湖南大學

HUNAN UNIVERSITY

# 毕业论文

| | |
|---|---|
| **论文题目：** | 基于化妆品在线评论的爬虫设计与文本挖掘 |
| **学生姓名：** | 朱文静 |
| **学生学号：** | 201307040129 |
| **专业班级：** | 电子商务 2013 级 1 班 |
| **学院名称：** | 工商管理学院 |
| **指导老师：** | 江资斌 |
| **学院院长：** | 马超群 |

**2017 年 5 月 25 日**

# 湖南大学

# 毕业论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

学生签名：　　日期：20　年　月　日

# 毕业论文版权使用授权书

本毕业论文作者完全了解学校有关保留、使用论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权湖南大学可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本论文。

本论文属于

1、保密，在年解密后适用本授权书。

2、不保密。

（请在以上相应方框内打“√”）

学生签名：　　日期：20　年　月　日

导师签名：　　日期：20　年　月　日

# 摘要

伴随着网购人数的快速增长，中国电子商务交易额持续增长，随之而来的便是人们对电子交易平台商品信息可靠信的需求。基于 Web2.0 的社交平台使得消费者不再仅是产品信息的被动接受者，用户们可以在平台对相关服务或商品做出及时评价，为之后的消费者提供更加全面的信息。

本文就用户评价信息的文本挖掘这一主题进行研究，首先介绍了一般电商平台提供的评价机制，然后分析了关于网络爬虫和文本挖掘技术的国内外研究现状，使用 Python 设计出一款适用于一般电商平台的爬虫流程。并以京东上的某款化妆品为例进行了实验。

用户评价信息抓取完成后，接着就是从这些文本信息中挖掘出有价值的信息。本文采用 LDA 主题分析对商品的评价信息进行文本数据挖掘。首先需要对这些文本评论数据进行预处理：文本去重、机械语料压缩以及短句删除。然后使用 ROSTCM6 软件对处理过的文件进行情感分析，得到“正面情感结果”和“负面情感结果”，然后用 Python 的“jieba”中文分词包和 Gensim 库分别对这两个文本文件进行分词处理和 LDA 分析。最后根据实验产生的数据得到相关的结论并分析研究中遇到的问题与有待完善的地方。

现阶段人们对于评论数据挖掘的研究，国内外还尚未成熟，还有许多有价值的地方值得我们研究，我相信国未来随着自然语言处理、文本挖掘技术、机器学习等技术的发展，用户评价数据挖掘将会产生更多对人们有价值的信息。

**关键词：在线评论；文本挖掘；爬虫；Python；LDA 主题分析**

# ABSTRACT

With the rapid growth in the number of online shopping, China's e-commerce transactions continued to grow, followed by person on the electronic trading platform for reliable information on the needs of commodity information. The Web 2.0-based social platform allows consumers to no longer be passive recipients of product information, and users can make timely comments on relevant services or products on the platform to provide more comprehensive information to consumers later.

In this paper, the author reviews the topic of text mining for user evaluation. Firstly, it introduces the evaluation mechanism provided by the general electronic business platform, after that analyzes the research status of web crawler and text mining technology at home and abroad.This text use Python language as the general electronic business platform crawler process, and use Jingdong's cosmetics as an example.

After the completion of the crawl of user evaluation , we can dig out the valuable information from the text. In this paper, LDA theme analysis is used by text data mining of evaluation . First of all, these text data need to be pre-processed: text to weight, mechanical corpus compression and phrase deletion. And then use the ROSTCM6 software to analyze the processed files, get the "positive emotional results" and "negative emotional results", and then use Python's "jieba" Chinese word segmentation and Gensim library respectively on the two text file word processing and LDA analysis.

Finally, according to the experimental data generated by the relevant conclusions ,we can analysis the problems encountered in the study and to make some improve. At present, there is a lot of valuable places worthy of our study, I believe that the future of the country with the natural language processing, text mining technology, machine learning and other technical development, user evaluation Data mining will produce more information about people.

**Keywords: online customer evaluation; text mining; Internet worm; Python; LDA theme analysis**

# 目录

# 插图索引

# 附表索引

# 一、绪论

## （一）研究背景

### 1.电子商务的发展

互联网和信息技术的飞速发展，将商品交易市场演变为实体市场和互联网虚拟市场，以网络交易为主体的电子商务使互联网虚拟市场得到快速发展。

1990-1993 年的电子数据交换时代，成为了中国电子商务的起步期。经过五年的雏形期，1998 年 3 月，中国第一笔互联网网上交易成功。此后电子商务在我国蓬勃发展[1]。2011 年，中国网络购物用户规模达 1.94 亿，用户渗透率由上年的 35.6%升至 37.8%;截至 2012 年，中国网络购物用户规模达到 2.42 亿人，网络购物使用率提升至 42.9%。到 2013 年，中国网络购物用户规模达 3.02 亿人，团购用户规模达 1.41 亿人。2014 年，中国网络购物用户规模为 3.61 亿人，网民使用率达到了 55.70%，较 2013 年底增加 6.8 个百分点。

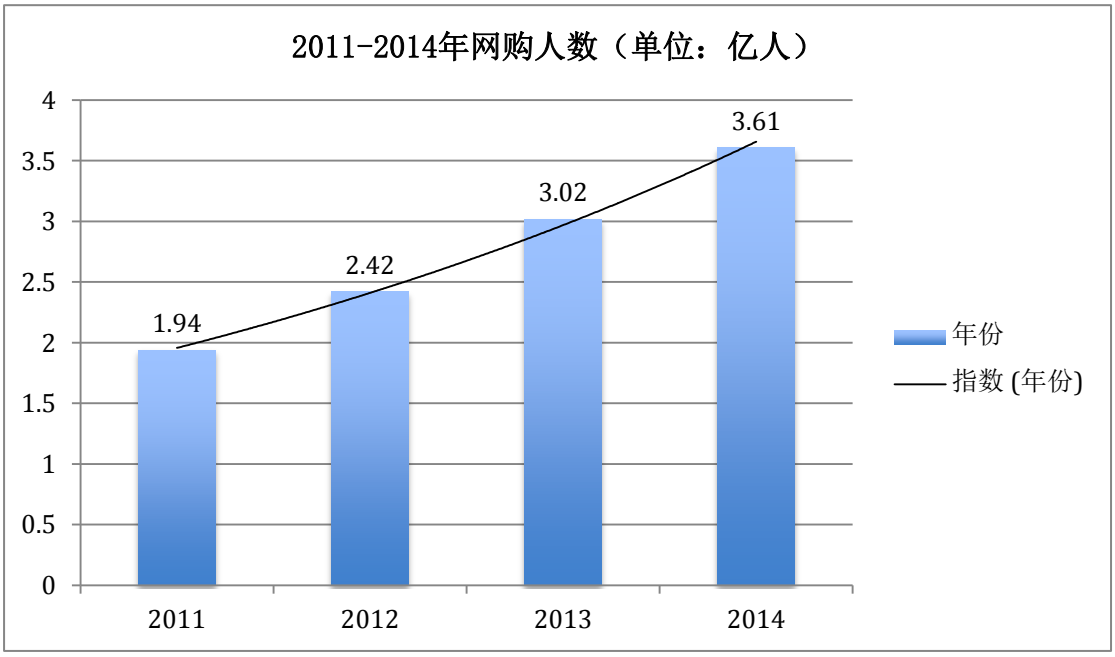


图 1 2011-2014 年网购人数增长

伴随着网购人数的快速增长，中国电子商务交易额持续增长，2009 年中国的电子商务交易额为3.7万亿元，到2012年增加至7.85万亿，同比增长30.83%。其中，B2B 电子商务交易额达 6.25 万亿，同比增长 27%。网络零售市场交易规模达 13205 亿元，同比增长 64.7%。2013 年中国电子商务交易额为 10.50 万亿元。2014 年，中国电子商务交易额 13 万亿元，同比增长 23.81%[2]。

