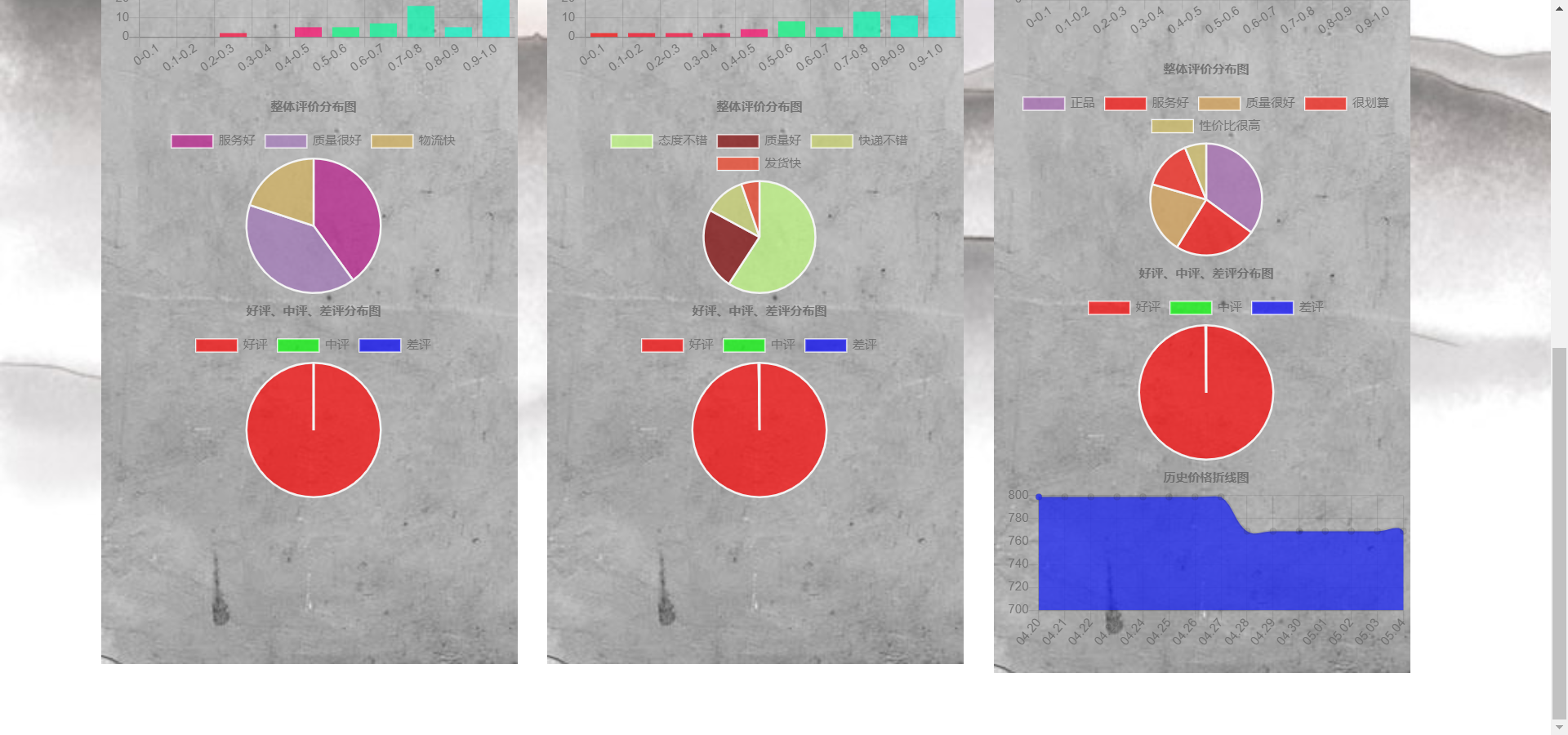


图 7‑2结果查看2

通过清除按钮可以清除页面上已显示的图表数据，如果系统之前未保存过该商品的数据，将会将在后台进行爬取、分析该商品的信息。



图 7‑3发送新商品

在输入新添加的商品URL地址时可以勾选输入邮箱的选框，输入自己的邮箱地址。



图 7‑4勾选邮箱发送

在商品评论数据分析完成后，将会收到提示邮件。

