

**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**

**年 月 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 指导教师：黄雷君

**摘要**:随着电子商务的快速发展，在线购物已经是我们生活中经常会见到的现象，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。但是网购平台的评价过于繁多杂乱，并且其中还参杂着部分水军，导致我们人工分析较为困难。

在线商品用户评论分析系统是将爬虫技术和自然语言处理技术结合起来实现的在线商品评论分析系统。系统的主要功能是使用者输入淘宝商品的URL链接，系统将会自动抓取商品的评论数据，获取每条评论的情感倾向度，通过图表进行展示，向用户直观地展示商品的评论情况和基本信息。系统首先通过截取、分析得到的接口，通过该接口来获取淘宝商品的评论。之后通过百度Al平台进行评论的情感倾向分析。系统内置了多种分析引擎，包括基于情感字典、基于机器学习以及基于百度Al的多种分析方式。通过综合分析，目前采取了百度Al的引擎进行评论分析。通过相应的图表展示商品的信息，给使用者一个良好的使用体验。通过多次使用的结果表明，该系统具有操作简便、易扩展等优点。

**关键词**：淘宝；评论；爬虫；自然语言处理；情感倾向分析；百度Al

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

**Abstract:** With the rapid development of e-commerce, online shopping has become a phenomenon we often see in our lives. In online shopping, we often analyze the pros and cons of products through the evaluation of others' products. However, the evaluation of the online shopping platform is too numerous and complicated, and it is also mixed with some water forces, which makes our manual analysis more difficult.

The online product user comment analysis system is an online product comment analysis system that combines crawler technology and natural language processing technology. The main function of the system is that the user inputs the URL link of the Taobao product. The system will automatically grab the comment data of the product, obtain the sentiment orientation of each comment, display it through the chart, and visually display the product review and basics to the user. information. The system first obtains the comments of Taobao products through the interface obtained by intercepting and analyzing. After that, the sentiment orientation analysis of the comments was conducted through the Baidu Al platform. A variety of analysis engines are built into the system, including emotion-based dictionaries, machine-based learning, and multiple analysis methods based on Baidu Al. Through comprehensive analysis, Baidu Al's engine is currently used for comment analysis. Display the information of the product through the corresponding chart to give the user a good experience. The results of repeated use show that the system has the advantages of simple operation and easy expansion.

**Keywords:** Taobao,comments,reptiles,natural language processing, sentiment orientation analysis, Baidu Al

目 录

[**本 科 生 毕 业 设 计（论文）** 1](#_Toc6924453)

[**本科生毕业设计（论文）诚信承诺书** 2](#_Toc6924454)

[1 绪论 1](#_Toc6924455)

[1.1 开发背景 1](#_Toc6924456)

[1.2 设计目的 1](#_Toc6924457)

[2 可行性研究 2](#_Toc6924458)

[2.1 可行性分析 2](#_Toc6924459)

[2.1.1 技术可行性 2](#_Toc6924460)

[2.1.2 经济可行性 2](#_Toc6924461)

[2.1.3 操作可行性 2](#_Toc6924462)

[2.1.4 法律可行性 2](#_Toc6924463)

[3 需求分析 3](#_Toc6924464)

[3.1 系统需求目标 3](#_Toc6924465)

[3.2 系统功能分析 3](#_Toc6924466)

[4 总体设计 4](#_Toc6924467)

[4.1 系统流程图 4](#_Toc6924468)

[4.2 数据流程图 5](#_Toc6924469)

[4.3 数据库设计 5](#_Toc6924470)

[4.3.1 商品信息表 6](#_Toc6924471)

[4.3.2 商品评论表 6](#_Toc6924472)

[4.3.3 评论分析表 6](#_Toc6924473)

[4.3.4 邮箱地址表 7](#_Toc6924474)

[4.4 系统模块设计 7](#_Toc6924475)

[4.4.1 管理模块设计 7](#_Toc6924476)

[4.4.2 图表显示模块设计 8](#_Toc6924477)

[4.4.3 爬虫模块设计 8](#_Toc6924478)

[4.4.4 分析模块设计 8](#_Toc6924479)

[4.4.5 交互模块设计 8](#_Toc6924480)

[4.4.6 邮件发送模块设计 9](#_Toc6924481)

[5 详细设计 10](#_Toc6924482)

[5.1 管理模块 10](#_Toc6924483)

[5.2 图表显示模块 11](#_Toc6924484)

[5.3 爬虫模块 12](#_Toc6924485)

[5.4 分析模块 15](#_Toc6924486)

[5.5 交互模块 16](#_Toc6924487)

[5.6 邮件发送模块 17](#_Toc6924488)

[6 编码设计 19](#_Toc6924489)

[6.1 编码设计风格 19](#_Toc6924490)

[6.2 编码设计思想 19](#_Toc6924491)

[7 总结 20](#_Toc6924492)

[8 致谢 21](#_Toc6924493)

[9 参考文献 22](#_Toc6924494)

[10 附录1英文原稿 23](#_Toc6924495)

[11 附录2英文翻译 29](#_Toc6924496)

# 绪论

## 开发背景

这是一个互联网的时代，随着网络的普及和电脑技术的发展，人们的生活越来越离不开互联网，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。

随着我国互联网普及率和人民生活水平的提高，网上购物已成为很多人愿意选择的购物方式。网络购物已经深入我们生活中的方方面面，在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，我们可以足不出户就购买到物美价廉的商品。据调查显示，近几年来网购所占购物数额的比例逐年大幅度提升。而与此同时，电子商务平台上商品的评论数量也呈几何式上升，使得想要购买商品的顾客很难在纷繁的评论中提取到自己所需要的信息。即使能得到信息也过于片面，不能系统合理评测商品。站在用户的角度上开发本系统，目的是希望能降低顾客网络购物风险。本系统主要采用爬虫技术，文本情感分析中的一些算法和一些当前比较先进的开源框架和Al平台来实现，用户可参考系统给出的分析结果从而选择是否购买该产品。

## 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，往往只会看到其中几页或者几条评论，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

## 可行性分析

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，以下将从几个方面进行分析本系统的可行性。

### 技术可行性

系统使用django作为开发基准框架。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，python具有丰富的第三方开发库和强大的社区论坛。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数据报文发送和抓取。对于数据挖掘，python具有jieba第三方开发库可以进行中文文本的分词，jieba是目前python最好的中文分词库。NLTK第三方开发库，是在NLP领域中最常使用的一个python库，SnowNLP，可以方便的处理中文文本内容分析和挖掘。更进一步，各大IT公司具有更多更强大的Al平台，如百度Al开放平台、腾讯Al开放平台等众多开放平台可以使用，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。本系统数据库使用Mysql数据库引擎，Mysql具有简单易用，对于个人开发者免费，小巧，性能较强，可移植性高、运行速度较快等优点，Mysql支持部署在Linux平台，可减少服务器的开支和性能占用。

### 经济可行性

本系统均采用开源软件设计开发，其中数据库使用Mysql数据库，对于个人开发者免费使用，开发语言和系统框架使用python和django，数据库部署平台Linux，Web应用部署平台Windows，开发软件使用Pycharm教育版，均为个人免费产品。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。并且系统的后台管理界面可以查看数据库信息，可以完成日常的过期数据清理和异常数据清理。操作方面可行。

### 法律可行性

本系统的设计和使用没有违反国家相关法律规定，法律方面可行。

# 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几个需求目标：

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。用python成熟框架编程，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：采用开放的接口和标准，便于系统向更大的规模和功能扩展。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

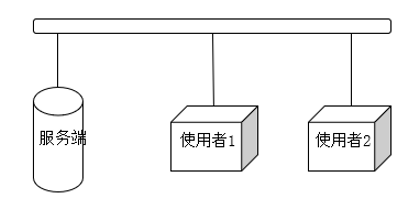
软件要求不断完善，功能也是逐步添加的，以下为该在线商品用户评论分析系统的基本功能需求：

* 用户在搜索栏输入需要进行评论分析的商品链接，系统将商品链接返回后台处理。
* 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。
* 设置后台管理账户，使管理员可以通过后台管理页面操作数据库和系统。