### 2.在线用户评论

基于 Web2.0 的社交平台使得消费者不再仅是产品信息的被动接受者，用户

们可以在平台对相关服务或商品做出及时评价，为之后的消费者提供更加全面的信息。Mudambi 较早在《What makes a helpful online review? A study of customer reviews on amazon.com》一文中提出了在线评论有用性这一概念，对消费者的购买行为中对于在线评论信息的采纳程度进行衡量[5]。

在线评论作为网络口碑的一种，一般指潜在或实际消费者在电子商务或第三方评论等网站上发表商品或服务的正面或负面观点。一般情况下，商家创造的产品信息一般会隐藏产品潜在的缺陷和不足,只展示产品的优势。而在线评论从用户角度出发，对产品做出相对客观的评价，因此内容会更加真实可靠[6]。

根据国际著名市场研究公司 Jupiter Research 的调查，超过 75%的消费者在线购买商品之前会参考在线评论信息[7]。特别是在当下数据爆炸的环境下，有价值的用户评论可以帮助消费者减少购买的不确定性，做出更好的购买决策，从而提高消费者对在线评论网站的粘性。因此，关于在线评论有用性的研究具有非常重要的理论与实践意义。

### 3.文本挖掘

互联网的便利,用户可获得的信息包含了从技术资料到娱乐资讯等各种各样的文档，从而构成了一个庞大的分布式数据库。这个数据库具有异构性、开放性等特点，而且存放的是非结构化的文本数据。想要从这些非结构化的文本数据中获取有价值的信息,必须结合人工智能研究领域中的自然语言理解和计算机语言学，于是从数据挖掘中派生了两类新兴的数据挖掘研究领域：网络挖掘和文本挖掘。

文本挖掘最早由 Ronen Feldman 等人提出，其作为一个新的数据挖掘领域，其目的在于把文本信息转化为人可利用的知识。人们利用智能算法，并结合文字处理技术,分析大量的非结构化文本源,抽取或标记关键字概念、文字间的关系，并按照内容对文档进行分类，从而获取有用的知识和信息[10]。

目前研究和应用最多的几种文本挖掘技术有：文档聚类、文档分类和摘要抽取[12]。利用文本挖掘技术处理大量的文本数据，无疑将给企业带来巨大的商业价值。因此，目前对于文本挖掘的需求非常强烈，文本挖掘技术应用前景广阔。

## （二）研究目的和意义

消费者通过网上购物，节省了大量时间和空间，提高了交易效率。但另一方面，由于消费者只能从线上浏览商品，不能实物感知商品，导致了卖家与买家之间的信息不对称。所以通常情况下，买家在决定买一件商品时会先阅读之前买家的在线评论，然后根据评价的好坏再决定是否购买。由此可见，在线评论与一件商品的销售量存在很大的关系。

现阶段电商平台的评价种类大致有信用积分、动态评分、标签评价以及内容评价[14]。而用户关注最高的应该是内容评价。而内容评价不仅仅局限于购买产品后及时评价，一般有很多种类型，以淘宝为例，如下表 1 所示。

表 1 电商平台评价种类

| 评价方式 | 评价特点 |
| --- | --- |
| 售后即评 | 一般是用户刚收到商品时做出的回复。 |
| 追加评价 | 用户在使用的过程中对商品出现各种问题追加的评论。 |
| 默认好评 | 未评价的用户在一定期限后系统会默认为好评。 |
| 双向评价 | 买方卖方相互评论。 |
| 定向提问 | 指定买过的用户进行提问。 |
| 好评有礼 | 商家通过“好评返现”的刺激用户好评。 |
| 退款评价 | 买方退款成功后作出的评价。 |

由上表可知，电商平台是鼓励用户评价的。此外，不少电子商务网站在提供撰写在线评论功能的同时，也设置有用无用投票选项，并会向浏览用户显示有用投票比例。

但目前的评价机制还是存在很多的问题，例如由于信息零碎，在线评论的作用没有得到充分利用。对于一个少则几百上则几万的产品评论而言，没有一个人是愿意一直看评论的，所以差评、图评、追评是大家追逐的重点，同时这部分内容也是最少的，很多评价的价值未被挖掘。而且目前的标签抓取技术还是很差，容错率很低，所以经常误解评价出错，因此对于在线评论的文本数据挖掘显得尤为重要。

针对这一情况，本文欲设计一个专门爬取电商交易平台用户评论的爬虫算法，对爬取的数据进行处理后进行文本数据挖掘，并以京东上化妆品牌“珂润”的一款面霜作为研究对象，探索大量评论后隐藏的有用信息，为商家和买家提供有价值的信息。

## （三）研究综述

### 1. 网络爬虫

#### （1）爬虫的文献综述

孙立伟、何国辉和吴礼发按照系统结构和实现技术，将网络爬虫大致分为以下几种类型：通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层网络爬虫[16]。实际的网络爬虫系统通常是几种爬虫技术相结合实现的。而因为聚焦爬虫在爬取过程的高效性，现在的研究主要集中在聚焦爬虫。

聚焦爬虫爬行策略实现的关键是评价页面内容和链接的重要性，不同的方法计算出的重要性不同，由此导致链接的访问顺序也不同。目前主要有基于内容评价的爬行策略、基于链接结构评价的爬行策略、基于经验的爬行策略和一些其他的爬行策略[18]。

在基于内容评价的爬行策略中，主要利用抓取网页中的锚文本、文本内容、URL 字符串信息。DeBra 采用了文本相似度的计算方法并提出了“鱼群搜索策略”（Fish Search）算法，在鱼群搜索策略中，每个网页相当于一条鱼，当它们发现相关信息时，这些鱼就沿链接方向继续寻找相关页面，若与主题不相关则放弃

当前链接。之后 Herseovic 对 Fish Search 算法进行了改进，引进了相似度度量方法，提出了“鲨鱼搜索策略”（Shark- search）算法，利用空间向量模型计算页面与主题的相关度大小。后来，Menczer F 等人又提出了“最佳优先”搜索策略，这一策略通过计算向量空间的相关性，把相关性“最好”的页面放入最优先下载的队列。但是该算法只能找到局部范围内的最优解，难以得到全局范围内的最优解。

基于链接结构评价的爬行策略，主要以 PageRank 和 HITS 算法为代表。PageRank 的概念是每个到页面的被链接得越多，就意味着被其他网站投票越多。HITS 算法通过两个评价权值—内容权威度和链接权威度来对网页质量进行评估。

基于经验的爬行策略，在经验爬行方法中，主题爬虫在再次爬行中能够参考先前爬行的经验，从而过滤不相关链接，并且通过对知识库的不断完善，使得对主题的定义更加准确，更易发现主题相关网页，当然在若想取得理想的效果，必须要加大训练集。

### （2）爬虫的应用领域

根据目前已有的文献可知，研究者用爬虫解决了不同领域的实际问题，如互联网舆情监控、负面情绪研究、电商网站商品评论收集和房地产行业等。如方星星、鲁磊纪、徐洋使用爬虫技术实现网络舆情监控系统的实现[19]，张明杰通过爬虫技术，设计了基于新浪微博的网络舆情数据采集系统[20]。彭纪奔、吴林等人在 Scrapy 爬虫框架下，实现对目标网络的周期性自动抓取，并对网页内容中包含的负面情绪进行度量，提出一个网络负面情绪挖掘系统 CyberCare[21]。周中华、张惠然、谢江为了快速采集到微博中的数据，设计了一款可以实现并行抓取微博数据爬虫[22]。董浩然、谢欢等人基于 GIS 主题爬虫，实现了一个可以实现房产地理位置查询、房屋基本属性查询、房屋价格估算、房产批量评估等功能的房产信息实时更新与处理系统的爬虫工具[23]。卢长宝和庄晓燕，将爬虫运用到了餐饮业，通过实证研究证实了 SERVQUAL 模型评测大众餐饮行业服务质量的适用性[24]。邓宏勇,许吉,张洋,袁敏,施毅分析了爬虫技术在中医药领域的运用情况[25]。