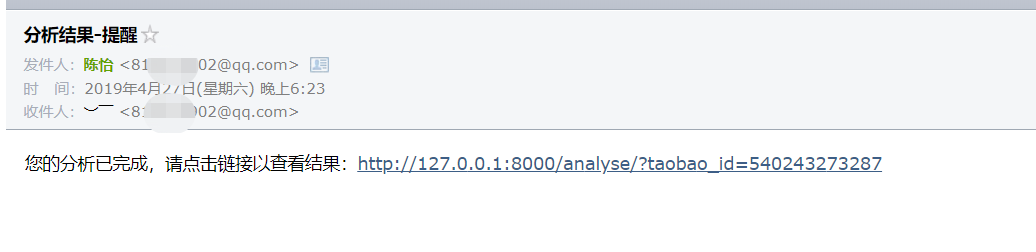


图 7‑5提醒邮件

通过点击邮件中的url连接可以查看分析结果。

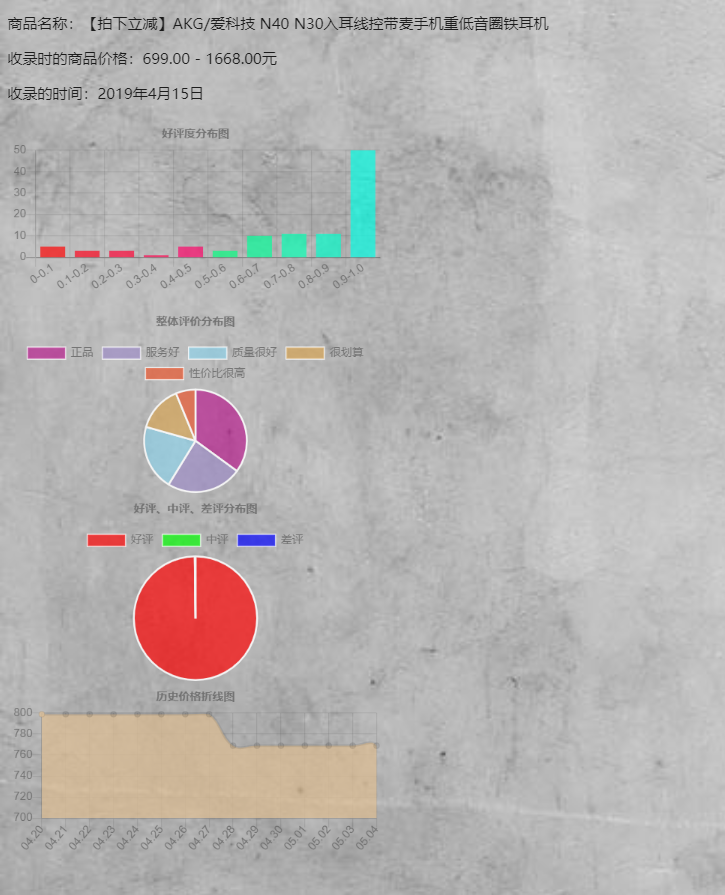


图 7‑6查看分析结果

## 管理员界面

管理员界面是基于django的admin界面上进行设置的，当前提供了对数据库的查看和管理功能。通过登陆/admin页面后可以查看数据库的信息和对管理人员的创建和授权。

