**网络评论过滤与分析系统**

**研究报告**

[一、选题背景 3](#_Toc3201252)

[*（一）* *研究目的* 3](#_Toc3201253)

[*（二）研究背景* 3](#_Toc3201254)

[*（三）研究意义* 5](#_Toc3201255)

[二 、实施进程 6](#_Toc3201256)

[三 、成果内容 7](#_Toc3201257)

[*（一）项目内容简介* 7](#_Toc3201258)

[1.1热点话题、敏感话题识别 7](#_Toc3201259)

[1.2评论主题追踪 8](#_Toc3201260)

[1.3自动摘要 8](#_Toc3201261)

[1.4评论趋势分析 8](#_Toc3201262)

[1.5评论统计报告 8](#_Toc3201263)

[*（二） 项目实际内容* 9](#_Toc3201264)

[2.1网络评论过滤与分析系统web端 9](#_Toc3201265)

[2.2 PC端分析报告 10](#_Toc3201266)

[四、创新点 13](#_Toc3201267)

[*（一）创新点* 13](#_Toc3201268)

[1.1技术创新 13](#_Toc3201269)

[1.2 功能创新 16](#_Toc3201270)

[1.3信息公开透明 16](#_Toc3201271)

[1.4 自动报告文档 17](#_Toc3201272)

[*（二）特色* 17](#_Toc3201273)

[2.1前沿性 17](#_Toc3201274)

[2.2 实用性 18](#_Toc3201275)

[2.3开放性 18](#_Toc3201276)

[2.4安全性 18](#_Toc3201277)

[*（三）学生主导表现* 19](#_Toc3201278)

[3.1分工明确 19](#_Toc3201279)

[3.2 独立开展 19](#_Toc3201280)

[3.3 积极讨论 19](#_Toc3201281)

[五、管理战略 20](#_Toc3201282)

[5.1项目概述 20](#_Toc3201283)

[5.2利益相关群体 21](#_Toc3201284)

[5.3 项目发展战略 22](#_Toc3201285)

[*六、市场营销* 23](#_Toc3201286)

[6.1目标市场 23](#_Toc3201287)

[6.2服务 23](#_Toc3201288)

[6.3推广策略 23](#_Toc3201289)

[七、 成果应用情况 24](#_Toc3201290)

[*（一）成果应用情况* 24](#_Toc3201291)

[*（二）应用前景* 24](#_Toc3201292)

[八、研究心得 25](#_Toc3201293)

[*（一）创新思维与成长的收获* 25](#_Toc3201294)

[*（二）可借鉴的经验* 25](#_Toc3201295)

[九、相关文献 26](#_Toc3201296)

# 一、选题背景

## *研究目的*

随着互联网的快速发展，网络媒体作为一种新的信息传播形式，已深入人们的日常生活。网友言论活跃已达到前所未有的程度，不论国内外重大事件，都能马上形成网上舆论。好的舆论可以成为发展的“推进器”，不好的舆论可以成为民众的“迷魂汤”、社会的“分离器”、杀人的“软刀子”、动乱的“催化剂”。由此可见，坚持正确的舆论导向，建设积极健康的网络空间必不可少。因此此次我们搭建网络评论过滤与分析系统，一方面期望净化评论环境，另一方面能够有效整理并分析评论，并加以正确应对和引导。

网络评论过滤与分析系统，通过对网页评论内容的采集处理、敏感词过滤、智能分类、主题检测、专题聚焦、统计分析，实现各团体及个人对自己相关网络舆情监督管理的需要，最终形成舆情简报、舆情专报、分析报告、移动快报，为决策层全面掌握舆情动态，做出正确舆论引导，提供分析依据。

## *（二）研究背景*

**1.社会现状分析：**

党的十八大以来，国家进入全面建成社会主义现代化强国的新征程。推动社会主义文化繁荣昌盛，牢牢掌握意识形态工作领导权具有重大意义。要掌握意识形态工作领导权，要建设好网络空间。互联网不是法外之地，要加强建设和管理，要本着对社会负责、对人民负责的态度，建立网络综合治理体系。尤其要加强网上舆论引导，培育向上向善的网络文化，营造一个风清气正的网络空间。

在当今新媒体强势兴起和传统媒体非常活跃的环境下，舆论场一方面呈现出自由开放的态势，有利于公民对公权力行为的监督，使一些腐败官员丢掉乌纱帽，受到了法律的制裁，媒体有效地行使了舆论监督的作用；另一方面不实信息混淆其中，甚至有的借机造谣，迷惑受众，误导社会舆论。舆论监督是新闻传播的基本功能，不仅代表着媒体的良知，也昭示着我国民主法治的进程。舆论对社会的导向性可见一斑.由此把握好芜杂信息的处理原则，去伪存真，净化舆论监督环境不可忽视。

可见，在当今的媒体时代，舆论监督的重要性不可忽视。因为它直接来源与群众，却又有区别于群众的呼声。同时我们应该注意到，近年来随着微博、贴吧、微信公众号等自媒体平台的不断发展，舆论导向不仅仅局限于新闻中，在这些允许用户自由讨论的、更贴近用户平常生活的平台中，面向用户的层次不同，评论内容也有着观点鲜明的特征，使得这些评论有着极大的研究价值。而如今大多数研究还局限在对新闻舆论的监督与分析，很容易忽略这些贴近生活，时下兴起的用户评论。此外，由于评论基数大，质量参差不齐，对这些评论不能很好地统计与分析。亦或是只能窥见其冰山一角，以偏概全。

开展这个项目之前我们对相关问题进行了一些调研，具体调研结果如下：

参加数据调研者150人，**分别**来自各行各业

**2.研究现状分析：**

国内外已有对评论文本的特点、文本的句法结构、机器学习等方法对垃圾内容的检测识别进行研究。Jindal[1]等认为垃圾文本内容具有重复性，故通过识别评论中重复或近似重复的内容与评论对象的语义关系进行识别；刁宇峰[2]等利用LDA模型对博客数据进行主题聚类，将博客文本包含的与垃圾评论相关的词汇聚类为垃圾主题，并用主题形式对隐含垃圾观点进行刻画，提高垃圾观点检测的准确率；Mukherjee[3]等够早了正对在线评论的与LDA类似的产生式模型，加入了评论作者、文本、评分等生成关系，提高了垃圾主题聚类性能