## 2. 文本挖掘

### （1）文本挖掘的文献综述

国外在文本挖掘的领域起步较早，50 年代末，H.P.Luhn 提出了词频统计思想，将之用于自动分类上。随后在 1960 年，Maron 发表了第一篇关于自动分类的文章。接着众多学者在这此领域进行研究工作，研究主要有围绕文本的挖掘模型、文本特征抽取与文本中间表示、文本挖掘算法(如关联规则抽取、语义关系挖掘、文本聚类与主题分析、趋势分析)、 文本挖掘工具等。目前，国外的文本挖掘研究已经从实验性阶段进入到实用化阶段，著名的文本挖掘工具有：IBM 的文本智能挖掘机、 Autonomy 公司的 Concept Agents、TelTech 公司的 TelTech 等[26]。

我国是近几年才正式引入文本挖掘这个概念，并开始关于中文的文本挖掘研究。从已有的研究成果看，目前我国文本挖掘研究还处在向国外相关的理论和技术学习的阶段，尚未形成系统的适合中文信息处理的文本挖掘理论与技术框架。在技术手段方面，也主要是借用国外针对英文语料的挖掘技术，没有完备的中文

信息处理与分析技术来构建针对中文文本的文本挖掘模型，很大程度上限制了中文文本挖掘的发展。

### （2）文本挖掘的应用领域

文本挖掘运用的领域范围广泛，就本人所查阅的文献就有数字人文、人文社会、生物医学、网络舆情监控等诸多领域。

郭金龙、许鑫总结了数字人文中领域中文本挖掘的应用。通过文本挖掘技术，对文章风格特征的分析来辅助鉴定作者身份。国外研究者除了利用各种分类和统计算法对作者归属进行了大量的研究外，还利用文本挖掘对作家的性别特征进行了挖掘，如 Helma 等对古希腊作品中的作者性别进行了分析研究，Shlomo 等对法国文学中男女作家的性别进行了研究，得出了男女作家不同的写作风格的定量佐证[28]；郭金龙、许鑫、陆宇杰总结了人文社会研究中文本挖掘的应用，研究者运用文本挖掘技术进行情感分析、热点发现、可视化技术等[29]；王浩畅，赵铁军总结的生物领域文本挖掘技术的研究中，文本挖掘提高了生物医学命名实体识别的效率，辅助进行不同实体间的关系抽取。目前国内在生物医学文本挖掘领域的研究相对还比较少 ，清华大学研究者在蛋白质关系抽取方面做了深入研究，哈工大研究人员主要致力于生物医学命名实体识别和关系的识别的研究[30]；黄晓斌，赵超研究的网络舆情方面，通过建立舆情信息的文本挖掘系统，对网络舆情信息进行分析，生成针对某一社会公共事件民众存在的不同的情绪和观点[31]。孟雪井，孟祥兰，胡杨洋利用文本挖掘探索百度指数投资者的情绪指数[32]；吴恒，陈燕翎采用文本挖掘技术，以携程蜜月游记为研究对象，探索对游客旅游时如何对目的地做出选择[33]；张玉峰，朱莹将文本挖掘运用到了企业竞争情报的获取[34]；熊伟，郭扬杰用文本挖掘，对酒店顾客的在线评论信息做出一系列研究[35]。

## （四）论文框架

本论文总共有五章，逐步介绍了如何利用爬虫与文本挖掘技术，然后就电商交易平台中的化妆产品进行数据挖掘。

第一章为绪论，说明了本文的选题背景、研究目的和意义与论文整体框架。简单介绍了我国电子商务的近年来的发展，电商平台中在线用户评论的现状、文本挖掘的流程和技术，进而得出关于在线评论文本挖掘的应用价值。

第二章介绍了开发数据挖掘爬虫的流程。在对一个网站进行数据采集之前，首先需要对该网站进行背景调研，检查 robots 文件、网站地图以及估算网站大小。在了解了网站的基本信息后，接着就是进行数据采集，由于商品评论一般采用 JavaScript 脚本，所以我们需要先找到评论页的真正 URL，然后分析网页的结构，对数据进行采集和存储。

第三章介绍了如何对采集的评论数据进行文本挖掘。首先需要对文本数据进行文本去重，然后使用 ROSTCM6 进行情感分析，得到正负向情感结果并分别进行分词和 LDA 主题分析。