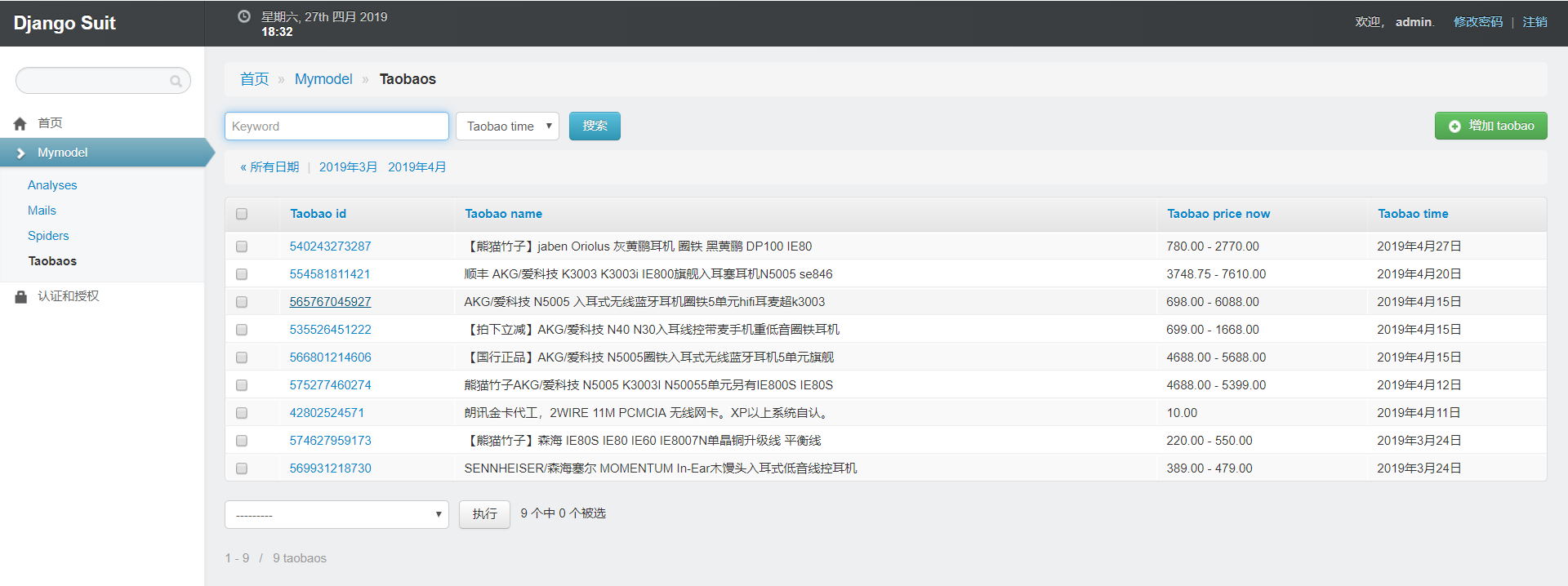


图 7‑7后台查看数据库

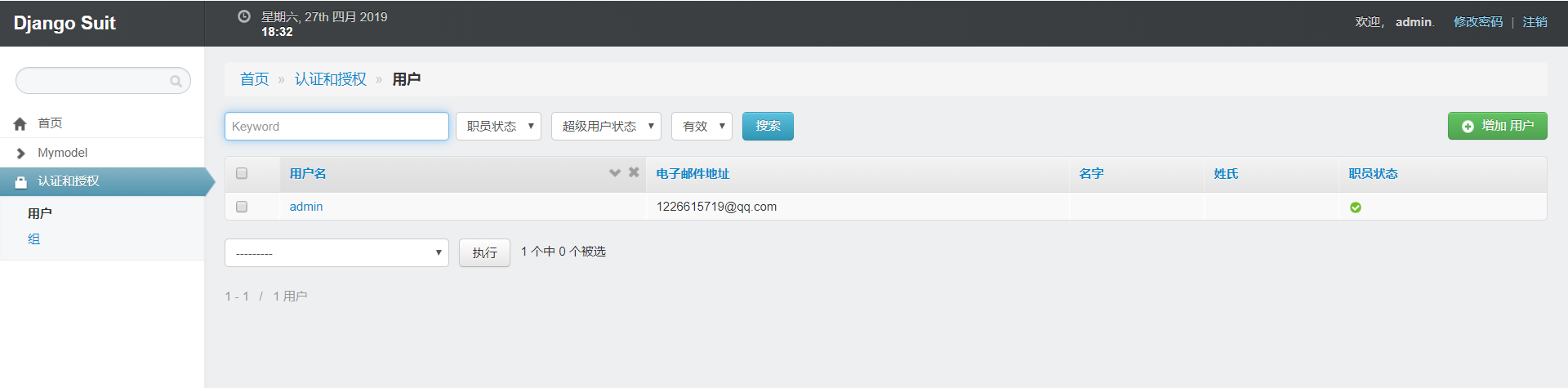


图 7‑8后台管理员配置

# 总结

我完成的毕业设计是在线商品评论分析系统，系统基本完成既定目标，已完善系统的相关功能和bug的修补，使用上用户体验良好，系统运行流畅。从系统的设计到完成的过程中，我体悟良多，不管是知识的获取能力、应用系统的设计能力、程序代码的编写能力还是项目进度和需求管理能力都感到受益颇深。特别是在文本情感倾向分析模块的设计和编写过程中，对于机器学习、自然语言处理都有了初步的理解和练习。在这次毕业设计的过程中，通过应用系统程序的设计和实现，使我在编程设计的理论和实践上都有了一个提高，尤其是增进了自己分析问题、研究问题、解决问题、进度管理等方面的能力。

通过大学期间的理论学习、每学期的课程设计以及这次的毕业设计的完成，充分的锻炼了自己的理论和实践，这其中除了自己的刻苦学习和认真专研外，也要归功于各位授课教师和指导教师的辛勤教育。感谢各位老师付出的努力和期望。

随着毕业设计的完成，我的大学生涯也面临终点。通过大学四年的学习，我不仅学到了学科里应该掌握的知识而且学会了为人处事的原则和方法。四年大学生涯将会一个美好回忆，而我在大学期间获得的点点滴滴也将受益终身。

# 参考文献

[1]于娟,刘强.主题网络爬虫研究综述[J].计算机工程与科学,2015,37(02):231-237.

[2]魏韡,向阳,陈千.中文文本情感分析综述[J].计算机应用,2011,31(12):3321-3323.

[3]王洪伟,郑丽娟,尹裴,史伟.在线评论的情感极性分类研究综述[J].情报科学,2012,30(08):1263-1271+1276.

[4]尹裴,王洪伟.面向产品特征的中文在线评论情感分类:以本体建模为方法[J].系统管理学报,2016,25(01):103-114.

[5]陈旻,朱凡微,吴明晖,应晶.观点挖掘综述[J].浙江大学学报(工学版),2014,48(08):1461-1472.

[6][美]韦斯·麦金尼（Wes McKinney）.利用Python进行数据分析(原书第2版) 机械工业出版社.