# 总体设计

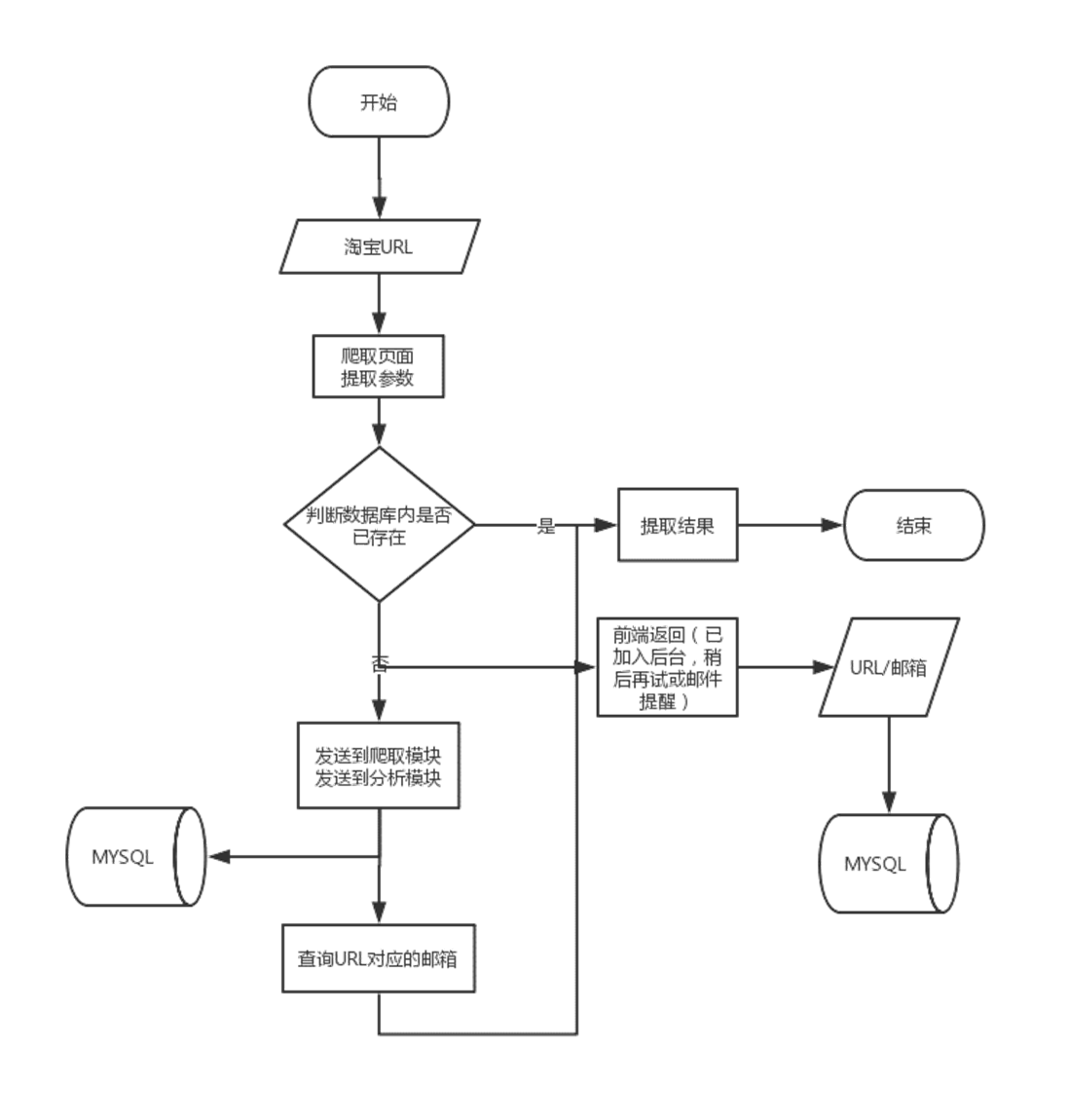
## 系统流程图

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：



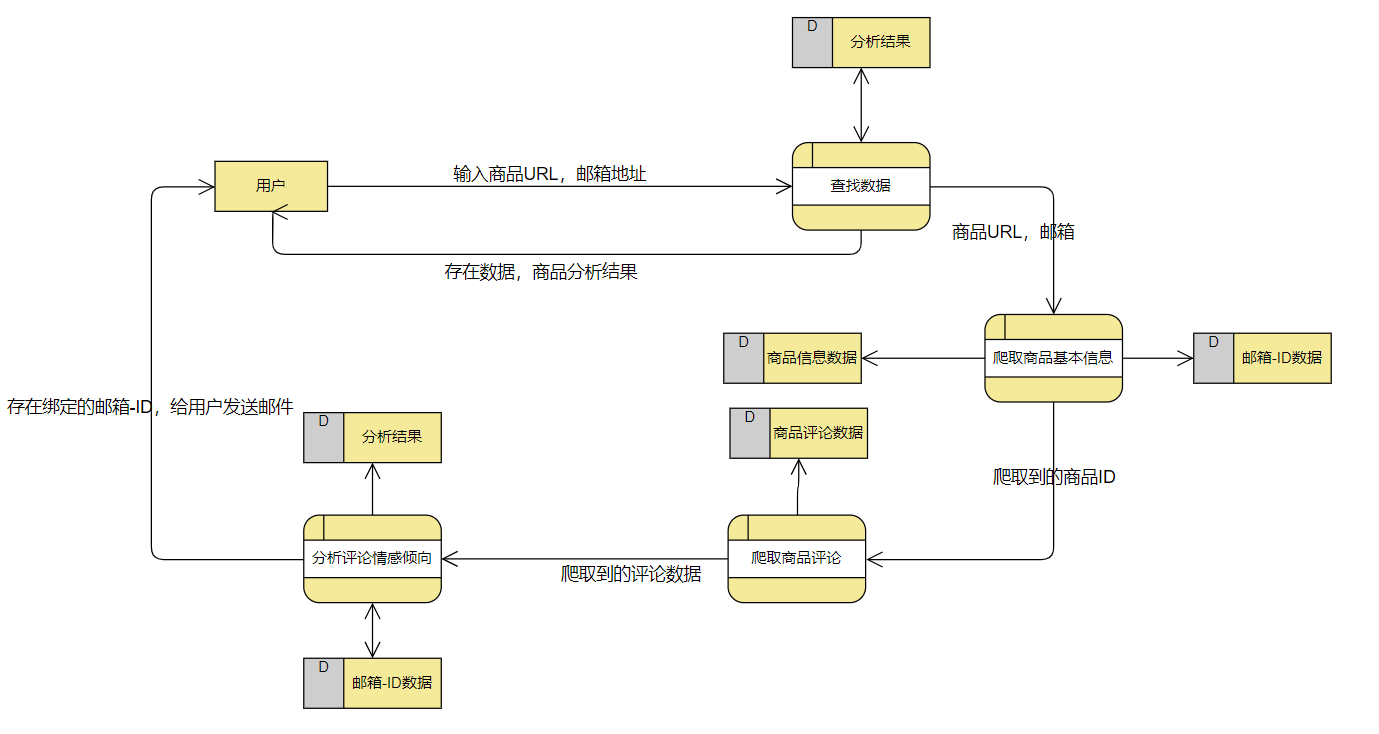
系统对于操作成功和失败，异常错误都应存在相应的提示，以便用来帮助用户使用操作系统，同时也可以帮助开发者和管理员更好的开发、维护系统。

用户通过web浏览器进行相关操作，系统的流程图如下：



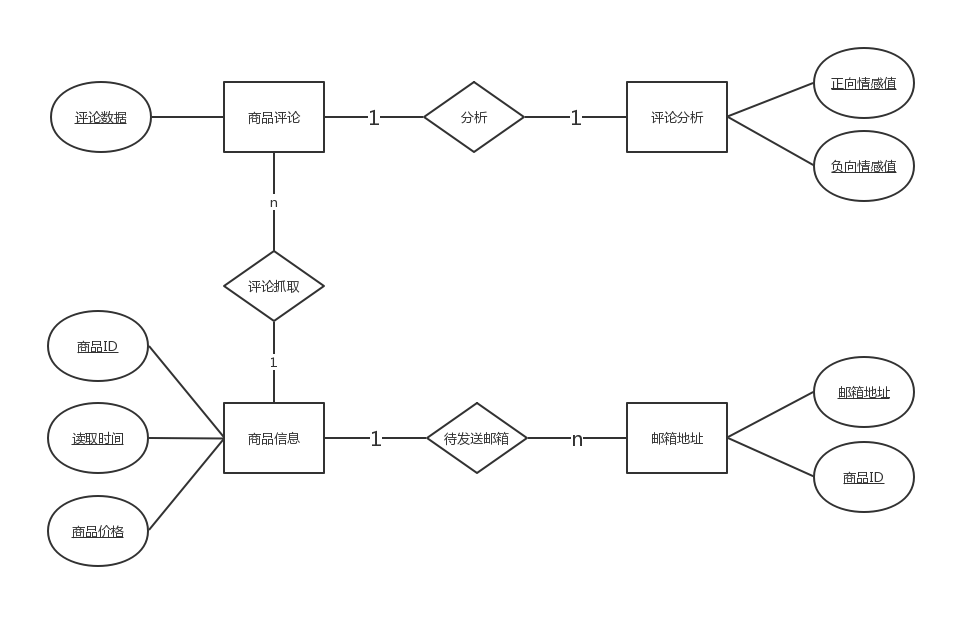
## 数据流程图

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：



## 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，数据库的E-R图如下图所示：



### 商品信息表

商品信息表用于保存商品的相关信息，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，下图为商品信息表的表结构设计：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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

商品评论表用于保存商品评论的相关信息，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，下图为商品评论表的表结构设计：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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

评论分析表用于保存评论的分析结果的数据，包括商品ID，评论正向情感倾向值，评论负向情感倾向值，保存时间等字段，表名为：MyModel\_analyse，下图为评论分析表的表结构设计：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

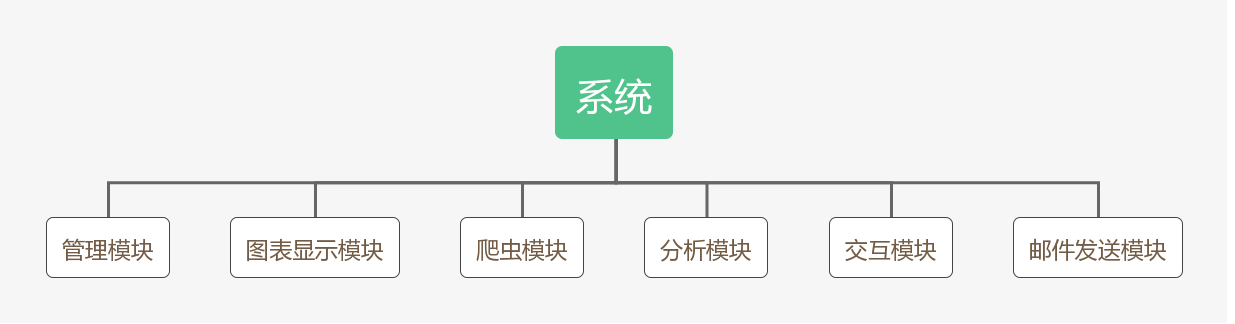
### 邮箱地址表

邮箱地址表用于保存待发送的邮箱的数据，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，下图为邮箱地址表的表结构设计：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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、邮件发送模块这六个主要模块，下图为系统功能模块图：