第四章是关于化妆品行业的应用分析，本文采取京东上一款面霜为例，对其进行商品评论数据采集、文本分析并得出相关结论。

第五章是总结和展望，对文章进行了总结，分析研究过程中遇到的难题以及

研究过程中存在的不足，为后来的数据文本挖掘提供一些有价值的结论，如图 2 所示。

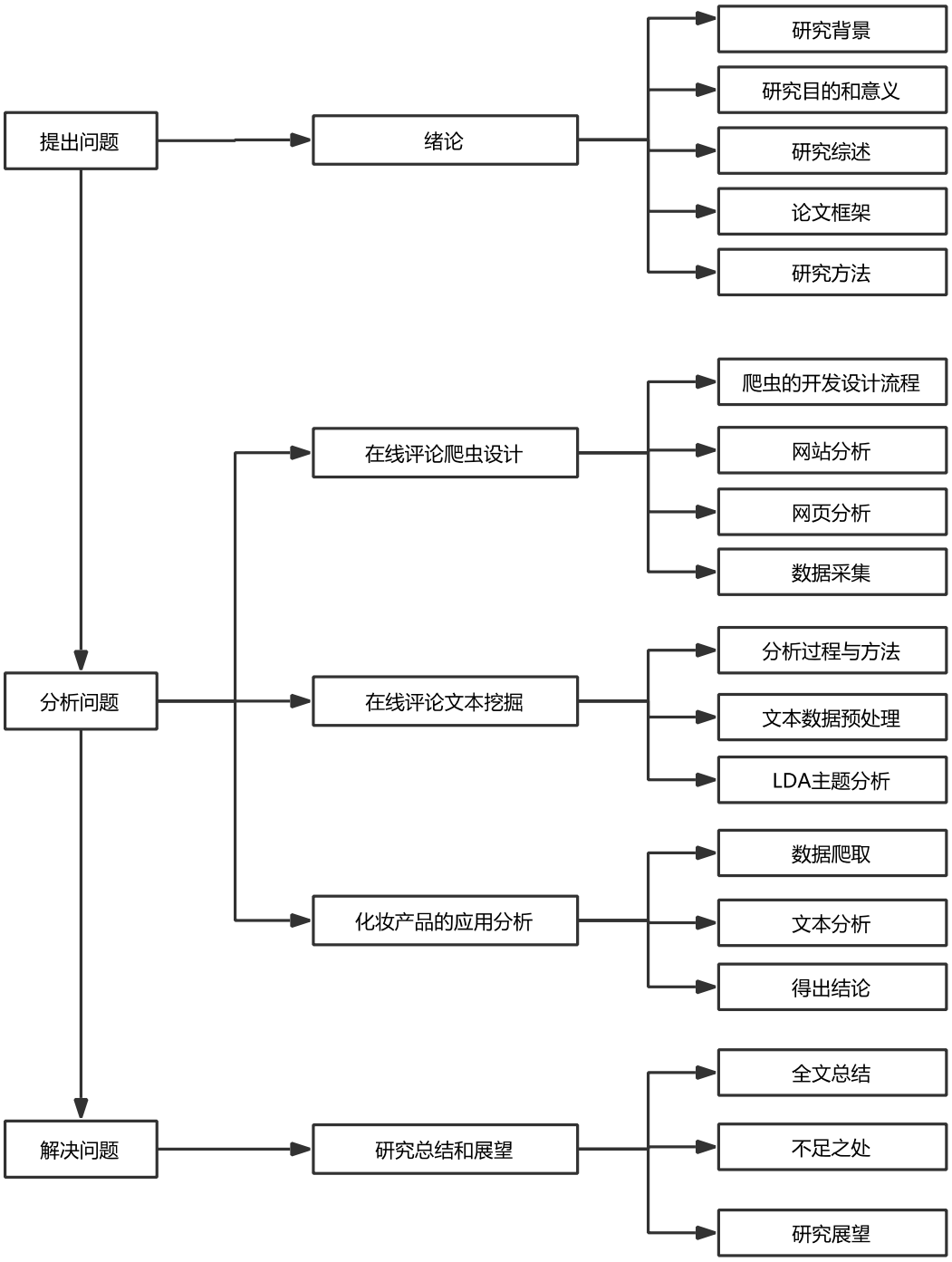


图 2 全文框架图

## （五）研究方法

### 1. 文献研究法

通过国内外科学文献数据库来获得与研究相关的调查文献，对在线评论发展、网络爬虫技术以及文本挖掘技术等文献进行收集与处理，从而全面地了解与掌握所要研究的问题。首先对研究对象就行较为明确的界定，而后对研究对象以及国内外研究现状进行综述，根据对研究现状的分析，提出解决本文研究问题的流程与方法，从而做到为之后的研究提供理论方面和实证方面的支持。

### 2. 信息研究方法

信息研究方法作为一种利用信息来研究系统功能的科学研究方法，揭示事物的更深一层次的规律，帮助人们提高和掌握运用规律的能力。本文通过网络爬虫爬取的数据，使用文本挖掘技术，对信息的收集、传递、加工和整理，从而挖掘出在线评论中隐藏的有价值的信息。

### 3. 个案研究法

个案调查是就研究对象中的某一特定对象，加以调查分析，弄清其特点及其形成过程的一种研究方法。本文根据分析研究后设计的研究思路，以化妆品行业为例，就某一品牌的某一商品进行实际的调查研究，从而得到有针对性的研究结果。

# 二、 在线评论爬虫设计

## （一）爬虫的开发设计流程

爬虫的设计开发，大致可以分为三步：网站分析、网页分析和数据采集，具体流程如图 3 所示。

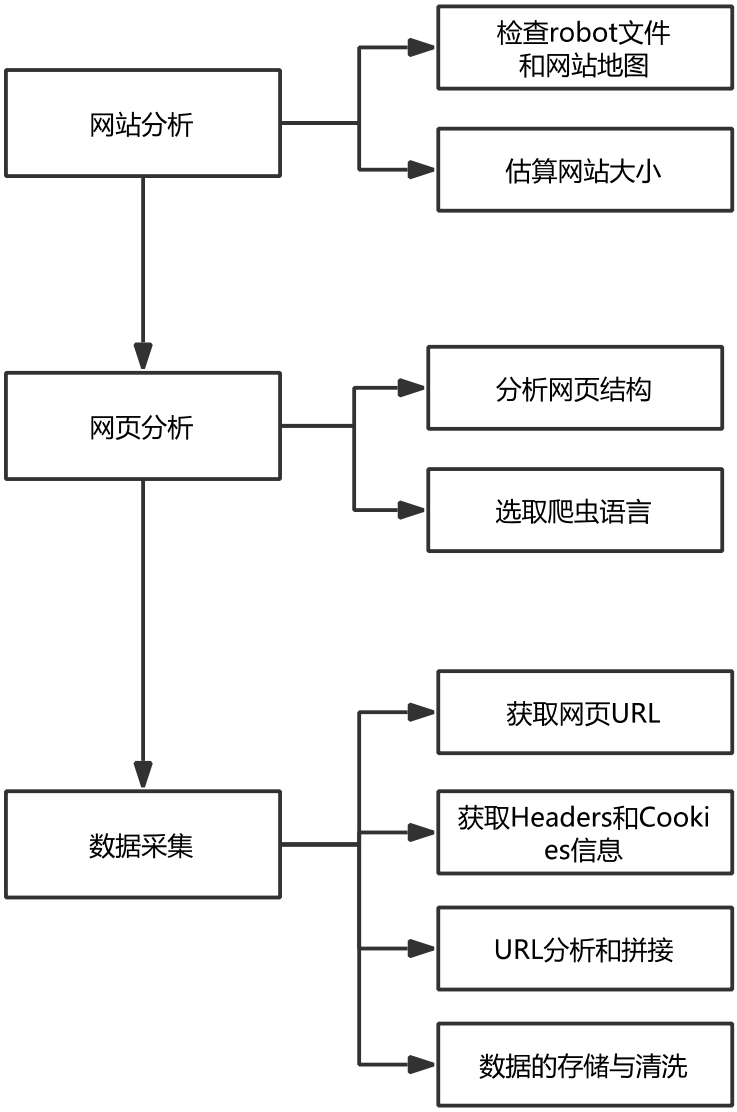


图 3 设计流程图

## （二）网站分析

在爬取一个网站前，我们需要对目标站点的规模与结构进行一定的了解。这时可以通过网站自身的 robots.txt 和 Sitemap 文件，为我们提供帮助。

### 1.检查 robots.txt 文件和网站地图

大多数网站会定义 robots.txt 文件，这样可以了解爬取该网站时都存在哪些限制。爬取前检查 robots.txt 文件，可以最小化爬虫被禁的可能，而且还可以发现和网站结构相关的其他线索。此外，网站的 Sitemap 文件可以帮爬虫定位到网站最新的内容，这样就不用爬取每一个网页，使得爬取工作更加高效。

### 2.估算网站大小

目标网站的大小决定我们如何进行数据的爬取，而估算一个网站的大小，简便的方法就是检查 Google 爬虫的结果：通过 Google 搜索的 site 关键词，过滤域名结果从而获取该信息。因为在我们进行数据爬取之前，Google 很可能已经爬取过我们感兴趣的网站。

## （三）网页分析

分析完网站后，我们还要对网页进行分析，然后根据网页语言的特点，选取合适的爬虫语言。

### 1.分析网页结构

通常我们通过浏览器所看到的页面信息，是由很多的页面元素组装在一起的，其中有我们可以实际看到的图片和文字，也有我们看不到的布局结构（div、tr、td 等），所有的这些页面元素组织成了我们能够看到的 Page 页面。而当爬虫进行爬取时，它看到的是包含着图片文字与布局结构的一堆 html 代码，这些 html 代码我们可以使用浏览器中的“查看源代码”看到。所以在对用户评论进行爬取工作之前，我们需要对电商网页的结构特点进行了解，然后根据其特点匹配相应的技术。

目前，大多数电商网站中商品评论的部分是使用 JavaScript 技术实现的。JavaScript 一种直译式脚本语言，广泛用于客户端的脚本语言，用来给 HTML 网页增加动态功能，为用户提供更流畅美观的浏览效果。通常 JavaScript 脚本是通过嵌入在 HTML 中来实现自身的功能的。