## *（三）研究意义*

在这个背景下，网络评论过滤与分析系统呼之欲出。该系统通过对热点问题和重点领域比较集中的网站信息，如：网页、论坛、BBS等，随时爬取最新的消息和意见。爬取后完成对数据格式的转换及元数据的标引。借鉴已有的工作中对文本特征与垃圾评论相关关系规律的总结，结合基于规则的分类模型以及神经网络架构，对爬取至本地的信息，进行初步的过滤和预处理。对热点问题和重要领域实施监控，此外，对信息阐述的观点、主旨进行倾向性分析，以提供参考分析依据。分析的依据可根据信息的转载量、评论的回复信息时间密集度来判别信息的发展倾向

通过网络评论过滤与分析系统，相关部门可以及时了解网络舆情动态，关注到自己在网络舆情中的状态，从而可以产生网络舆情预警，及时纠正应对网络上关于自己的负面舆论影响，为部门网络危机公关或品牌形象营销提供数据依靠。对于个人用户与研究机构而言，该系统可以帮助用户高效、完整的了解网络上的其他用户对各种事物的评论数据统计。

# 二 、实施进程

2018年10-11月：

学习相关技术，HTML5、CSS、XML、JavaScript、PHP、SQL等，查阅相关开发书籍。

爬取大量数据，搜集大量数据集与论文，对评论筛选模型进行设计与测试。其中查阅NLP内相关知识，联系被试人员对一些无效评论进行人为设计以及分类，并测试筛选模型。

2018年12月至2019年1月：

筛选模型的进一步测试，以及研制评论分类模型。由上一阶段的成果先进行划分，并在一些主领域进行分类（如支持、反对等）。收集数据集对模型进行优化，同时考虑到人文因素，仍需联系被试人员进行人为测试以及准确率检验

2019年2月至2019年3月：

将得出的数据结果进行统计分析或者是采用决策分析算法，通过阅读有关统计知识采取科学的展示方式，数据库设计与建表。

2019年4月至2019年6月：

优化系统操作体验。考虑联系部分媒体以及商业公司。试采用该模型进行信效度检验，测试在现实情况下的使用情况以及是否存在疏漏等。

2019年7月至2019年8月：

对模型进行优化以及修复bug.进入测试阶段，主要通过邀请学生参与试用评测的方式，记录每次测试的相关数据，根据数据结果对出现的bug或不足之处进行修改加工，并拿出最终的成品。

2015年9月:

开始撰写关于本项目的研究报告，其中将包括整个研发过程的概述，最初的想法，初步的构思，明确框架，着手研究等，重点在于技术实现过程中遇到的难题及解决办法，最后有小组成员的心得体会等。进入结题及答辩阶段，向组织方提供完整的项目成果，并就专家组的问题进行答辩。

# 三 、成果内容

## *（一）项目内容简介*

网络评论过滤与分析系统搭载在web平台，首页操作简单，易于识别。用户可在服务窗内选择自主查询生成报告，亦或是查询已有热门评论报告等操作。用户通过输入网页url链接,系统将对网页评论进行爬取，并反馈给用户相关评论报告。或者是在已有数据库中搜索当前热门事件的评论报告。同时管理员每日也会更新热门话题内容，对用户的热门输入进行归纳，定期对数据库进行增删操作。

### 1.1热点话题、敏感话题识别

热点话题、敏感话题识别，一方面指的是管理员方定期自主对热门平台（微博、贴吧、知乎等）的热门话题进行评论爬取，并产生有效的数据报告加载至数据库。另一方面指的是用户方对感兴趣的话题进行输入，系统将对热门输入的网页进行关注，并引入数据库中方便其他用户查阅。系统在关注热点的同时，对话题的敏感度也会密切关注。当该话题涉及政治敏感，以及大量不良信息冗杂其中（如无意义评论、水军等），为防止误导用户，系统将及时对评论报告进行更新与警示处理。

### 1.2评论主题追踪

评论主题追踪，指的是每篇报告都是在不侵犯用户隐私的情况下生产的，对于相关话题报告将详细记录该平台以及生成日期，方便用户自行查阅。此外，对评论主题系统会定期更新评论数量并导入模型，由模型输出来更新表单，实现评论主题追踪功能。

### 1.3自动摘要

自动摘要。对于事件以及评论内容，系统在避免主观诱导的情境下，对内容进行简单摘要，对评论则既凸显宏观局势，又相对突出热门个体。

### 1.4评论趋势分析

评论趋势分析，是该系统的特色功能之一。系统不是简单地对评论进行统计、展现。而是动态地针对有效评论数、人员分布比、评论热门趋势等众多因数，由科学的统计算法归纳出评论趋势走向，以确保评论的实效性、真实性。

### 1.5评论统计报告

评论统计报告，是该系统重要特色的直观体现。评论统计报告形式上以web端呈现，内容上可分为事件简介、事件趋势、热点评论、情绪分析、用户分析以及评论分析。既能将评论的真实特性完美展现给用户，又能满足不同用户的多方面需求。