### 管理模块设计

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 | |

### 图表显示模块设计

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 | |

### 爬虫模块设计

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 | |

### 分析模块设计

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 | |

### 交互模块设计

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。用户可以填写邮箱，将分析结果通过邮件发送，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 | |

### 邮件发送模块设计

通过设置好的邮箱地址和密码，以此作为发件人，向需要发送提醒的邮箱发送提醒邮件，提醒用户评论分析已完成，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 | |

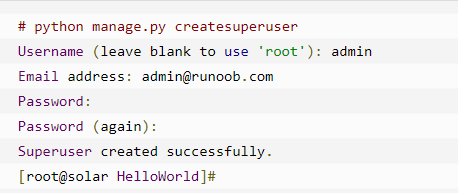
# 详细设计

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。

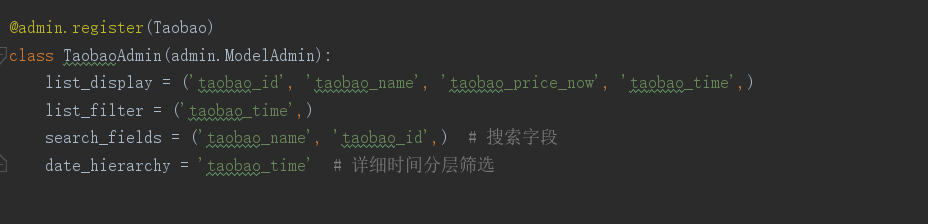
## 管理模块

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。

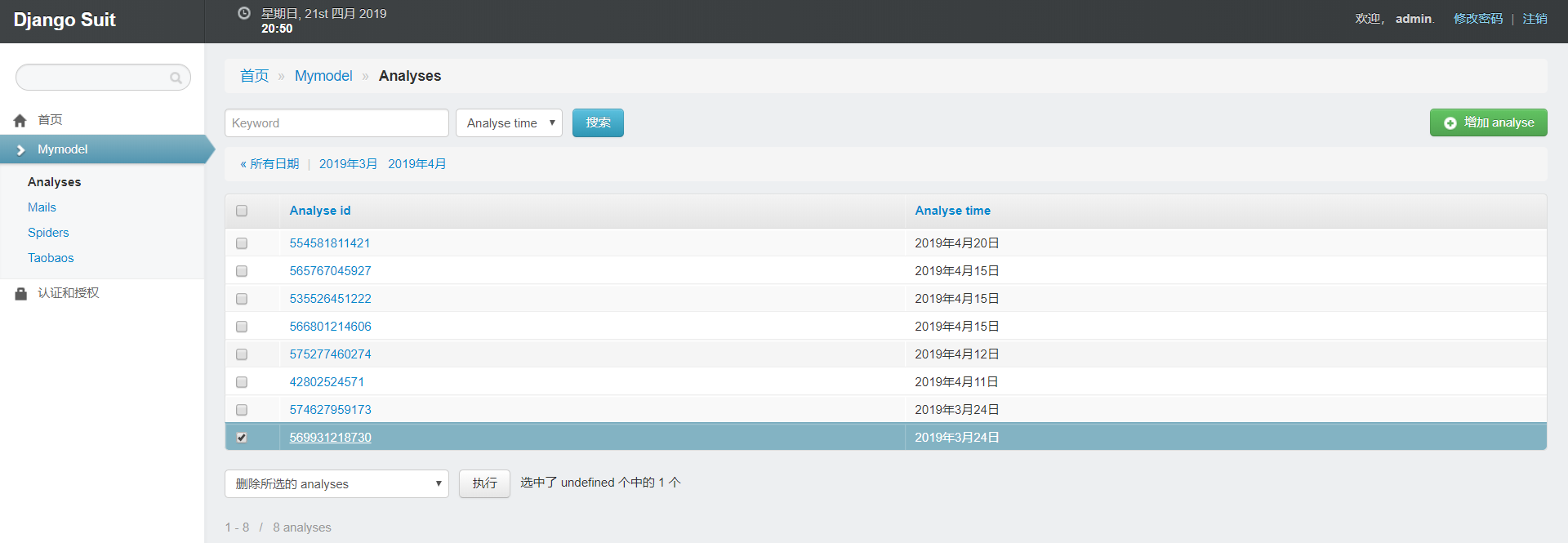
django提供了基于web的后台管理工具，可以通过django的自动后台管理工具快速地生成一个管理页面。通过激活管理工具和创建超级用户即可使用：