### 2.选取爬虫语言

所以基于以上分析，本研究拟采用 Python 语言开发网络爬虫。与其他语言相比，Python 的优点是方便高效。

其一是在网页抓取本身的接口时，相比其他静态编程语言，如 java，c#，C++,Python 抓取网页文档的接口更简洁;相比其他动态脚本语言,如 perl,shell,Python 的 urllib2 包提供了较为完整的访问网页文档的 API。此外，抓取网页有时候需要模拟浏览器的行为，很多网站对于生硬的爬虫抓取都是封杀的，需要我们模拟 user agent 满足合适的请求，譬如模拟用户登陆、模拟 session/cookie

的存储和设置。这些在 Python 里都有很优秀的第三方包可以实现，如 Requests，mechanize。其二是网页抓取后，抓取的网页一般需要处理，比如过滤 html 标签，提取文本等。Python 的 Beautifulsoup 提供了齐全的文档处理功能，帮你用很少的代码完成大部分文档的处理[36]。

## （四）数据采集

### 1. 获取爬取网页的 UIL

一般网页中商品评论信息是由 JavaScript 脚本动态加载的，所以直接抓取商品详情页的 URL 信息，并不能得到商品评论的信息。这时可以使用 Chrome 浏览器里的开发者工具，对商品评论信息的文件进行查找。具体操作是在开发者工具中选择 Network，设置为禁用缓存和只查看 JS 文件，然后刷新页面。页面加载完成后找到商品评价部分，等商品评价信息全部显示后，在 Network 界面的左侧筛选框中输入 productPageComments（这里以京东平台为例），此时下面的加载记录中只有一条信息，这就是我们要抓取的 URL 地址，里面包含的就是商品详情页的商品评论信息。

### 2. 将爬虫伪装成浏览器

为了防止爬虫被禁，我们需要设置头文件和 Cookie 文件。头文件信息比较容易找到，在 Chrome 的开发者工具中选择 Network，刷新页面后选择 Headers 就可以查看到本次访问的头文件信息，头文件信息的旁边还有一个 Cookies 标签，里面就是本次访问的 Cookies 信息。

### 3. 评论页面的结构分析

第一步获取的 URL 中包含两个重要信息，一是商品 ID，另一是页码。因为实验中我们只抓取一个商品的评论信息，所以商品 ID 不需要更改。我们的目的在于抓取商品的评价信息，所以商品评论的页码不是固定值。所以我们可以将获取的 URL 分成两部分，通过生成随机页码，然后拼接 URL 的方式对评论数据进行抓取。实际操作是在抓取过程使用 for 循环，每次循环都从随机数中产生一个生成页码编号，然后两部分的 URL 进行拼接，生成要抓取的 URL 地址。然后与前面设置的头文件信息和 Cookie 信息一起发送请求，从而获取页面信息。为了避免频繁的请求导致返回空值，所以每次请求休息间隔设置为 5s。

### 4. 数据的存储与清洗

抓取完的数据可以以 CSV 的格式存储在本地，接下来的工作就是对这些爬取的数据进行清洗和分析工作。导入 re 包，使用正则表达式对我们需要的‘content’字段进行提取和清洗。完成字段信息的提取与清洗后，再将这些字段生成京东商品评论数据 CSV 格式的汇总表。

# 三、在线评论的文本挖掘方法

## （一）分析方法与过程

数据采集成功后，接下来需要对采集到的数据进行文本挖掘，从而得到用价值的信息。分析流程如图 4 所示。

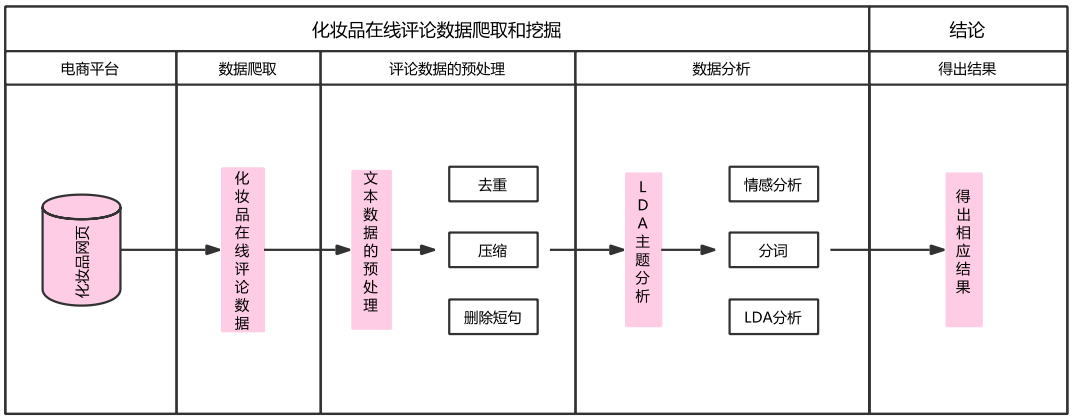


图 4 文本挖掘流程图

## （二）文本评论数据的预处理

初步获取的文本评论数据中，存在许多重复或者没有价值的信息，如果把这些数据也进行分词、词频统计以及情感分析等工作中，则会对分析的结果产生误差，得到的结论质量会降低。所以对这些文本评论数据进行分析之前，必须先进行文本预处理，将其中夹杂的的无价值含量的评论去除。

文本评论数据的预处理主要由三部分组成:文本去重、机械语料压缩以及短句删除。这里主要讲解文本去重和机械语料压缩[44]。

### 1. 文本去重

在前人的研究中，许多的文本的去重算法都是先通过计算文本间的相似度，再以此为基础进行去重，方法包括编辑距离去重、Simhash 算法去重等等，但是大多存在一些缺陷。拿编辑距离算法来说，编辑距离算法去重实际上就是先计算两条语料的编辑距离，然后进行阈值判断，若编辑距离小于某个阈值则进行去除重复处理。这种方法对于重复度较高而又无意义的评论文本，去除效果是很好的，但是当这种方法测到都有意义，但是表达方式相近的时候，可能也会采取删除操作，从而造成错删问题。

因为编辑距离去重的算法容易误删有用的数据，所以在处理一些相对简单的

文本时，相近的语料之间有不少有用的评论，去除这类语料显然不合适，所以为了留下更多的有用语料，只能从完全重复的语料下手。采用两两对比的方法，完全相同就去除。

由此可知，存在文本重复问题，其本质是此语料是否有用。通过观察评论判断是否重复，但是最终会保留一条评论有用的语料。而运用比较删除法只能留一条或者全去除，因此只能设为留一条，以确保尽可能存留有用的文本评论信息。

### 2.机械语料压缩

由于电商平台的文本评论数据的参差不齐，无意义的文本数据很多，因此仅仅进行文本去重还不够，经过文本去重后的评论仍然需要加工处理。机械压缩去词实际上就是去除语料中有连续重复的部分。

从一般的评论偏好讲，评论中无意义的连续重复大多出现在开头或者结尾。连续重复的判断，可以通过建立两个存放国际字符的列表实现。先放第一个列表，再放第二个列表，逐一读取国际字符，并按照不同情况，将其放入第一第二个列表或触发压缩判断，若判断重复则进行压缩去除。在机械压缩去词处理连续重复时，判断和压缩规则的设定必然还要考虑词法结构的问题。

## （三）文本评论数据的 LDA 主题分析

LDA 模型也被称为三层贝叶斯概率模型，包含文档（d）、主题（z）、词（w）三层结构，能够有效对文本进行建模，和传统的空间向量模型（VSM）相比，增加了概率的信息。通过建立 LDA 主题模型，能够挖掘数据中的潜在主题，进而分析数据集的集中关注点及其相关特征词。

虽然 LDA 可以直接对文本进行主题分析，但是如果文本的正面评价和负面评价混淆在一起，由于分词粒度的影响（否定词或程度词等），可能一个主题下会生成一些令人迷惑的词语。因此在进行 LDA 主题分析之前，最好将文本分为正面评价和负面评价两个文本。