## *（二） 项目实际内容*

### 2.1网络评论过滤与分析系统web端

因为本产品是基于web平台的，所以只要通过浏览器进入系统首页，即可进行一系列的使用。为方便用户体验，以及考虑到功能识别度，我们采用了类搜索引擎的界面设计。

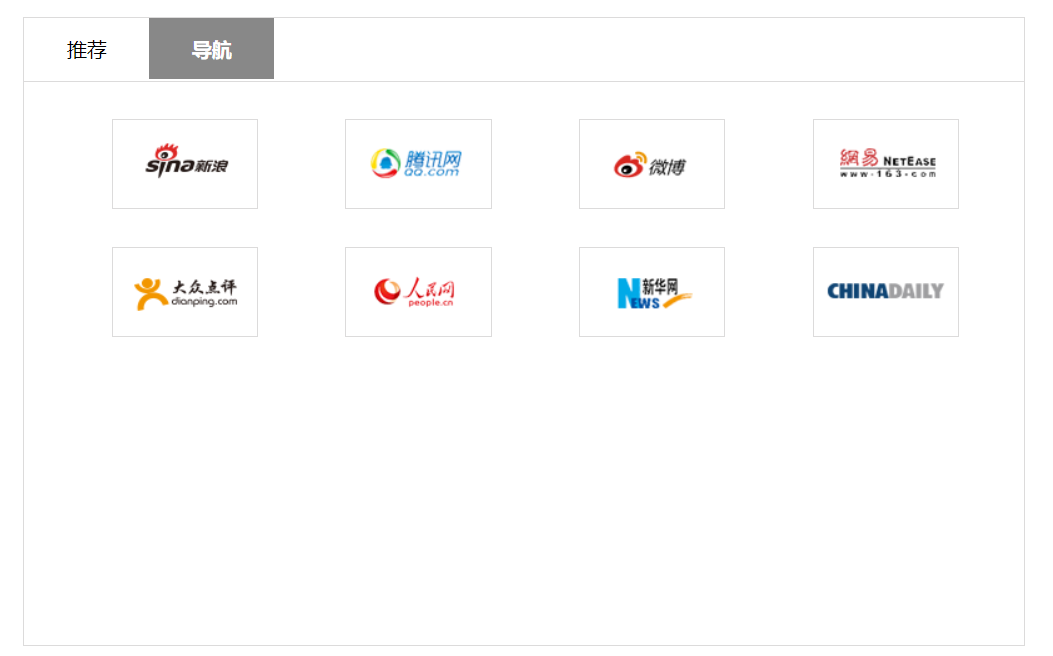


界面的人性化设计。采用红黄蓝三原色，象征着包罗不同的评论以及不同的思想

除独特的logo设计外，搜索栏在外观和功能上都十分符合用户期望。一方面支持用户自主查询网页，另一方面支持用户对已有报告进行查询。



推荐界面，是对近期热门话题由热度进行的系统自主推荐排序，方便用户快速查询，了解当下热门话题。



导航页面，方便用户快速进行查阅以及了解本系统所支持页面。

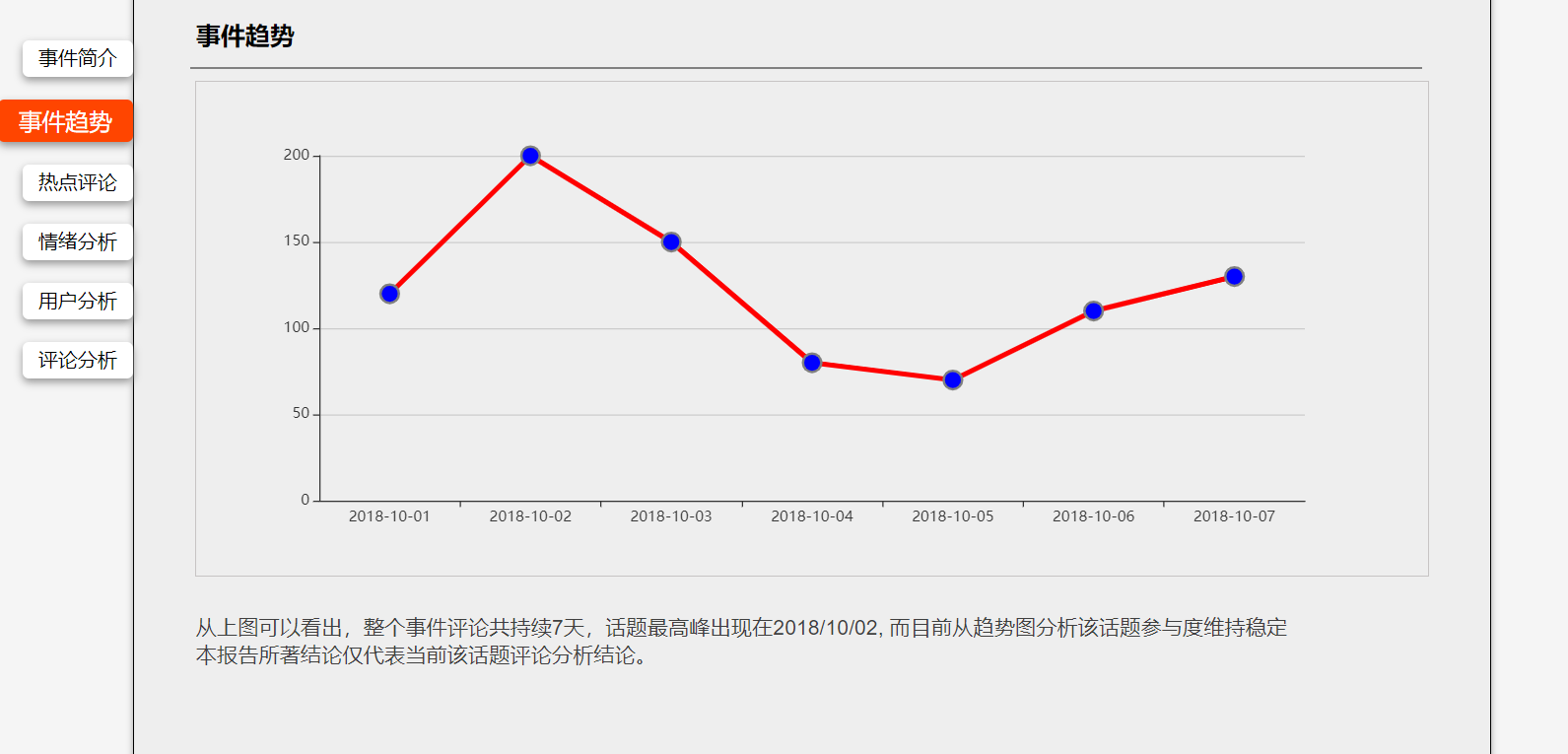
系统首页简洁明了，目前已在web端实现，将来考虑是否可移植成APP形式方便移动端用户使用。