在admin.py文件中注册模型，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：



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：



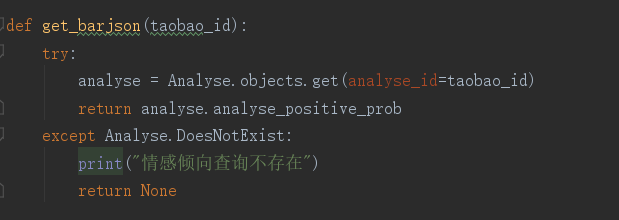
## 图表显示模块

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。

这里图表使用了Chart.js下的图表显示插件，方便地进行柱状图、饼状图、折线图等的生成。

首先需要通过商品ID从数据库中读取数据。



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。



最后将数据返回到前端界面，通过Chart.js生成相应的图表。

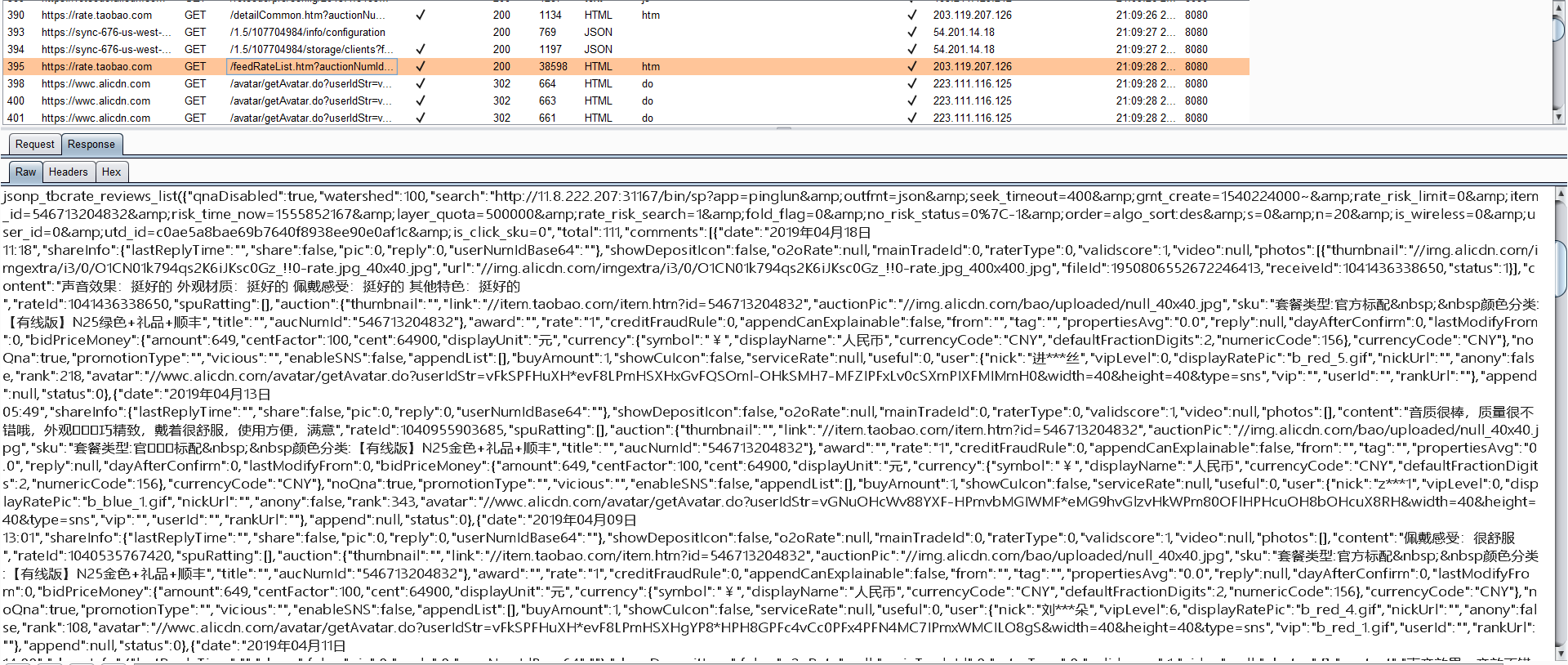


## 爬虫模块

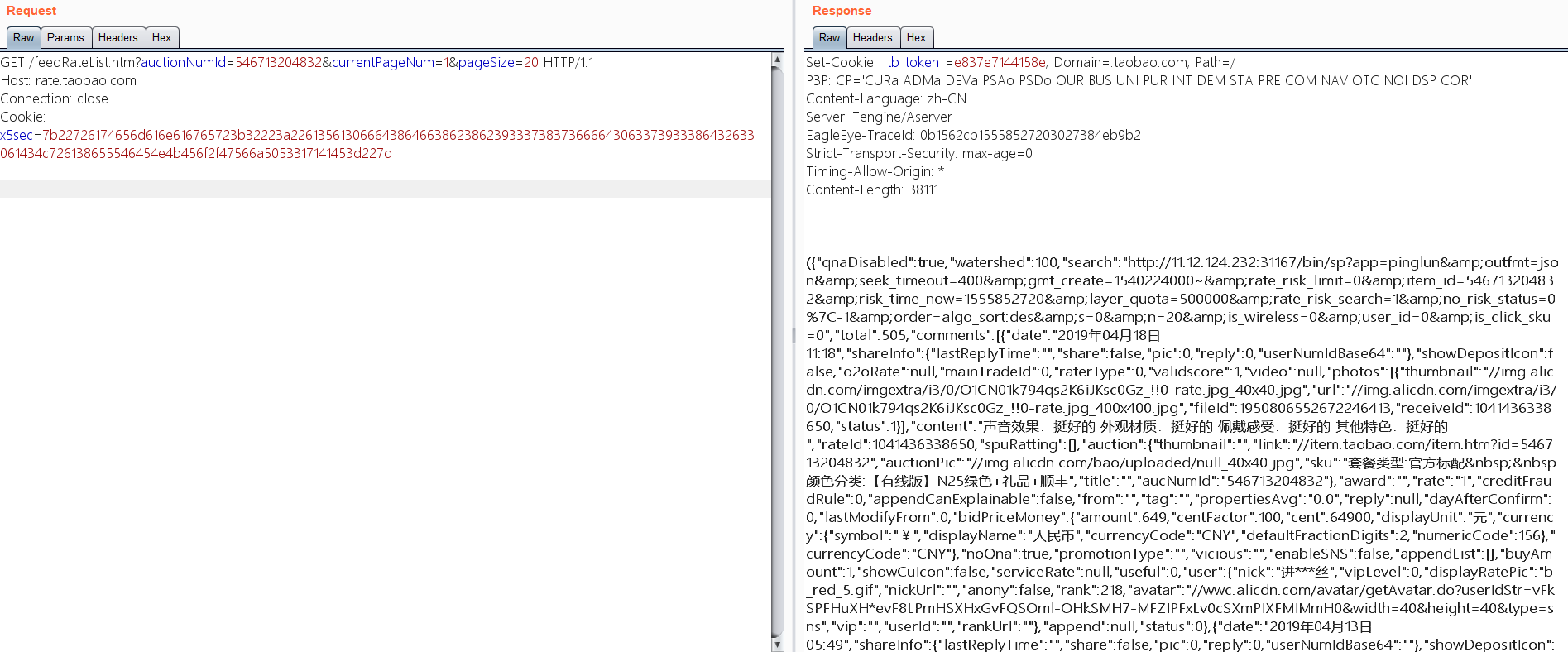
将商品链接发送到爬虫引擎，爬取商品链接下的评论，存入数据库等待分析模块运行。适当使用IP代理池和反爬技术，设置合理的爬取速度防止IP、账户被封。

爬虫模块需要将商品的评论和总体评价数据进行爬取，常规做法是使用scrapy进行爬取数据，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。

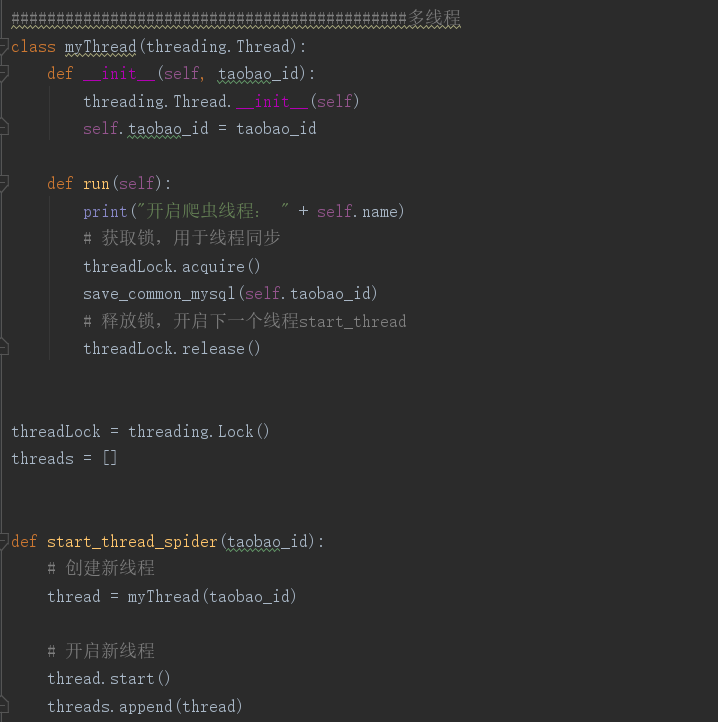
这里通过使用burp suite和火狐浏览器进行配合查看在浏览淘宝商品的过程中，会产生哪些请求，首先要对火狐浏览器设置代理和导入证书，之后开启burp suite的监听功能，浏览一个淘宝商品的评论后，再查看burp suite抓取到的请求，这里我们可以看到https://rate.taobao.com/feedRateList.htm请求是淘宝返回商品评价的JSONP请求



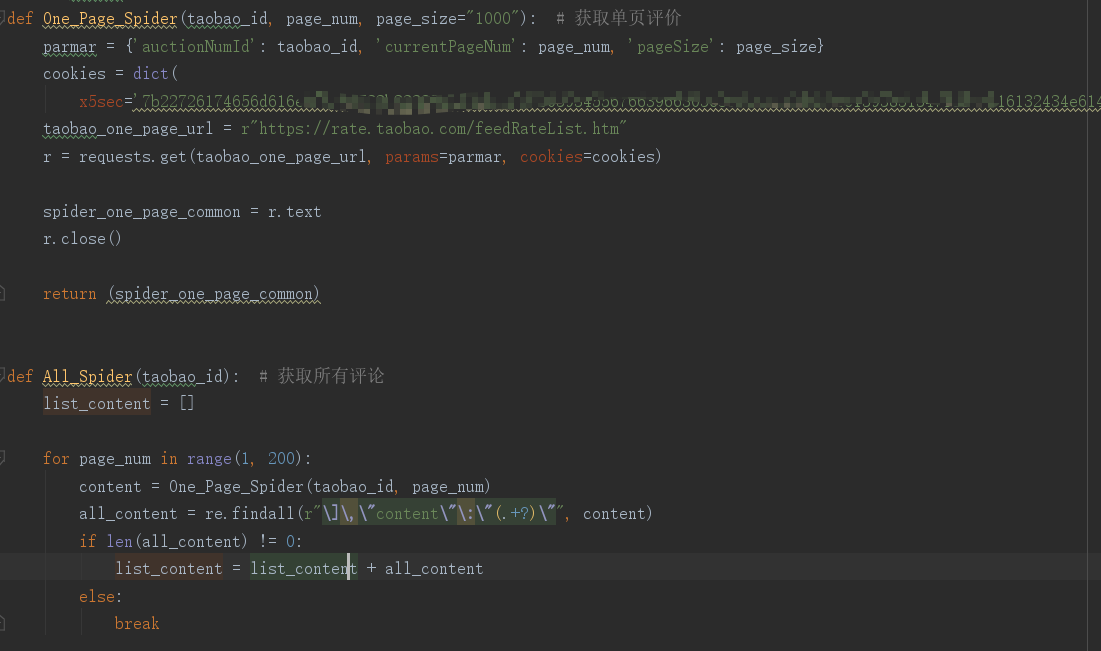
通过逐步删除请求中的相关参数，最后分析得到GET参数中的auctionNumId控制商品的ID，控制JSONP返回的评论数据是哪一个商品的，currentPageNum控制商品评论的页数，控制返回的评论数据是第几页的数据，pageSize控制商品评论的页面大小，控制返回的评论数据的多少，Cookie参数中的x5sec控制身份认证，在多次请求后，如果这个参数不存在或为无效参数，将会跳转到验证码机制的页面，但是通过分析发现，这个x5sec参数的数值不会过期，长时间或退出登陆任然有效，所以可以通过获取一个x5sec参数的数值并使用这个数值长期的获取商品评论的JSON数据。



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



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，调用分析模块进行评论的情感倾向值分析。



## 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数据进行分析，这里设置了四种分析方式，并可以独立使用。

情感倾向值分析是自然语言处理中的一个研究方向，分析一个文本情感倾向值，可以知道这个文本的倾向值是正向的还是负向的，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，通过一定数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SnowNLP的分析模式

SnowNLP是目前比较推荐的中文情感分析库，通过SnowNLP可以方便快捷地完成分类器和情感分析的工作。首先通过收集正负向的训练评论数据，通过SnowNLP训练分类器并进行保存。