### 1.文本分词处理

中文中，字、句和段落可以通过分界符进行划界，而对于“词组”来说，它们的边界模糊，没有一个形式上的分界符。因此进行中文文本挖掘时，需要对文本进行分词处理，即将连续的字序列按照一定的规范重新组合成词序列的过程。

本论文中使用的是 Python 的中文分词包“jieba”，提供分词、词性标注、未登录词识别，支持用户词典等功能，可以实现对 TXT 文档中的商品评论数据进行中文分词。

### 2.情感倾向性模型

#### （1）训练生成词向量

为了将文本情感分析转化为机器学习问题，首先需要将符号数学化。在 NLP

中，最常见的词表示方法就是 One-hot Representation，即将一个词映射成一个很长的单位向量，向量的长度就是词表的大小。

### （2）评论集子集的人工标注与映射

利用词向量构建的结果进行评论集子集的人工标注，正面评论为 1，负面评论为 2，然后将每条评论映射为一个向量，将分词后评论中所有词语对应的词向量相加后做平均，使得一条评论对应一个向量。

### （3）训练栈式自编码网络

自编码网络是由原始的 BP 神经网络演化而来。原始的 BP 神经网络中从特征空间输入到神经网络中，并用类别标签与输出空间来衡量误差，用最优化理论不断求得极小值，从而得到一个与类别标签相近的输出。但编码网络是用从特征空间的输入来衡量与输出空间的误差。

## 3. LDA 主题分析

### （1）LDA 主题模型介绍

LDA 模型采用词袋模型，将每一篇文档视为一个词频向量，从而将文本信息转化为易于建模的数字信息。定义词表大小为 L，一个 L 维向量(1，0，0，…，0，0)表示一个词。由 N 个词构成的评论记为 $d=(w_1, w_2, \cdots, w_N)$。假设某一商品的评论集 D 由 M 篇评论构成，记为 $D=(d_1, d_2, \cdots, d_M)$。M 篇评论分布着 K 个主题，记为 $z_i(i=1,2,\cdots, K)$。记α 和β 为狄利克雷函数的先验参数，θ 为主体在文档中的多项分布的参数，其服从超参数为α 的 Dirchlet 先验分布，Φ为词在主题中的多项分布的参数，其服从超参数β 的 Dirichlet 先验分布。

LDA 模型假定每篇评论由各个主题按一定比例随机混合而成，混合比例服从多项分布，记为：

$$Z|\theta = \text{Multionmial}（\theta） \qquad 公式（1）$$

而每个主题由词汇表中的各个词语按一定比例混合而成，混合比例也服从多项分布，记为：

$$W|Z，\emptyset = \text{Mutionomial}（\varphi） \qquad 公式（2）$$

在评论 $d_j$ 条件下生成词 $w_i$ 的概率表示为：

$$P\left(w_i|d_j\right) = \sum_{S=1}^{K} P（w_i|z = s） \times P（z = s|d_j） \qquad 公式（3）$$

其中，$P(w_i|z=s)$ 表示词 $w_i$ 属于第 s 个主题的概率，$P(z=s|d_j)$ 表示第 s 个主题在评论 $d_j$ 中的概率。

### （2）LDA 主题模型估计

LDA 模型对参数θ 、Φ的近似估计通常使用马尔科夫链蒙特卡洛算法中的一个特例 Gibbs 抽样。利用 Gibbs 抽样对 LDA 模型进行参数估计，依据下式：

$$P（z_i = s|Z_{-i}, W） \propto （n_{s,-i} + \beta_i） \div （\sum_{i=1}^{V} n_{s,-i} + \beta_i） \times \left(n_{s,-j} + \alpha_s\right)$$

公式（4）

其中，$z_i$=s|表示词 $w_i$ 属于第 s|个主题的概率，$Z_{-i}$ 表示其他所有词的概率，$n_{s,-i}$ 表示不包含当前词 $w_i$ 的被分配到当前主题 $Z_s$ 下的个数，$n_{s,-j}$ 表示不包含当前 $d_j$ 的被分配到当前主题 $z_s$ 下的个数。

通过对上式的推导，可以推导词 $w_i$ 在主题 $z_s$ 中的分布的参数估计$\Phi_{s,j}$，主题 $z_s$ 在评论 $d_j$ 中的多项分布的参数估计$\theta_{j,s}$，如下：

$$\emptyset_{s,i} = （n_{s,i} + \beta_i） \div （\sum_{i=1}^{V} n_{s,i} + \beta_i）$$ 公式（5）

$$\theta_{j,s} = （n_{j,s} + \alpha_s） \div （\sum_{s=1}^{K} n_{j,s} + \alpha_s）$$ 公式（6）

其中，$n_{s,i}$ 表示词 $w_i$ 在主题 $z_s$ 中出现的次数，$n_{j,s}$ 表示文档 $d_j$ 中包含主题 $z_s$ 的个数。

### （3）运用 LDA 模型进行主题分析的实现过程

商品评论关注点的研究，即对评论中的潜在主题进行挖掘。一般来说，每则评论中都存在一个主题。如果一个潜在主题同时是多则评论中的主题，则这一潜在主题很可能是整个评论语料集的热门关注点。在这个潜在主题上特征词词频越高，就越可能成为热门关注点中的评论词。

为提高主题分析在不同情感倾向下，热门关注点的精确度，在情感分类结果的基础上，对不同情感倾向下的潜在主题分别进行文本挖掘，从而得到不同情感倾向下的用户情况。接着分别统计整个评论语料库中，正负情感倾向的主题分布情况，并在两种情感倾向下，对各个主题出现的次数从高到低进行排序，根据需要，选择排在前面的主题作为评论集中的热门关注点，然后根据潜在主题特征词的概率分布，得到对应的热门关注点的评论词。

## （四）本章小结

LDA 主题模型在文本聚类、主题挖掘、相似度计算等方面都有广泛的应用，相对于其他主题模型，此模型的泛化能力较强，不易出现过拟合现象。其次它是一种无监督的模式，只需提供训练文档就可以自动训练出各种概率，无需任何人工标注过程，节省大量人力及时间。但要注意的是，研究过程中，分词结果的准确性对后续文本挖掘算法有很大的影响，如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。

# 四、化妆品在线评论的文本挖掘

## （一）爬取数据

此次论文进行实证研究的对象是京东平台上的花王珂润官方旗舰店的“珂润润浸保湿滋养乳霜 40g”商品，第三章分析过，京东商品评论信息是由 JavaScript 动态加载的，所以直接抓取商品详情页的 URL 并不能获得商品评论的信息。因此我们需要使用 Chrome 浏览器里的开发者工具进行查找。

### 1．设置头文件和 cookie

为了在爬取数据的时候不被查封，我们还需要对爬虫进行伪装，使爬虫看起来更像是来自浏览器的访问。这里主要的两个工作是设置请求中的头文件信息以及设置 Cookie 的内容。

具体方法是打开开发者工具，在开发者工具界面中选择 Network，设置为禁用缓存和只查看 JS 文件，然后刷新页面。页面加载完成后向下滚动鼠标找到商品评价部分，并在 Network 界面找到 productPageComments，这里包含的就是商品详情页的商品评论信息。

点击此信息，在右侧的 Headers 标签，可以看到其中包含了当前页面中的头文件。


图 5 Headers 信息

```
#设置头文件参数
headers = {
    'User-Agent': ' Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.87 Safari/537.36',
    'Accept': '*/*',
    'Accept-Language': 'zh-CN,zh;q=0.8,en;q=0.6',
    'Connection': 'keep-alive',
    'Referer': 'https://item.jd.com/1750036.html'
}
```