### 2.2 PC端分析报告

分析报告是本系统的特色功能的直观表现，目前由五大块组成：事件简介、事件趋势、热点评论、情绪分析、用户分析以及评论分析，页面美观且操作简洁



报告重点突出，侧边栏方便用户进行查询，结合科学图表直观表现评论分析结论。



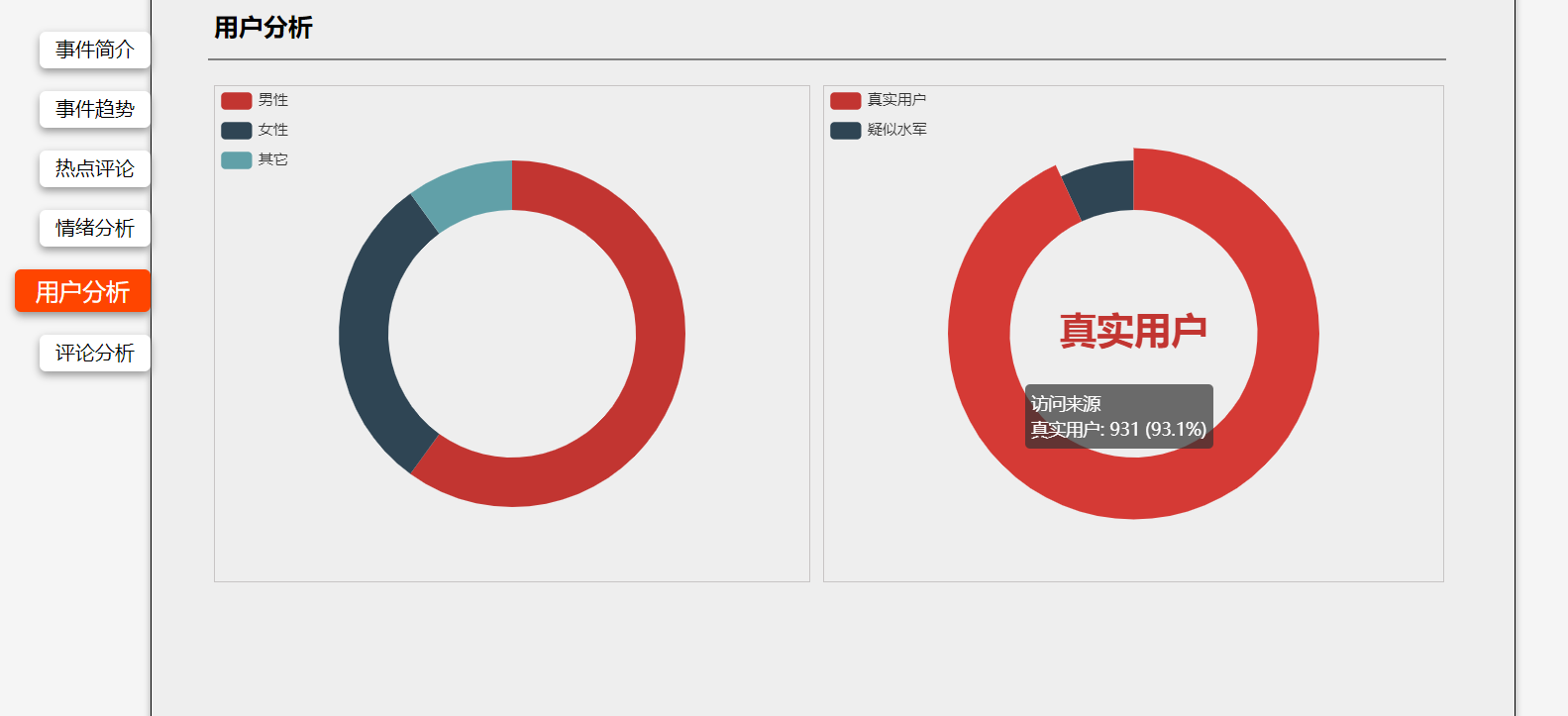
“事件趋势”板块，统计每日用户量，以折线图表的形式直观展现，并对数据进行客观分析以及趋势分析。



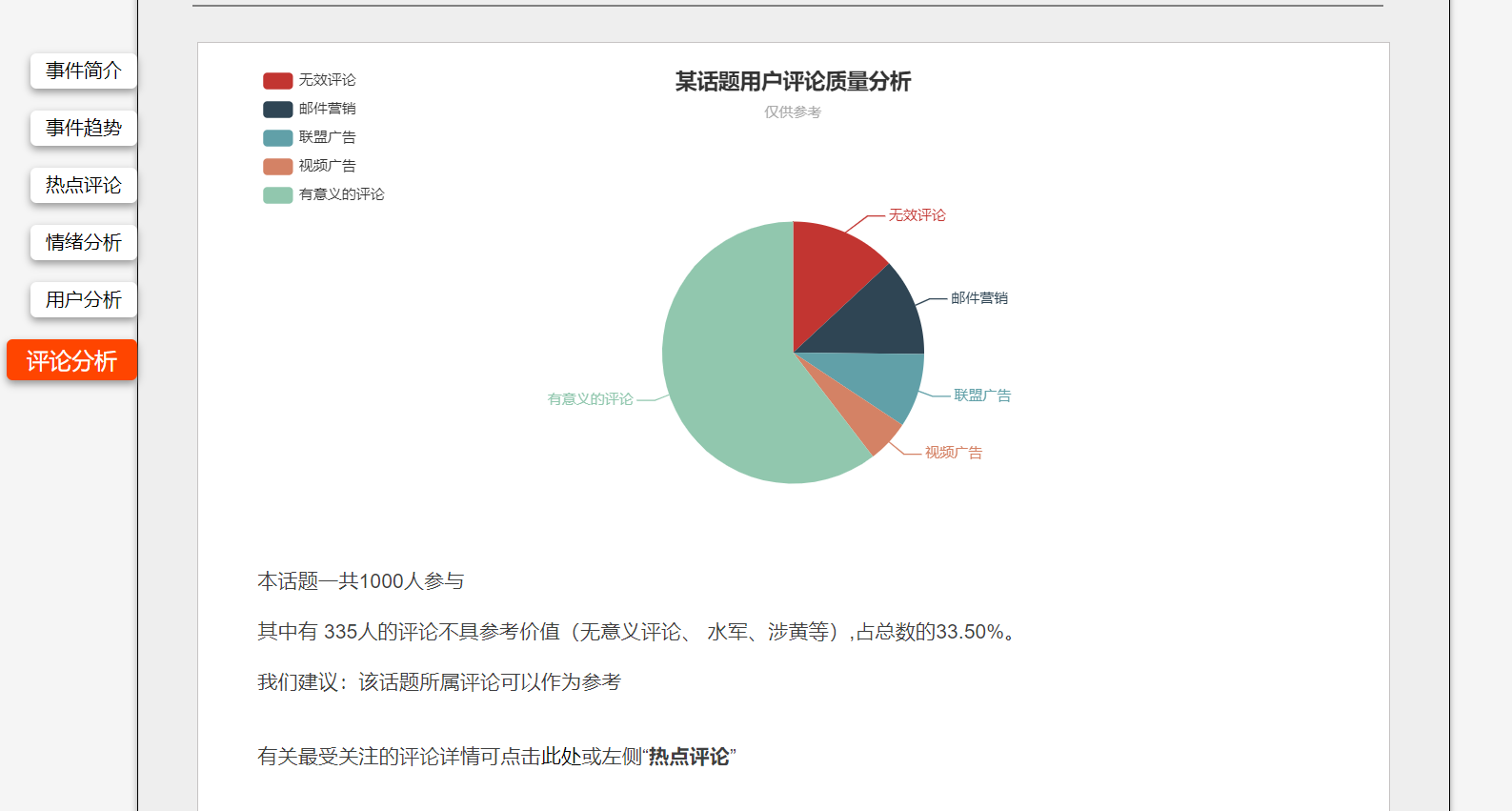
“热点评论”板块，在“抓大”的同时也关注“抓小”，针对实实在在的个体用户进行热点评论摘要，方便查阅用户了解当前用户评论形式以及动向。