[7]Kumar Ravi,Vadlamani Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications[J]. Knowledge-Based Systems,2015,89.

[8] Steven Bird,Ewan Klein,Edward Loper.Natural Language Processing with Python, 2009-06-01

[9] [Nitin Hardeniya](https://www.google.ae/search?hl=zh-CN&tbo=p&tbm=bks&q=inauthor:%22Nitin+Hardeniya%22), [Jacob Perkins](https://www.google.ae/search?hl=zh-CN&tbo=p&tbm=bks&q=inauthor:%22Jacob+Perkins%22), [Deepti Chopra](https://www.google.ae/search?hl=zh-CN&tbo=p&tbm=bks&q=inauthor:%22Deepti+Chopra%22), [Nisheeth Joshi](https://www.google.ae/search?hl=zh-CN&tbo=p&tbm=bks&q=inauthor:%22Nisheeth+Joshi%22), [Iti Mathur](https://www.google.ae/search?hl=zh-CN&tbo=p&tbm=bks&q=inauthor:%22Iti+Mathur%22). Natural Language Processing: Python and NLTK, 2016-11-22

# 致谢

本设计是在指导教师黄雷君老师的亲切关心和细心指定下完成的。黄雷君老师从设计方案的选定、设计计划的安排、设计编码疑难问题的解惑、论文的书写都给予了精心的指导和严格的要求。对我的系统设计开发和论文的结搞都给予了莫大的帮助和关心。在此，我向黄雷君老师表示由衷的感谢和诚挚的谢意。

同时也感谢我的同学和朋友提供的大力支持和帮助。

最后向评审本论文，参加论文答辩的各位老师表示最衷心的感谢。