图 6 Headers 代码说明

在查看头文件信息的旁边还有一个 Cookies 标签，点击进去就是本次访问的 Cookies 信息。

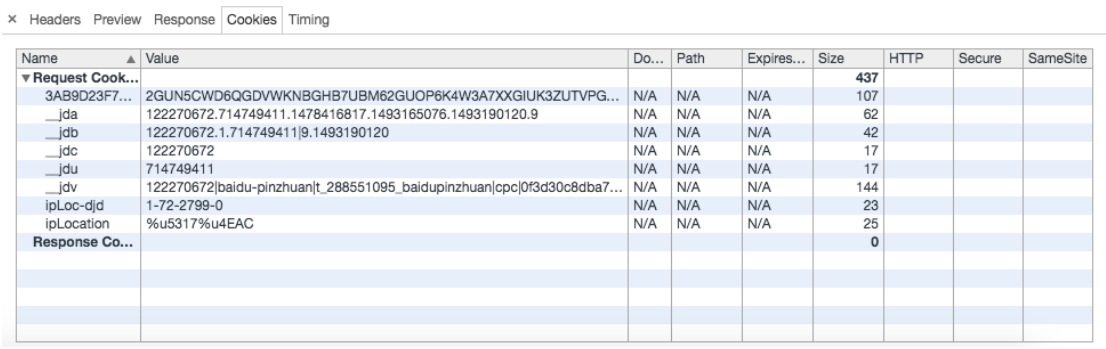


图 7 Cookies 信息

```
#设置cookie参数
cookie = {
    '__jda': '122270672.714749411.1478416817.1493085844.1493103307.5',
    '__jdb': '122270672.1.714749411|5.1493103307',
    '__jdc': '122270672',
    '__jdu': '714749411',
    '__jdv': '122270672|baidu-pinzhuan|t_288551095_baidupinzhuan|cpc|0f3d30c8dba7459bb52f2eb5eba8ac7d_0_b8becb7c06044ad79ffa370e259e8b4a|1492474906949',
    'ipLoc-djd': '1-72-2799-0',
    'ipLocation': '%u5317%u4EAC'
}
```

图 8 Cookies 代码说明

## 2. 找到商品评论的 URL

打开 Headers，找到其中的 Request URL，复制 URL 并把 URL 地址放在浏览器中打开，里面包含了当前页的商品评论信息。这就是我们要抓取的 URL 地址。

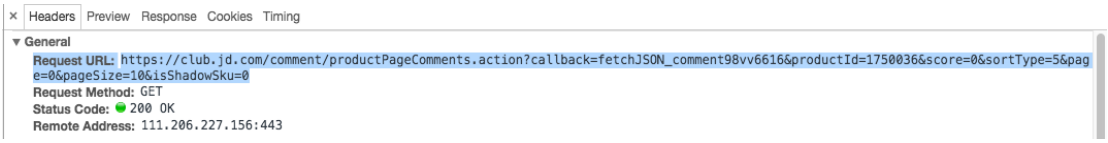


图 9 商品评论 URL

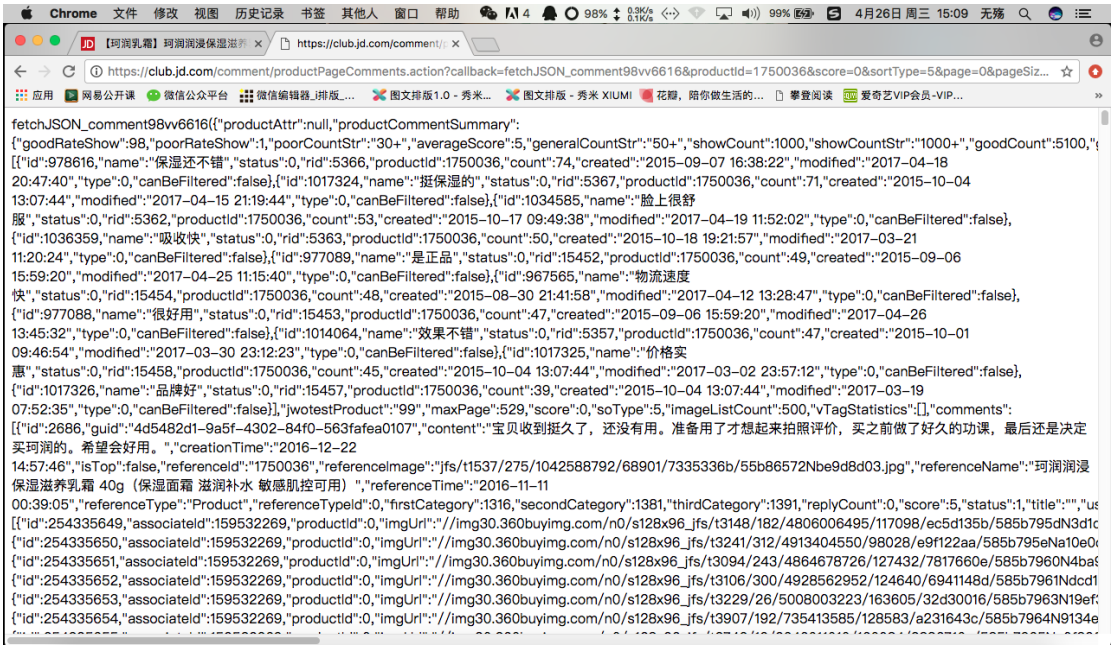


图 10 原始商品评价信息

## 3. 分析 URL

仔细观察这条 URL 地址可以发现，其中 productId=1750036 是当前商品的商品 ID。与商品详情页 URL 中的 ID 一致。而 page=0 是页码。如果我们要获取这个商品的所有评论，只需要更改 page 后面的数字即可。

我们分析的“珂润润浸保湿滋养乳霜 40g”商品的评论有 5300+条，也就是有近 530 页需要抓取，因此页码不是一个固定值，需要在 0-530 之间变化。这里我们将 URL 分成两部分，使用 random 生成 0-530 的唯一随机数，也就是要抓取的页码编号。然后再将页码编号与两部分 URL 进行拼接。

```
#设置URL第一部分
url1 = 'https://club.jd.com/comment/productPageComments.action?callback=fetchJSON_comment98vv6601&productId=1750036&score=0&sortType=5&page='
#设置URL第二部分
url2 = '&pageSize=10&isShadowSku=0'
#生成随机数
ran_num = random.sample(range(530), 530)
```

图 11 URL 代码

## 4. 实现商品评论的翻页循环爬取

下面是具体的抓取过程，使用 for 循环每次从 0-530 的随机数中找一个生成页码编号，与两部分的 URL 进行拼接。生成要抓取的 URL 地址并与前面设置好的头文件信息和 Cookie 信息一起发送请求获取页面信息。将获取到的页面信息进行汇总。每次请求间休息 5 秒针。

抓取完后，我们还需要对页面进行编码。完成编码后就可以看到其中所包含的中文评论信息了。后面的工作就是要对这些评论信息进行不断提取和反复的清洗。

```
#循环抓取页面
for i in ran_num:
    a = ran_num[0]
    if i == a:
        i = str(i)
        url = (url1 + i + url2)
        r= re.get(url=url,headers=headers,cookie=cookie)
        html=r.content
    else:
        i=str(i)
        url = (url1 + i + url2)
        r= re.get(url=url,headers=headers,cookie=cookie)
        html2=r.content
        html=html+html2
        time.sleep(5)

#编码
html=str(html,encoding='utf-8')


#存储
file = open("page5.txt", "w")
file.write(html)
file.close()
```