“情绪分析”图表板块，对有效评论进行情感文本分类，统计不同用户对该话题的情绪动向，并结合走势图做出客观总结与建议。



“用户分析”图表板块，对用户本身进行性别、真实性分析。在不触及用户隐私的前提下，对评论本身的真实性进行检验，并及时反馈在评论报告图表当中。



“评论分析”图表板块，对评论进行总体、全局把握分析，让用户了解当前评论文本分类分布，以及该评论是否具有可参考价值。及时做到主题跟踪、敏感话题预警。

# 四、创新点

## *（一）创新点*

### 1.1技术创新

我们选择基于CNN的TextCNN模型进行情感分类与文本特征提取。卷积神经网络在数字图像处理领域取得了巨大的成功，从而掀起了深度学习在自然语言处理领域的狂潮。2015年以来，有关深度学习在NLP领域的论文层出不穷。在情感分类中各类模型也鳞次节比地出现，例如，Fast Text、TextCNN, RCNN等。相较于Fast Text与RCNN, TextCNN 在模型准确率上有大幅度提高，并且更加高效。因此我们小组决定采用TextCNN来实现对文本中的情感分类。

同时，我们还引入了残差学习[4]。残差学习能够解决退化现象的问题，使其在计算机视觉领域大获成功。同时，从goole提出的transformer模型[5]中我们也可以看到残差学习也在自然语言处理领域得到了很好的表现效果。因此我们小组希望通过在TextCNN中引入残差学习，使其获得更加优异的表现成果。以下是我们的模型设计：

A）基于VGG[6]模型

与VGG采用了一样的设计模式：

i）对于相同的输出特征图尺寸，层具有相同数量的滤波器

ii）如果特征图尺寸减半，则滤波器数量加倍，以便保持每层的时间复杂度。我们

iii）通过步长为2的卷积层直接执行下采样。网络以全局平均池化层和具有softmax的1000维全连接层结束。

B）采用BN方法

我们可以采用随机梯度下降算法对神经网络进行训练，但是这样的带来的问题就是难以收敛，因为每一个mini-batch的梯度方向与整个数据集的梯度方向不一定一致，所以会导致训练时间大大加长。我们可以通过增加batch size使得每一次训练的梯度方向尽可能与整体的方向一致。另外，我们也可以采用BN，使每次mini-batch时每个层的输入数据都是均值0方差1，那么就可以解决这个问题。

C）提出残差学习的思想。

传统的卷积网络或者全连接网络在信息传递的时候或多或少会存在信息丢失，损耗等问题，同时还有导致梯度消失或者梯度爆炸，导致很深的网络无法训练。ResNet在一定程度上解决了这个问题，通过直接将输入信息绕道传到输出，保护信息的完整性，整个网络只需要学习输入、输出差别的那一部分，简化学习目标和难度。VGG和ResNet的对比如下图所示。ResNet最大的区别在于有很多的旁路将输入直接连接到后面的层，这种结构也被称为shortcut或者skip connections[4]。如下图所示：

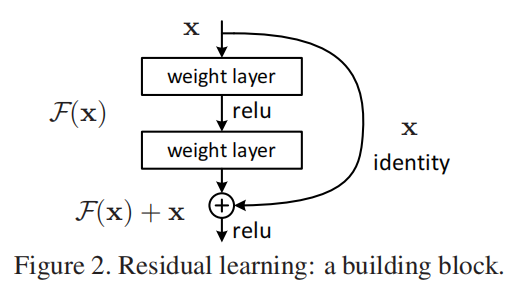


图 2

在面对退化问题上，如果我们能使得深层网络的后面那些层(deeper counterpart[4])是恒等映射，那么原本的复杂模型就退化为一个浅层网络。但是对于恒等映射的学习是相当困难的，相反，对于一个零矩阵的学习则显得较为容易。





若将输入设为X，将某一有参网络层设为H，那么可以将以X为输入的此层的输出y记为H(X)。一般的CNN网络如Alexnet[7]、VGG等会直接通过训练学习出参数函数H的表达，从而直接学习X -> H(X)，这里H的学习目标就是恒等映射。令F(x)=H(x) − x，残差学习则是致力于使用多个有参网络层来学习输入、输出之间的参差即H(X) - X，即学习X -> F(x)+ X。其中X这一部分为直接的恒等映射，而F(x)则为有参网络层要学习的输入输出间残差，这里F的学习目标就是一个零矩阵。

D）使用深层神经网络

近年来，研究表明卷织神经网络深度在决定模型训练效果上扮演了至关重要的作用[8],[9]。网络越深，往往伴随着表现效果更好，尤其是在ResNet网络结构提出后，大幅度减少了退化现象的出现。因此，我们小组决定采用深层次的神经网络来实现对情感特征的精确提取。

最后我们分别用Ren-CECps 1.0中的文本语料与NLPCC2013微博情绪分析样例数据对模型进行测试，结果如下图所示：

|  |  |  |
| --- | --- | --- |
| **model** | Ren-CECps 1.0 | NLPCC2013 |
| TextCNN | 87.3% | 81.6% |
| TextCNN + shortcut | 92.7% | 85.2% |