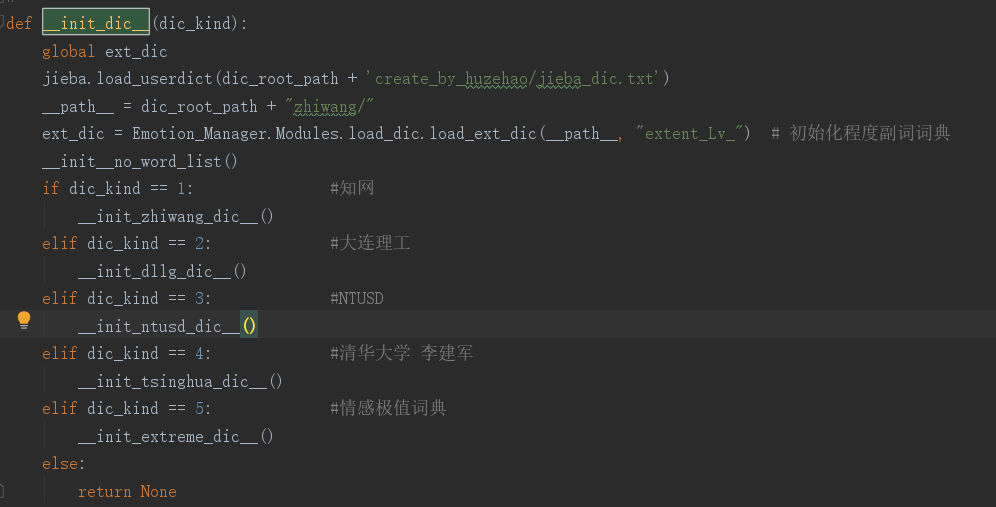
通过训练好的模型进行文本的情感倾向分析，获得正向和负向的情感倾向值



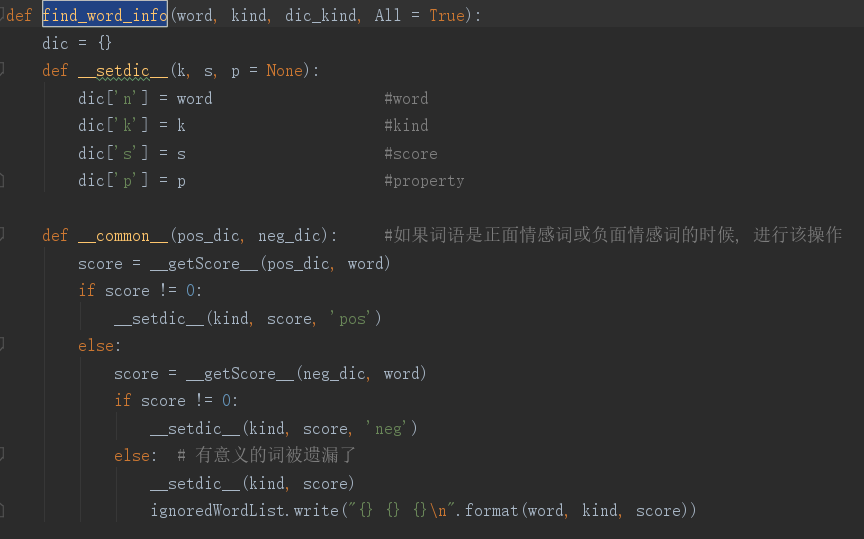
### 基于情感字典的分析模式

情感字典是一种比较通用的分析模式，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。使用情感字典进行分析的基本流程为：

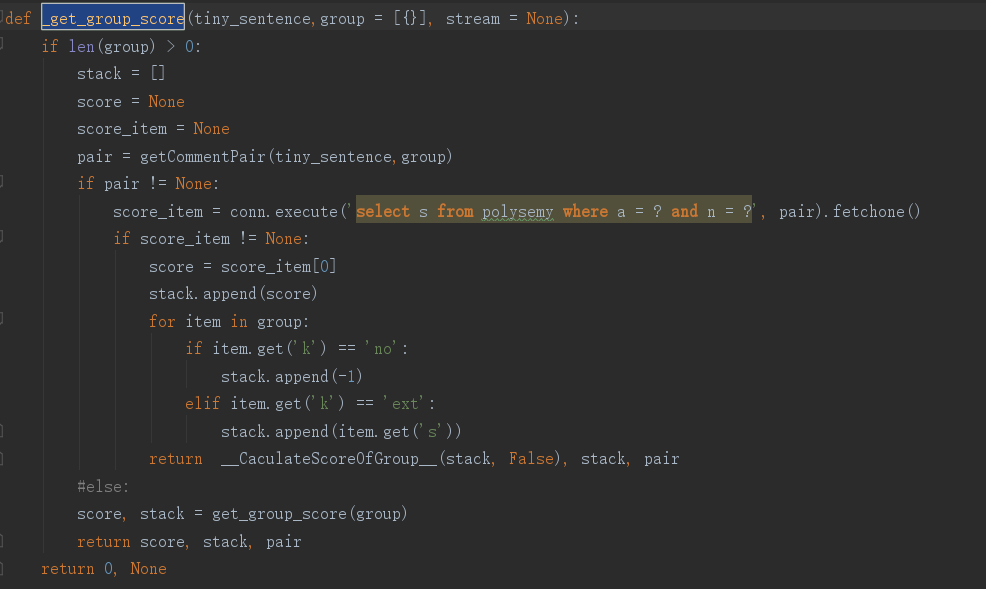
1. 加载情感字典



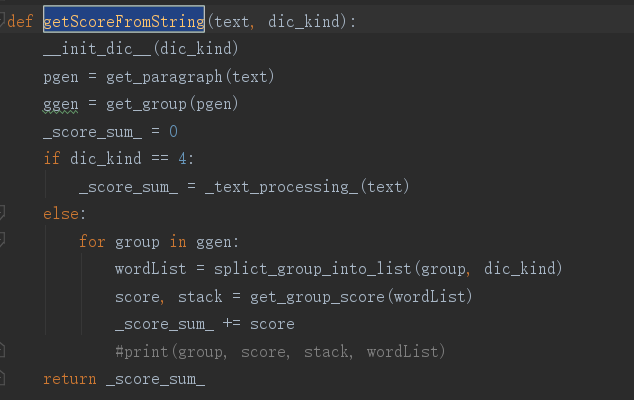
1. 得出词语详细信息



1. 计算短句子情感值

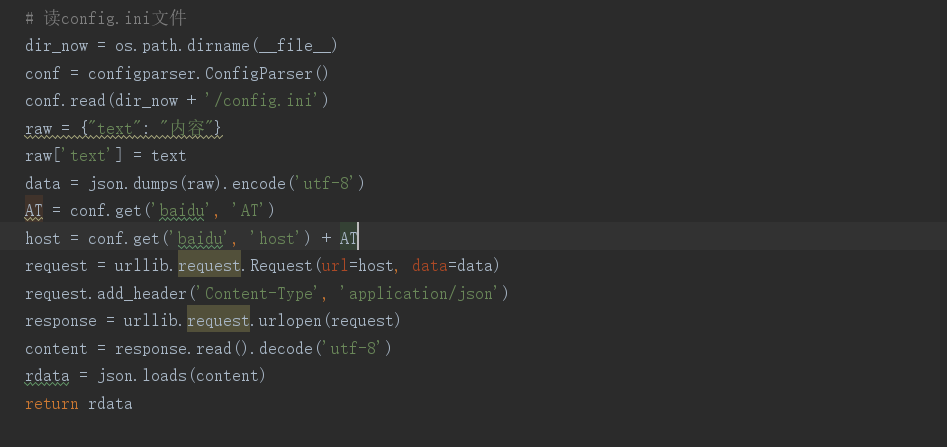


1. 获取整个中文文本的情感值



### 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，缩短对于情感倾向分析的开发的时间。首先在百度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。

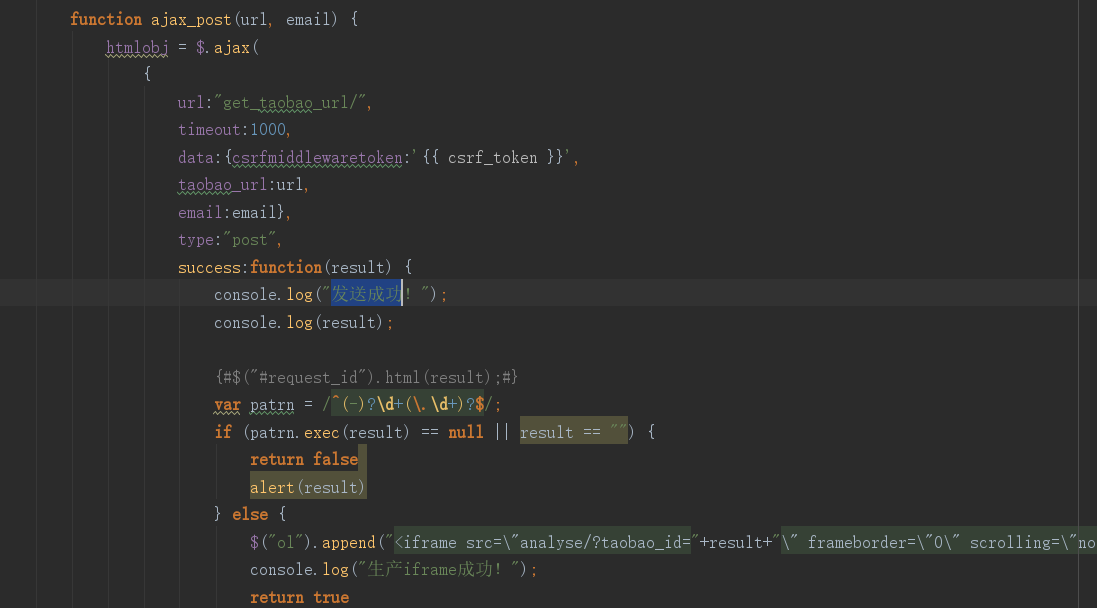


## 交互模块

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。用户可以填写邮箱，将分析结果通过邮件发送，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，这里使用AJAX技术进行数据的交互，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



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



## 邮件发送模块

通过设置好的邮箱地址和密码，以此作为发件人，向需要发送提醒的邮箱发送提醒邮件，提醒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产生拼接后的查看分析结果的图表显示URL地址，将URL地址通过邮件发送给用户，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。



# 编码设计

## 编码设计风格

本系统编码设计风格应遵循以下四条基本规范：

1. 对于系统中的变量名、函数名、数据库表名、字段名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. 程序编码书写具有缩进，复杂过程的实现都应在代码中有分步注释，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。这样的好处是可以清晰地体现代码的层次，便于进行代码的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，如对Django系统框架的熟悉，对Python代码编写语法的熟悉，对文本情感倾向分析的了解和编码过程等都应有相应的了解和Demo编写，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。

# 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修改，用户体验良好，系统运行流畅。在从设计直到完成的这段时间内，我感觉学习到了很多东西，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。在完成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较大幅度的提高，尤其是培养了自己独立分析问题、解决问题、研究问题、自我管理的能力。

本系统其中一个模块采用了自然语言处理和机器学习这两项技术，在现在一个人工智能和大数据的时代背景下，自然语言处理和机器学习将会是一个非常重要的技术领域。特别对于现在政企极其关心的网络安全领域和目前非常火热的自动驾驶领域将会是一个重要的工具，对于安全领域中的态势感知、动态防御，自动驾驶领域中的智能行车、危险躲避都将十分依赖于人工智能、自然语言处理和机器学习这几个学科知识的发展。多学习了解这些领域的知识将会对未来的工作学习有一个更好的助力。