### 1.2 功能创新

传统的评论分析都是基于特定的规则的，例如，我们可以根据所选筛选条件，列出舆情中提及地域的舆情量分布，一篇文章如果提及多个地域，被提及的地域的次数越多，提及该地域的舆情量越多。或者，我们也可以使用云热词，即根据所选筛选条件，显示该主题相关的热词，该热词相关的舆情量越大，字越大。

但是这样的分类是往往是不精确的，因为机器是没有办法理解我们的语义的。因此传统的评论分析没有办法分析评论中所夹杂的情感，更难以检测评论的走势。因此我们决定采用深度学习的方法，对评论中的情感进行分析。同时，我们还考虑到，不是所有的评论都是有效评论，反之，当下互联网上有着多种内容无意义的评论，例如,“我随便一评论就是标准的十五个字”、“水经验水经验水经验水经验水经验”等。这些评论的存在加大了我们分析的难度，因此我们需要通过一定的数据处理来删除可能会对我们结果产生影响的无关评论。深度模型的使用，让我们的分析更具有说服力，并一定程度上使机器具备了理解人类语言的能力。

同时，通过对评论进行分析，我们可以根据相关的时间戳来判定评论的走势，起到一个预警的作用，避免爆吧现象的产生，给用户带来较好的使用体验的同时保证了评论质量。杜绝了有组织地对目标贴吧发送大量水帖、垃圾帖，以将吧内各帖挤下去，使用户不能在贴吧内正常交流的现象的产生。

### 1.3信息公开透明

本系统将公开所有的信息源，并面向所有用户开放。如今市面上的许多舆情系统仅面向政企开放，许多关键的大数据也被互联网巨头、政府机构所垄断，但本系统将致力于捕捉互联网上尽可能多的公开的观点，从量级趋势上判断该观点的“灌水”程度、传播趋势、情感走向，同时面向所有个人用户免费开放，保证信息的公开透明，使个人也可以感受到大数据带来的便捷。

### 1.4 自动报告文档

本系统将会针对用户输入的话题自动生成相应的报告文档。文档内容将包括：

一、话题的“灌水”程度，将显示有多少人的评论不具参考价值（无意义评论、水军、涉黄等），占总数的多少，并列举其中最受关注的评论；

二、话题的热度，将使用图表显示话题参与人数与热度的关系，并预测该话题在未来的走势；

三、在过滤完那些不具参考价值的评论以后，对剩下的有效评论中进行情感分类，得出用户的态度总体是正向或负向或中立。

通过报告与可视化数据，可以让用户完整而效率地看到由网络上的其他用户对各种事物的有效评论生成的统计数据。

## *（二）特色*

### 2.1前沿性

评论的分类与过滤技术近年来非常火热。国内外目前对于广告评论的分类检测方法可以归结为三类：一是基于内容特征的广告评论检测，二是基于上下文特征得到广告评论检测，三是基于用户行为特征的广告评论检测。其中那于内容特征的广告评论检测方法中，基于向量空间模型（VSM）的文本分类方法应用最普遍。而目前用于广告文本过滤中的分类方法可以分为两类，一类是基于规则的方法，另一类是基于统计的方法。 Jialong Zhang等提出了一种内容无关的、基于利用发表者发表评论结构信息进行评论垃圾分类的方法，兰丹媚提出了基于同义词词林相似度的AdaBoost分类方法和基于内容特征提取的SVM分类方法对垃圾评论进行过滤。

本系统将致力于汲取前人的经验，尝试使用更好的机器学习、自然语言处理方法来完成评论的过滤与分类，并通过网页形式服务于用户，帮助个人、企业高效、完整的了解网络上的其他用户对各种事物的评论数据统计。

### 2.2 实用性

本系统拥有强大的实用性。一方面，本系统的数据将取自全网三大信息源，包括微博、知乎、论坛等，信息源可靠，收集得到的评论将有效服务用户。另一方面，通过网络评论过滤与分析系统，相关部门可以及时了解网络舆情动态，关注到自己在网络舆情中的状态，从而可以产生网络舆情预警，及时纠正应对网络上关于自己的负面舆论影响，为部门网络危机公关或品牌形象营销提供数据依靠。而对于个人用户而言，该系统可以帮助用户过滤垃圾评论，从而完整、效率地看到由网络上的其他用户对各种事物的有效评论生成的统计数据。

### 2.3开放性

我们在系统开发过程中使用的是Django框架，保证了系统的可扩展性、灵活性、可维护性。Django相对于flask框架，提供了众多功能组件，包括项目工程管理的自动化脚本工具、模板、表单、admin管理站点、认证权限、session机制等。MTV设计模式简单、易扩展，代码与样式分离，使得修改更容易。同时Django本身也很安全。默认情况下，Django比PHP确实更好地防止了许多常见的安全错误，Django会通过动态生成网页并通过模板向Web浏览器发送信息，“隐藏”网站的源代码（CSS和html文件除外），直接在互联网上查看。

### 2.4安全性

本系统将部署在阿里云服务器上来保证网站、数据库安全性。阿里云RDS支持适用TLS1.2 SSL协议，写入磁盘时数据是加密的。阿里云云数据库RDS版还提供 SQL 审核，可以实时跟踪用户在何时访问了实例以及采取了何种操作。同时具备DDos基础防护功能，实时监控网络接入点的流量，如果发现攻击，会将源IP进行清洗或者列入黑名单，且支持白名单IP配置，从源头直接控制风险，同时具备拦截 SQL 注入、暴力破解攻击及其他类型的数据库攻击。

## *（三）学生主导表现*

### 3.1分工明确

以下是我们的分工情况：

成员A：搜集并筛选文本分类模型相关论文，搜集文本分类相关数据集。学习前端部分知识（HTML+CSS+JS等)，设计并制作分析系统前端页面

成员B：设计数据算法，筛选并分析文本过滤、文本分类模型论文，搭建评论过滤与文本分类模型，编写技术文档报告，搭建爬虫与模型、模型与后端部分连接

成员C：设计数据算法，筛选并分析文本过滤、文本分类模型论文，学习textCNN、ResNet相关知识，搭建评论过滤与文本分类模型，编写技术文档报告。

成员D：学习Java后端相关知识，搭建评论过滤与分析系统后端，搭建与模型、数据库、前端部分的连接，维护网页以及数据库管理。

成员E：搜集并筛选文本分类模型相关论文，搜集文本情感分类相关数据集。学习爬虫、数据库相关知识，编写爬虫部分算法以及搭建数据库。

### 3.2 独立开展

我们通过讨论决定每个人的分工情况，安排每个模块所要完成的相应的任务。然后我们通过学习，独立完成自己所要完成的工作。当遇到无法应付的技术难点时，我们首先会查阅相关的资料寻找相关的实现方案。如果无法解决，我们会通过小组讨论想出相应的解决方案，当讨论出现冲突时，我们会寻求指导老师的帮助。

### 3.3 积极讨论

我们小组每周五下午都会定期展开激烈讨论，设计整个项目逻辑，并制定每个部分的实现细节。在讨论会上，我们除了总结当前任务完成情况与下阶段任务分配，讨论在项目开发过程中遇到的问题外，我们还会制定下一步的计划，决定下一个开发周期所要做的工作，并定下时间期限，尽量确保在期限内完成任务。对于在会上提出的争议性较大的问题，我们会咨询指导老师，然后根据指导老师的意见并综合考虑计划安排，使我们的项目进程及完成情况符合预期。

# 五、管理战略

### 5.1项目概述

“网络评论过滤与分析系统”是应用于分析民众对于社会热点事件的情感分析的非盈利系统，具有积极把控舆论导向的性质。

#### 5.1.1 项目基本情况

网络评论过滤与分析系统，通过对网页评论内容的采集处理、敏感词过滤、智能分类、主题检测、专题聚焦、统计分析，实现各团体及个人对自己相关网络舆情监督管理的需要，最终形成舆情简报、舆情专报、分析报告、移动快报，为决策层全面掌握舆情动态，做出正确舆论引导，提供分析依据。

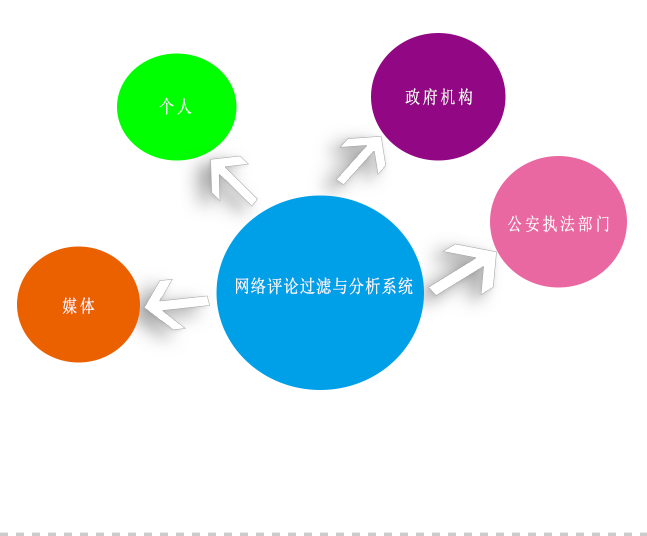
#### 5.1.2 项目主要特征

网络评论过滤与分析系统，在垃圾信息过滤方面，不同于市面上大多数基于关键字筛选进行过滤的系统，而是采用文本分类的技术，在关键字筛选的基础上进一步提高准确率，为大家提供文明的网络环境。目前市面上并没有单独的评论分析分类的相关项目，所以热点信息下面的评论一直鱼龙混杂，民众的准确态度也难以得知，因此，网络评论过滤与分析系统，将评论进行大数据分析，从而准确分析民众的态度。

#### 5.1.3 组织结构

目前团队内为五人小组，大体上可分为算法组两人和网页组三人。算法组负责设计数据算法与搭建文本分类模型，对评论进行分析，生成相关统计数据；网页组一人负责利用爬虫技术完成对热点事件评论的收集，一人负责分析成果网站后端的构建与维护，一人负责网站前端的构建与维护。

### 5.2利益相关群体



#### 5.2.1 个人

网络评论过滤与分析系统有助于帮助个人了解当前社会热点事件，以及大众的舆论导向，从而更全面的看待热点事件，避免被片面之词和有心人士误导，树立正确的个人价值观。

#### 5.2.2 媒体

网络评论过滤与分析系统有助于帮助媒体分析当前的热点事件，客观准确地报导舆论舆情，及时把控舆论导向避免误导人民价值观。也有助于帮助自媒体把握当下热点，以增加其知名度。

#### 5.2.3 政府机构

通过网络评论过滤与分析系统，政府机构可以及时了解网络舆情动态，关注到自己在网络舆情中的状态，从而可以产生网络舆情预警，及时纠正应对网络上关于自己的负面舆论影响，为部门网络危机公关或品牌形象营销提供数据依靠。同时把握好舆论导向，有利于掌握意识形态工作领导权，构建和谐社会。

#### 5.2.4 公安法制部门

通过网络评论过滤与分析系统，公安法制部门可以及时对危害社会安定的言论进行制止，对相关肇事者进行依法惩戒，以净化网络空间，维护社会的安定。

### 5.3 项目发展战略

#### 5.3.1 项目初期（2-4月）：基础功能实现

第一阶段注重完成基础功能的实现与调试。首要任务为实现文本分类模型的训练，并搜集数据集，进行各项功能的测试调整。其次对网站的搭建也十分重要：前端页面直接决定着用户的第一印象；爬虫的优劣影响着输入数据以及性能；后端联系各部分功能不可忽视。在项目初期阶段，需要不断汲取已有的文献以及项目经验，及时做出总结与更改、测试，直到形成较为成熟的网络评论过滤与分析系统。

#### 5.3.2项目中期（4-6月）：持续维护，进一步提高分析准确性

第二阶段开始对爬虫算法进行更新，爬取大量网页评论数据作为输入测试系统，同时对文本分类模型进行进一步优化，实现更精确、更完善的分析结果。开始对一部分（如在校学生）进行人为测试，记录用户体验反馈，总结反馈并基于此优化系统。之后开始大力推广此系统，使得其被大多数人所接受利用，实现物尽其用。此时需要团队人员时刻把控网站情形，预防未知因素对网络评论过滤与分析系统的影响。