通过大学期间理论学习和平时的课设实践以及这次的毕业设计的实践，使得我在理论和实际上有了很大的提高，除了自己的刻苦学习和认真专研外，这些成绩的取得和指导老师和各位授课的老师的辛勤教育是分不开的。感谢各位老师付出的努力和期望。

随着毕业设计的结束，我的大学生活也即将面临终点。回顾这四年的大学学习生涯，我不仅学到了学科里的知识而且学会了为人处事的原则和方法。四年大学生涯将会是我一生的美好回忆，而我在大学期间学习的点点滴滴也将受益终身。

# 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。在应用系统的开发和调试中给予了我很大的支持与帮助。这个系统的设计和论文的完成，凝结着黄雷君老师的心血和汗水。老师开拓的工作作风、科学的思维和严谨的治学态度都使我受益颇深。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。

在系统程序的设计和实现过程中，也得到了Python社区各位网友的帮助，特别是应用系统开发工程中遇到的技术性难题时候，各位社区的朋友及时答复我发出的求助帖。进行相关的疑难解答。对我问题的顺利解决提供了很大的帮助。

同时也要感谢实习公司的PM和各位同事，在我进行毕业设计的应用系统设计和开发过程中的帮助和关心，及时的减少工作量，使我有更多的时间投入到毕业设计的相关工作中。

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。

# 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

# 附录1英文原稿

china region

**Video Consumption, Social Networking, and Influence**

REVENUE FROM CHINA’S online entertainment market reached approximately $200 billion this year.a It is not surprising that China’s video market is comparable to the U.S.;b,c in fact the number of online video users in China is 2.5 times more than that of the U.S. (that is, 212 million U.S.-based usersd compared to 579 million users in China).e Due to advancements in broadband and mobile technology, online video is the fastest growing area for China’s Internet, with a growth of around 50% over the past five years.

The landscape of China’s online video industry has as many similarities as differences with the U.S., presenting extremely interesting observations and insights. This article provides an overview of the market, dominant players, and business models, as well as presents intriguing product nuances and technical advances in this area. Like the U.S., there are two major categories of services for online video: the head and the tail. The heads are the premium players that stream copyrighted shows and movies. They are the Netflixes and Hulus of China. The tails present professionally generated content (PGC) and user-generated content (UGC) for different market segments. They are the YouTubes and Snapchats of China.

The Head: Premium Video Platforms

The top three players in premium online video are iQiyi, Tencent Video, and Youku. These companies are affiliated with Baidu, Tencent, and Alibaba, respectively. Statistics from QuestMobile, China’s big data services provider, shows that both Tencent Video and iQiyi recorded around 500 million monthly active users by the end of 2017, and around 300 million for Youku. The iQiyi video platform, that went public on NASDAQ last March, also leads in total watch hours.

Unlike the premium video services in the U.S., the major players in China all started with free services supported by advertising. Total ad revenue for online video was approximately $10 billion, catching up to ad revenue generated by commercial television.

However, we have seen a huge take off in subscriptions over the past two years, when users started gravitating toward (and paying for) platforms that were ad-free and offered additional features such as access to higher-quality video and memberonly original content. iQiyi counted 60 million subscribers as of Feb.

Due to the prevalence and popularity of video streaming, both infrastructure and application companies in China have invested heavily in video technology.

2018, and Tencent Video has over 40 million. By comparison, Netflix had about 55 million U.S. subscribers and 63 million international subscribers as of January 2018.

Payment practices have been forming rapidly among the middleclass and young Internet users in China over the past few years. The major driving forces behind this movement include a concerted crackdown on pirated content, affordable prices, and, most importantly, the ease of online payment.iThe monthly subscription price is between 20 RMB ($3.16) to 40 RMB ($6.32) for each of the three services. The total market size of Internet video subscription services has increased dramatically, from about $63 million in 2012, to $2.1 billion in 2016, and an estimated $11.5 billion in 2022.

Unlike premium content distributions in the U.S., many of the TV dramas and movies are non-exclusive, available on all three major services and elsewhere once they aired on TV or in movie theaters. Exclusive content is usually far more costly. All services invest heavily in copyrighted movies and TV shows, resulting in very high production prices for this content. Platforms often must make a calculated bet on what shows will prove popular, and make an offer before production. According to its prospectus, iQiyi’s annual content cost is about $1.9 billion, while the other two platforms spent nearly double that. These purchases are far beyond subscription fees and advertising income. According to its public filing, iQiyi lost $169 million Q1 2018,j and the other two services also lost similar sums.k This situation will continue for the next few years.

To reduce costs, and to stand out among their peers, all three premium services have started to make original content exclusive on their own platform.l The lead player for original shows is iQiyi, although all three had different hits. Unlike U.S. platforms where TV dramas tend to reign supreme, variety shows garner a greater audience in China. According to a recent Wall Street Journal article,m “The Rap of China,” a 12-episode hip-hop rap competition reality series created and shown by iQiyi targeting younger audience, has become “China’s most popular entertainment program in 2017.” The show attracted 2.7 billion views during its run from late June to early September. Short videos clips gleaned from the show were watched eight billion times on the social media platform Weibo.

Another fact about China’s premium online video services that differs from their U.S. counterparts is that they all participate in PCG and UGC short video markets. But, as we will discuss, other emerging players are increasingly dominating these segments.

The Tail: UGC and PGC Video Platforms

With more than 100 players and 400 million users in 2017, the short video landscape in China is hugely dynamic, and far from settling down. The user base is huge, fast growing, and extremely active. There are several major players, and most of them came into prominence over the last year or two.

Unlike the U.S. market, the initial dominant short video platforms were the premium players, like iQiyi and Youku. For example, YouKu claims to have invested approximately $1.6 billion in user-generated content since 2015.n They are modeled after YouTube and had a large number of viewers watching a mixture of premium and UGC content. Their short-form videos include movie clips and music videos as well as free-form user-created content, and they provide channels created for professional content producers.

However, in the past two years, we have witnessed the phenomenal growth of several mobile short video apps not associated with the premium players. Thanks to their ease of use, these apps became super popular and prevalent, penetrating a huge number of users in massive areas of China (see Figure 1). We saw a 311% increase in short video traffic in Q3 2017 compared to one year earlier.o Analysts labeled the China Internet era of 2017 as “the year of short videos.”

New popular short video services include Kuaishou, Huoshan, Xigua, Douyin, Miaopai, Meipai, Weishi, and many more. The most popular of them have monthly active users in the 100 million–200 million range. The app experience is mostly a flow of mobile video feeds, with videos running seconds to minutes long. The app provides good tools for users to shoot, edit, beautify, and add special effects to the videos. These video apps can be divided into two categories: those more like the “dubsmash” mobile app, where users can record their own video dubbing over music, and those more like Snapchat with free-form videos. All video apps provide strong discovery and follow functionalities, encouraging interaction and social connections among users.

More specifically, videos in Xigua are mainly PGC content and video clips running 1–5 minutes long. Content in both Huoshan and Kuaishou are short UGC videos of about 15 seconds. Many everyday users, especially those in the rural area and small cities, record short videos and share them via these apps.p The apps provide a method for users to present themselves, compete to gain fans and eyeballs, and eventually profit from the viewership. Kuashou, with daily active users of close to 100 million, is the current leader in this category. Using a slightly different format, Douyin has recently received considerable attention and gained a significant user base, claiming more than 150 million monthly active users.

Douyin, as well as competing apps like Weishi, allows users to create and share short music videos using provided templates.

Another large set of online video services focuses on end user livestreaming: they provide live-broadcasting capabilities for consumers, popular pop idols, and the general public. According to Pandaily, eight live-streaming platforms in China raised approximately $11.6 billion in the first half of 2018. The leading platforms—Huya Inc. and Douyu TV—account for nearly 70% of the total.

Just like all premium players provide UGC, all short-video players also working on live streaming. The situation is highly dynamic; the landscape was quite different several months ago, and it is likely to be very different a few months from now.

Still experimenting with business models, the new mobile video apps make money through advertising, affiliate marketing to e-commerce, gifting, and other creative methods. The top “VIPs” on these platforms may have millions of fans, and they profit by advertising or selling goods. Brands also start to create channels on these platforms. According to the China’s Short Video Industry Report from iResearch, the revenue from the short video arena reached $860 million in 2017, and is expected to exceed $4.5 billion by 2020.

We are just beginning to see the great potential of China’s short video market; it will evolve in the coming years. There is a trend of going vertical, with different players specialized in a domain or a demographic. With fierce competition within China, many of these short video apps made their ways overseas, in particular, to other Asian countries. For example, Kuaishou was the Number One video app in the Korean app store in November 2017. According to an April 2018 report from 36kr, almost half of the popular short video apps in Asia were “made in China.

Mobile and Social via Video

Large screens, including TV and other OTT television devices, are heavily regulated in China. Due to the high penetration of mobile devices, the usage of online video tilts strongly toward mobile. Even for the premium users, the dominant media preference is mobile and personal computers. According to iQiyi, it has on average 421 million monthly active users (MAU) on mobile and 424 million MAU on PCs in Q4 2017, while TV and OTT devices are negligible. Many short videos applications are designed for and used exclusively on mobile devices.

Analysts have labeled the China Internet era of 2017 as “the year of short videos.”

Most people using online video— premium or UGC—are quite young. According to iResearch, about 83% of iQiyi’s mobile users in August 2017 were younger than 35. The demographic is similar for short videos, with more than 80% of users younger than 35.

A major difference between the premium video products in China and in the U.S. is the emphasis on social features. If you look at the video discovery or display pages carefully, you will find sharing buttons on every page, linking to every possible social network. Excerpts from the longer shows, like songs or jokes, are often very popular clips for users to share. These services also provide interaction such as screen bullets, comments and love buttons, to improve the social experience while watching (Figure 2).

Another type of social interaction for the premium video service is with media stars. iQiyi PaoPao is a place for celebrities to interact with their fans. Hundreds of movie and TV stars have their home pages set there, with 20 million fans following the top stars. iQiyi claims 600 million active users in PaoPao, who are proven to be more sticky, watching 20% more videos on average than other users.

For short video apps, a significant emphasis is on social functions. These short video apps provide many ways for interaction, including discovering people with similar interests, follows, claps, love, comments, and gifting. The app platforms compete by paying large sums of money to the popular “VIPs,” encouraging them to set up channels on their platform.

For a short video app user, the number of fans defines success, and the top “VIPs” have more than 30 million fans. To become a “VIP,” users compete to create attractive, frequent content, and stimulate excitement with their fans. A slogan popular among the top users is “300 clips a day!”

Technical Opportunities and Challenges

With the prevalence and popularity of video streaming, both infrastructure and application companies in China have invested heavily in video technology. Generally, a video startup time of less than two seconds is considered “good;” of course, one needs to consider the video quality, device type, network situation, and other factors to make a specific judgment. According to a 2017 report from China Broadband Development Alliance, the average video startup time (VST on broadband) was between 0.6 to 1 second, which is better than the world standard.

What is unique about China’s Internet market that fosters technical advancements and innovation? We can call out a few examples: mobile dominance, the huge number of users and available user data, the massive scale of user-generated content, regulatory requirements, ever-changing user interests, and extremely fierce competition.

Features more suitable for mobile viewers, like usage scenarios that support watching movies while commuting on public transit, add to the success of these platforms. For example, all premium video platforms provide offline viewing as a default feature, while short video apps provide offline information such as news for people to view while not connected.

All video providers rely on personalized recommendation; it is especially important for short videos as the main method to discover videos of interest. A commercial recommendation system employs largescale machine learning on real-time streaming data and tries to optimize metrics such as the click-through rate, time spent, and user retention. With the availability of large amounts of data from both first- and third-party vendors, one thing that distinguishes recommendation systems on China’s Internet from the rest of the world is its sheer scale: Alibaba claims its machine-learning platform—eXtreme Parameter Sever (XPS)—processes 10B samples and 100B features daily,t while the Toutiao platform claims tens of billions of features and billions of vectors.u By the same token, strong recall technology is developed to select the top few thousand results from millions of potentially low quality or redundant user-generated content.

In general, machine learning is used to annotate, classify, and analyze video content, and to build user profiles based on the user’s geological location and browsing history. It then uses such information to match a user to videos that reflect their interests. Instead of a ‘pull’ or ‘subscribe’ model, many of the Chinese short video apps ‘pushes’ the relevant content to viewers’ home pages. Since most Chinese viewers are quite receptive to pushed information, recommendation technology is very effective, and users can indulge in content they like for hours a time.

For short forms, users upload videos to platforms every day, which may contain inappropriate content such as pornography, content that infringes on copyrights, and duplications. Although many of the media companies have thousands of human editors, it is difficult to manually examine all contents in real time. To react quickly to the market, video companies developed adaptive machine models that work together with human editors to prescreen and filter out potentially problematic videos. Necessity also prompted the rise of high-tech providers such as SenseTime that specialize in image and video reviewing technology.

Another important application of AI is to add special effects to usergenerated videos. For example, Chinese users often want features such as beautifying faces, adding special costumes, and changing backgrounds. To do so, one needs to detect facial key points and perform highly accurate face and hair segmentation. The related technologies, such as style transfer and object segmentation, are active areas of research in computer vision. Significant progress has been made in recent years using deep learning. In addition, augmented reality (AR) also has a lot of interesting usage on video, like blending fake objects with video backgrounds. These technologies help users create more interesting videos that are better fit for sharing.

To monetize the contents, all video platforms use computational advertising to display ads targeting users based on their personalized interests. In addition, many innovative video technologies are used in video ads, for example, ads overlaid on the videos, ads inserted in live broadcasts, ads on bullet screens, 360 ads integration with direct sell, and more. For example, one video ad service provider—Video++—claims to have more than 9,000 clients using their technology to integrate direct sell ads into streaming video.

Mobile video apps have also presented interesting technical challenges to machine learning and computer vision. Instead of inventing new algorithms, the practical solution was to apply the PC applicable algorithm to mobile. For example, to improve speed, or to use local data, a ML model inference for object recognition may need to be done on a mobile phone. In such circumstances, the size of the ML model must be small enough to fit, and even better, optimized to the phone hardware.

China’s streaming video companies continue to explore many other ways to innovate with technology. For example, Tencent tried to use robots to write articles for live news, and Youku tested auto caption translation. Technology is also used to predict user reactions and suggest content investment in media. iQiyi, for example, claimed to have casted actors in their original shows based on AI predictions.

Summary

Online video, or online media in a larger setting, is one of the best places where technology creatively meets user experience. Through online media, there are infinite ways to connect hundreds of millions of users to billions of pieces of information. This is an exciting time for China’s online video market, as it has just realized mass-market adoption. We look forward to the new technological innovations to come.

# 附录2英文翻译

中国地区

**视频消费，社交网络和影响力**

今年中国的在线娱乐市场收入达到约2000亿美元。中国的视频市场与美国相当这并不奇怪。事实上，中国的在线视频用户数量是美国的2.5倍（即美国用户数为2.12亿，而中国用户为5.79亿）。由于宽带和移动技术的进步，在线视频是中国互联网发展最快的领域，过去五年增长率约为50％。

中国在线视频行业的情况与美国有许多相似之处，呈现出非常有趣的情况和见解。本文概述了市场，主导参与者和商业模式，并介绍了该领域产品之间有趣的细微差异和技术进步。

与美国一样，在线视频呈现两大主流服务：高端的和低端的。高端的是有流媒体版权的节目和电影的付费服务。它们是中国的Netflixes和Hulus。低端的呈现针对不同细分市场的专业生成内容（PGC）和用户生成内容（UGC）。它们是中国的YouTubes和Snapchats。

高端：付费视频平台

付费在线视频的前三名是爱奇艺，腾讯视频和优酷。这些公司分别隶属于百度，腾讯和阿里巴巴。

根据来自中国大数据服务提供商QuestMobile的统计数据显示，截至2017年底，腾讯视频和爱奇艺的月活跃用户数约为5亿，优酷约为3亿。并且去年3月在纳斯达克上市的爱奇艺在总观看小时数上领先。

不同于美国的付费视频服务，中国的视频服务提供商都从有广告的免费服务开始。在线视频的总广告收入大约为100亿美元，赶上了商业电视产生的广告收入。

然而，在过去两年中，当用户开始倾向于（并支付）无广告的服务平台并提供其他额外功能，例如访问更高质量的视频和会员原创内容时，我们看到了订阅量的大幅增加。截至2月份，爱奇艺累计订阅用户达6000万。

由于视频流服务的普及和流行，中国的基础设施和应用公司都在视频技术方面投入了大量资金。

2018年，腾讯视频用户已超过4000万。相比之下，截至2018年1月，Netflix拥有约5500万美国用户和6300万国际用户。

过去几年，中国的中产阶级和年轻的互联网用户的支付行为迅速形成。 这一行为背后的主要推动力包括全面打击盗版内容，实惠的价格，以及最重要的是在线支付的便利性。这三家服务每个订阅价格在20元人民币（3.16美元）到40元人民币（6.32美元）之间。 互联网视频订阅服务的总市场规模已大幅增加，从2012年的约6300万美元增加至2016年的21亿美元，估计在2022年达到115亿美元。

与美国的付费内容发行商不同，许多电视剧和电影都是非独家的，一旦在电视或电影院播出，就可以在所有三家主要服务商和其他地方使用。独家内容通常要贵得多。所有服务商都大量投资于受版权保护的电影和电视节目，导致这些内容的价格非常高。平台通常会对哪些节目会受欢迎，进行有计划的押注，并在制作前提供报价。根据其招股说明书，爱奇艺的年度资源成本约为19亿美元，而其他两个平台的成本几乎翻了一番。这些购买费用远远超出订阅费和广告收入。根据其公开文件，爱奇艺在2018年第一季度损失了1.69亿美元，其他两家也损失了类似的数额。而且这种情况将在未来几年继续。

为了降低成本，并在同行中脱颖而出，所有三家付费服务商都开始在自己的平台上制作独家原创内容。原创节目的领导者是爱奇艺，尽管三家都有不同的热点。与美国的电视剧往往占据主导地位的情况不同，综艺节目在中国获得了更多的观众。根据华尔街日报最近发表的一篇文章“The Rap of China”表明，由爱奇艺创作并展示的针对年轻观众的12集嘻哈说唱竞赛真人秀，已成为“2017年中国最受欢迎的娱乐节目”。 该节目在6月下旬至9月初期间吸引了27亿次观看。在社交媒体微博上，从节目中截取的短片视频被观看了80亿次。

关于中国付费在线视频与美国不同的另一个事实是，他们都参与了PCG和UGC的短视频市场。但是，正如我们将要讨论的那样，其他新兴者日益主导这些方面。

低端： UGC和PGC视频平台

2017年中国的短片视频非常活跃，有超过100名玩家和4亿用户，并且远未稳定。它们的用户群庞大，增长迅速，非常活跃。有数个主要的平台，而且其中大多数在过去一两年内开始流行。

与美国市场不同，最初占短片视频平台主导地位的是高端付费平台，如爱奇艺和优酷。例如，优酷声称自从2015年以来，在用户生成的内容上投入了大约16亿美元。这些内容以YouTube为蓝本，并且有大量的观众观看混合了优质内容和UGC内容的视频。他们的短片包括电影剪辑和音乐视频，以及自由形式的用户创建的内容，并且他们为专业制作者提供创建的渠道。