#### 5.3.3项目成熟期（6-8月）：扩大规模，扩大服务对象

第三阶段稳定在校学生项目运行的基础上，扩大评论抓取途径，分析多种舆论导向，同时加大宣传，使得其适用范围更广。与时俱进，积极了解社会最新变化，秉承领先时代半步的宗旨，着眼于服务发展。根据新的形势与契机，推广新产品，新服务，提供新咨询、新信息，展现新面貌。

## *六、市场营销*

### 6.1目标市场

媒体、政府部门等社会舆论导向的引导及监管群体，以及对社会热点事件感兴趣的组织及个人。

### 6.2服务

#### 6.2.1 对组织及个人

帮助组织和个人了解当前社会热点事件，以及大众的舆论导向，从而更全面的看待热点事件，分析热点事件。

#### 6.2.2 对媒体和政府部门

帮助媒体客观准确地报导舆论舆情，帮助政府部门关注到自己在网络舆情中的状态，从而可以产生网络舆情预警，及时纠正应对网络上关于自己的负面舆论影响，为部门网络危机公关或品牌形象营销提供数据依靠。

### 6.3推广策略

在项目初期先在校园内进行小范围推广，并广泛收集意见和建议，进行进一步改善。在项目较为成熟之后，计划利用媒体进行推广，使得此系统被更多人熟知并使用。在项目完全成熟后，计划与微博、知乎等平台达成合作从而更方便地获取数据源，更精确地展示网络评论分析结果。

# 七、 成果应用情况

# **（一）成果应用情况**

网络评论过滤与分析系统即将正式上线web平台，将先在本校师生内预推广，吸取并采纳用户意见，再逐步推广至社会群体，以免费的形式提供给学生、团体、社会各界人士进行推广使用。

## *（二）应用前景*

项目网站目前搭建在“阿里云”服务器上，其安全性与效率得以保障，同时使用部署在服务器上的MySQL数据库，能够在保证用户体验的同时支持大量访问。

网络评论过滤与分析系统不仅能够满足一般用户的基本要求，对不同的用户又有着不用的适应性特点。其本身方便快捷、安全放心、功能完备。项目首页搜索页面简单易操作，支持多种网页格式并提供大量链接方便不同用户进行查询。后台管理页面能及时对最新舆论事件进行更新，查询并搜集评论数据，生成报告存入数据库，方便管理员整理以及用户需求。报告页面功能完备齐全，并具有一定可扩展性，大大节省了用户的时间和搜集查询的烦恼。此外其本身免费、方便等特征，更加吸引和迎合不同年龄段的用户。在校学生能够更好地分析和把握舆论动向，统计和分析评论用户的情感动向；诸如公关人员的社会人士能时刻把握评论动向，更好地应对和处理人员事件；政府部门等团体可以加强对网络评论的及时监测、有效引导，以及对网络舆论危机的积极化解，对维护社会稳定、促进国家发展具有重要的现实意义，这也是创建和谐社会的应有内涵。

综上本项目并未使用成本很高的软硬件设备，其本身免费、开放的特性将不断地吸引更多用户参与使用并进行功能优化提升，放在实际使用中可以预计会有较高的实用价值与社会效益。

# 八、研究心得

## *（一）创新思维与成长的收获*

首先非常感谢学校为我们提供了一次自主创新，深入课题研究的机会。在这次团队合作、自我锻炼、在实践中强化编码能力的过程中，我们收获了很多，也有一些心得体会。

在学习方面，通过这次项目强化了我们学习新知识的能力，使我们更深入地了解网站结构的搭建，使我们积累更多对模型的搭建与运用的经验，使我们的知识结构更全面。通过从相关论文中学习他人宝贵经验以及在其基础上进行改进与创新，这对于我们培养创新性思维和严谨的研究态度大有裨益。并且，在每一次调试参数的过程中，无论是尝试自己解决、组内交流，亦或是寻求网络资源及相关专业人士的帮助，都使我们积攒了许多自主学习自主解决问题的经验。

在团队合作方面，我们在课题初期需求并不完全确定的时候比较迷茫，成员间有着各自的想法，但对于是否能完整实现存在困惑。但在谢瑾奎教授的宏观指导和细节指点下，一次次会议让我们明确了需求和功能模块，也明确了努力的方向；同学之间不断交流彼此心得，互相帮助解决遇到的问题，不仅学习到很多知识，还让彼此间建立了深厚的友谊。这对于我们今后日常学习、团队合作乃至将来的日常工作都有着不可估量的启迪。

## *（二）可借鉴的经验*

在项目编码正式开始之前，首先最重要的是要把需求明确下来：包括功能模块的设计、前端页面的设计、模型的设计、数据库的设计、爬虫算法的设计，更细化的比如：代码包、类的方法和成员函数的设计等等，由于每位成员负责的板块有所不同，如果一开始这些问题比较模棱两可，并没有严格的规定，那么到了后期改动会比较麻烦，而且往往越底层的修改越麻烦。

编码过程中当自己原有的方法无法解决问题时，需要积极的学习新知识而不是拘泥于自己所熟悉的方法，只有这样才能推动项目的进展，同时也提高自身的技术水平。另外在编码过程中遇到困难时，除了从有关文献、资料中搜集现有的方法以外，及时与团队交流非常重要，有时候来自不同视角的看法可以提供新的解决方案，也可以避免网上一些良莠不齐的“答案”对我们学习的影响。在实际操作中我们秉承一个原则：既要避免“不懂装懂”，也要避免“一口吃成胖子”，这对我们的研究，甚至是将来的学习、工作、生活，都是有害无益的。

当项目编码内容基本完成后，测试以及部署是保证项目正常运行的极其重要的两个步骤。测试需要注意的问题是，集中测试后将需要修改的地方一起重新编码，并编以版本号，这样一来方便查看每个版本修改的问题。在部署的时候，要注意服务器的环境并做好记录，必要时需要做端口映射。

# 九、相关文献

[1]JINDAL N, LIU B.Analyzing and detecting review spam[C]//Proceeding of 7th IEEE International Conference on Data Mining ICDM 2007) .New York:IEEE, 2007:547-552

[2]刁宇峰,杨亮,林鸿飞.基于LDA模型的博客垃圾评论发现[J].中文信息学报,2011,25(01):41-47.

[3]MUKHERJEE A, VENKATARAMAN V.Opinion spam detection:an unsupervised approach using generative models[J].Technical Report, UH, 2014(07):1-11

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[8] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.

[9] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.