然而，在过去的两年里，我们目睹了几个与高端付费平台无关的移动短视频应用的惊人增长。由于它们的易用性，这些应用变得超级流行，在中国广大地区拥有大量用户（参见图1）。2017年第三季度，短视频平台的流量同比增长311%。分析人士称2017年的中国互联网时代为“短片之年”。

新的流行短视频服务平台包括快手，火山，西瓜，抖音，秒拍，美拍，微视等等。其中最受欢迎的应用每月活跃用户数在1亿至2亿之间。这些应用主要是移动视频流，视频长达几秒到几分钟。这些应用程序为用户提供了很好的工具来拍摄、编辑、美化和添加特殊效果。这些视频应用程序可以分为两类：一种是更像移动应用程序“dubsmash”的，用户可以录制自己的视频和配音，另一种更像是带有自由格式视频的Snapchat。所有视频应用程序都提供强大的发现和关注功能，鼓励用户之间的互动和社交联系。

更具体地说，西瓜的视频主要是PGC内容和时长1-5分钟的视频剪辑。火山和快手的都是约15秒长的UGC视频短片。许多日常用户，特别是农村地区和小城市的用户，通过这些应用程序录制短视频并分享它们。应用程序为用户提供了一种展示自己，竞争获得粉丝和眼球，最终从收视率中获利的途径。每天活跃用户接近1亿的快手是目前这一领域的领先者。抖音最近受到了相当多的关注并获得了丰富的用户群，声称每月活跃用户超过1.5亿。

抖音以及微视等其它竞争应用允许用户使用提供的模板创建和分享短视频。

另一大在线视频服务平台专注于终端用户直播：它们为消费者，流行偶像和普通大众提供直播功能。根据Pandaily的数据，2018年上半年，中国的8个直播平台共筹集了约116亿美元。其中领先的平台—虎牙和斗鱼—占总数的近70％。

就像所有高端服务都提供UGC一样，所有短视频平台也都提供直播。而且形势非常活跃；与几个月前的情况完全不同，而且从现在起到几个月后情况可能又会有很大差异。

新的移动视频应用程序仍在尝试商业模式，通过广告，联盟营销到电子商务，赠送礼物和其他新方法赚钱。这些平台上的顶级“VIP”可能拥有数百万粉丝，他们通过广告或销售商品获利。各知名品牌也开始在这些平台上开创渠道。根据iResearch发布的中国短视频产业报告，短视频领域的收入在2017年达到8.6亿美元，预计到2020年将超过45亿美元。

我们刚刚开始看到中国短视频市场的巨大潜力，并且它将在未来几年内发展。不同于人们在专门的领域或者人口统计学，它有垂直分割的趋势。随着中国国内的竞争日益激烈，许多这类短视频应用平台都将走向海外，尤其是亚洲其他国家。例如，2017年11月快手是韩国应用商店的头号视频应用。根据2018年4月来自36kr的一份报告，亚洲流行的短视频应用中几乎有一半是“中国制造”。

移动和社交视频

大屏幕，包括电视和其他OTT电视设备，在中国受到限制。而由于移动设备的高普及率，在线视频的使用强烈倾向于移动设备。即使是付费用户，主流的偏好也是移动设备和个人电脑。根据爱奇艺的数据，2017年第四季度，移动设备上平均每月活跃用户数（MAU）为4.21亿，个人电脑上平均为4.24亿，电视和OTT设备可以忽略不计。许多短视频应用程序专为移动设备设计和使用。

分析人士将2017年的中国互联网时代称为“短视频年”。

大多数使用付费或UGC在线视频的用户都很年轻。根据iResearch的调查，2017年8月爱奇艺的移动用户中有83%的人年龄小于35岁。短视频的用户统计与之相似，超过80%的用户年龄小于35岁。中国的付费视频产品与美国之间的主要区别在于更强调社交功能。如果你仔细观察视频的发现或显示页面，你会发现每个页面上都有共享按钮，链接到每个可能的社交网络。从较长的节目截取的短视频，如歌曲或笑话，通常是非常受欢迎和用户乐于分享的剪辑。这些服务还提供诸如弹幕，评论和关注按钮之类的交互，以改善观看时的社交体验（图2）。

付费视频服务的另一种社交是和媒体明星互动。爱奇艺的泡泡是明星与粉丝互动的好地方。数以百计的电影和电视明星都设置了自己的主页，其中有2000万粉丝关注这些顶级明星。爱奇艺声称泡泡有6亿活跃用户，他们被证明比其他用户平均多看20%的视频。

对于短视频应用来说，重点是社交功能。这些短视频应用程序提供了许多交互方式，包括发现具有相似兴趣的人，关注，称赞，喜欢，评论和赠送礼物。应用程序平台通过向流行的“VIP”支付大笔报酬来相互竞争，以鼓励他们在自家的平台上建立渠道。

对于短视频应用用户，粉丝数量意味着成功，顶级“VIP”拥有超过3000万粉丝。为了成为“VIP”，用户们创造有吸引力的、丰富的内容，并激发他们的粉丝们的兴奋来竞争。在顶级用户中流行的口号是“每天300个剪辑！”

技术机遇和挑战

随着视频流的流行和普及，中国的基础设施和应用公司都在视频技术方面投入了大量的资金。一般来说，一个视频启动时间少于两秒被认为是“好的”；当然，需要考虑视频质量、设备类型、网络状况等因素做出具体的判断。根据中国宽带发展联盟2017年的报告，平均视频的启动时间（宽带VST）在0.6到1秒之间，这一水平优于国际标准。

中国互联网市场的的哪些独特之处促进了技术进步和创新？我们可以举几个例子：移动优势，大量用户和可用的用户数据，大量的用户生成的内容，监管要求，不断变化的用户兴趣以及极其激烈的竞争。

更适合使用移动设备的观众的功能，例如适合在公共交通上通勤时观看电影，增加了这些平台的成功。例如，所有付费视频平台都提供离线观看作为默认功能，而短视频应用也提供离线信息，比如人们在未连接网络时查看的新闻一样。

所有的视频提供商都依赖于个性化的推荐，作为发现感兴趣视频的主要手段，它对短视频尤为重要。商业推荐系统对实时流数据采用大规模的机器学习并尝试优化点击率，花费的时间和用户保留度等指标。随着来自第一方和第三方供应商的大量可用的数据，中国互联网上的推荐系统与世界其他地区的一个区别在于其规模庞大：阿里巴巴声称其机器学习平台—eXtreme Parameter Sever (XPS)—每天处理10B样本和100B特征，而Toutiao平台声称有数百亿个特征和数十亿个向量。出于同样的原因，开发了强大的召回技术，以从数百万个潜在的低质量或冗余的用户生成内容中选择优质的结果。

通常，机器学习用于批注，分类和分析视频内容，并基于用户的地理位置和浏览历史来构建用户档案。然后，它使用此类信息将用户与用户感兴趣的视频进行匹配。许多中国的短视频应用程序将相关内容推送到观众的主页，而不是“拉取”或“订阅”模式。由于大多数中国观众都非常乐于接受推送信息，因此推荐技术非常有效，用户可以沉迷于自己喜欢的内容。

简单的说，用户每天都会上传视频到平台，其中可能包含不当内容，如色情内容，侵犯版权的内容以及重复内容。尽管许多媒体公司拥有数千名编辑人员，但很难实时手动检查所有内容。为了对市场做出快速反应，视频公司开发了自适应机器模型，与编辑人员一起预先筛选并过滤掉可能存在问题的视频。这方面的必要性也促使专注于图像和视频审查技术的SenseTime等高科技提供商的崛起。

AI的另一个重要作用是对用户生成的视频添加特效。例如，中国用户经常需要诸如美颜，添加特殊服装和更改背景等功能。为此，需要检测面部关键点并执行高度准确的面部分割。风格转移和对象分割等相关技术是计算机视觉研究的活跃领域。近年来，通过利用深度学习取得了重大进展。此外，增强现实（AR）在视频上也有很多有趣的用途，比如将虚拟对象与视频背景混合在一起。这些技术帮助用户创建更适合分享的有趣视频。

为了通过内容获利，所有视频平台都使用广告计算技术来根据用户的个性化兴趣展示符合用户定位的广告。所有视频平台都使用广告技术来显示基于用户个性化兴趣的具有目标性的广告。此外，视频广告中使用了许多创新的视频技术，例如，覆盖在视频上的广告，直播中插入的广告，弹幕上的广告，直接销售的360广告集等等。例如，一家视频广告服务提供商-Video++声称拥有9000多家客户使用他们的技术将广告整合到流视频中。

移动视频应用也对机器学习和计算机视觉提出了有趣的技术挑战。实际的解决方案是将PC应用算法移植到移动领域，而不是发明新的算法。例如，为了提高速度或使用本地数据，可能需要在移动电话上使用用于对象识别的ML模块。在这种情况下，ML模块的尺寸必须足够小甚至更小，以适应手机的硬件大小。

中国的流媒体视频公司将继续探索更多其它的创新技术。例如，腾讯试图使用机器人为实况新闻撰写文章，而优酷则测试了自动字幕翻译。还有技术用于预测用户的反应并建议媒体来如何投资内容。例如爱奇艺声称在他们的原创节目中演员的表演是基于人工智能预测的。

总结

在线视频，或者更广泛的说在线媒体是使用创造性技术让用户体验的最佳场所之一。通过在线媒体，可以通过无限的方式将数亿用户连接到数十亿条信息。对于中国的在线视频市场而言，这是一个激动人心的时刻，因为它刚刚实现了大众市场的使用。我们期待看到新的革命技术的到来。