图 12 循环代码

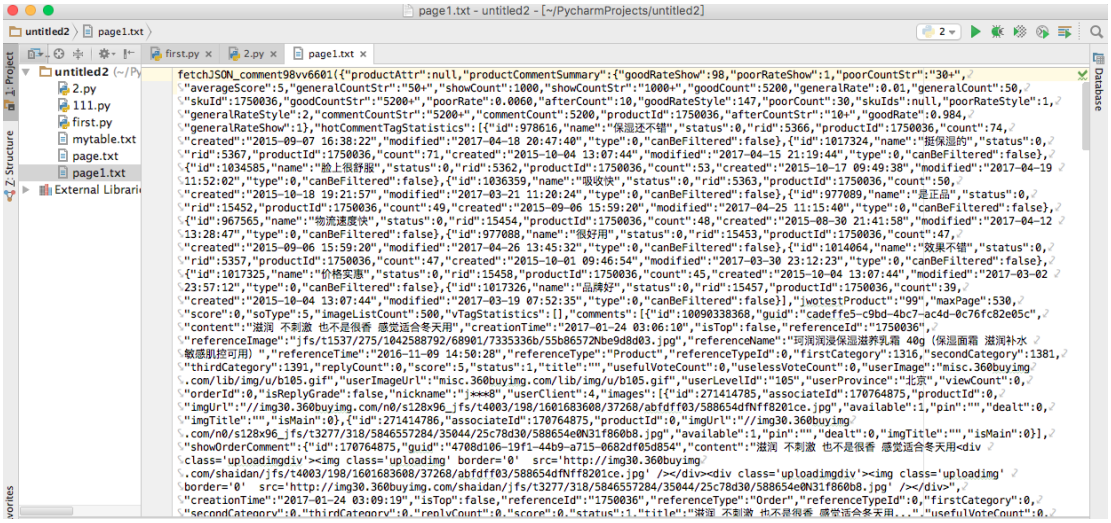


图 13 抓取内容截选

## （二）文本分析

### 1. 评论数据的筛选和去重

采集到的用户评论共有 5300+，在对采集到的数据进行了初步的筛选和去重处理后，得到了有效评论 4335 条。

```
#-*- coding: utf-8 -*-
import pandas as pd

inputfile = '用户评论.csv' #评论文件
outputfile = '用户评论去重.txt' #评论处理后保存路径
data = pd.read_csv(inputfile, encoding = 'utf-8', header = None)
l1 = len(data)
data = pd.DataFrame(data[0].unique())
l2 = len(data)
data.to_csv(outputfile, index = False, header = False, encoding = 'utf-8')
print(u'删除了%s条评论。' %(l1 - l2))
```

图 14 评价数据筛选代码

### 2. ROSTCM6 的情感分析

接着进行情感分析，本论文用 ROSRCM6 来完成该项操作。ROST 系统是由武汉大学开发的，可以实现包括情感分析以及语义网络构建等。ROSTCM6 软件的情感倾向分析适用的是基于优化情感词典的方法。就目前来讲，其准确率会比基于向量神经网络的情感分析方法的准确率高。ROSTCM6 进行情感分析后会生成“正面情感结果”、“中性情感结果”、“负面情感结果”三个文本文档。

### 3. LDA 模型实现

这里我们抛弃“中性情感结果”文本，分别对“正面情感结果”和“负面情

感结果”文本进行 LDA 分析，挖掘商品的优点和不足。

### （1）删除评分前缀

由于 ROSTCM6 得到的结果还有评分前缀，需要对前缀进行评分删除。

```
#-*- coding: utf-8 -*-
import pandas as pd

#参数初始化
inputfile1 = '用户评论去重负面情感结果.txt'
inputfile2 = '用户评论去重正面情感结果.txt'
outputfile1 = 'EA_neg.txt'
outputfile2 = 'EA_pos.txt'

data1 = pd.read_csv(inputfile1, encoding = 'utf-8', header = None) #读入数据
data2 = pd.read_csv(inputfile2, encoding = 'utf-8', header = None)

data1 = pd.DataFrame(data1[0].str.replace('.*?\d+?\\t ', '')) #用正则表达式修改数据
data2 = pd.DataFrame(data2[0].str.replace('.*?\d+?\\t ', ''))

data1.to_csv(outputfile1, index = False, header = False, encoding = 'utf-8') #保存结果
data2.to_csv(outputfile2, index = False, header = False, encoding = 'utf-8')
```

图 15 删除评分前缀代码

### （2）分词处理

接下来需要对两文本进行分词处理，保存为两个文本文档，并和停用词文档一起作为 LDA 程序的输入。

```
#-*- coding: utf-8 -*-
import pandas as pd
import jieba #导入结巴分词，需要自行下载安装

#参数初始化
inputfile1 = 'EA_neg.txt'
inputfile2 = 'EA_pos.txt'
outputfile1 = 'EA_neg_cut.txt'
outputfile2 = 'EA_pos_cut.txt'

data1 = pd.read_csv(inputfile1, encoding = 'utf-8', header = None) #读入数据
data2 = pd.read_csv(inputfile2, encoding = 'utf-8', header = None)

mycut = lambda s: ' '.join(jieba.cut(s)) #自定义简单分词函数
data1 = data1[0].apply(mycut) #通过“广播”形式分词，加快速度。
data2 = data2[0].apply(mycut)

data1.to_csv(outputfile1, index = False, header = False, encoding = 'utf-8') #保存结果
data2.to_csv(outputfile2, index = False, header = False, encoding = 'utf-8')
```

图 16 分词处理代码

### （3）LDA 分析

在分好词的正面评价、负面评价以及过滤用的停用词表的基础上，使用 Python 的 Gensim 库完成 LDA 分析。

```
#-*- coding: utf-8 -*-
import pandas as pd

#参数初始化
negfile = 'EA_neg_cut.txt'
posfile = 'EA_pos_cut.txt'
stoplist = 'stoplist.txt'

neg = pd.read_csv(negfile, encoding = 'utf-8', header = None) #读入数据
pos = pd.read_csv(posfile, encoding = 'utf-8', header = None)
stop = pd.read_csv(stoplist, encoding = 'utf-8', header = None, sep = 'tipdm')
#sep设置分割词，由于csv默认以半角逗号为分割词，而该词恰好在停用词表中，因此会导致读取出错
#所以解决办法是手动设置一个不存在的分割词，如tipdm。
stop = [' ', ''] + list(stop[0]) #Pandas自动过滤了空格符，这里手动添加

neg[1] = neg[0].apply(lambda s: s.split(' ')) #定义一个分割函数，然后用apply广播
neg[2] = neg[1].apply(lambda x: [i for i in x if i not in stop]) #逐词判断是否停用词，思路同上
pos[1] = pos[0].apply(lambda s: s.split(' '))
pos[2] = pos[1].apply(lambda x: [i for i in x if i not in stop])

from gensim import corpora, models

#负面主题分析
neg_dict = corpora.Dictionary(neg[2]) #建立词典
neg_corpus = [neg_dict.doc2bow(i) for i in neg[2]] #建立语料库
neg_lda = models.LdaModel(neg_corpus, num_topics = 3, id2word = neg_dict) #LDA模型训练
for i in range(3):
  neg_lda.print_topic(i) #输出每个主题

#正面主题分析
pos_dict = corpora.Dictionary(pos[2])
pos_corpus = [pos_dict.doc2bow(i) for i in pos[2]]
pos_lda = models.LdaModel(pos_corpus, num_topics = 3, id2word = pos_dict)
for i in range(3):
  neg_lda.print_topic(i) #输出每个主题
```

图 17 LDA 分析代码

## （三）结论分析

经过 LDA 主题分析后，根据生成的词频表，并通过人工筛选排除实验中存在的误差，评论文本的分析情况（表中的数字为各个分词的词频）如表 2，3 所示。

表 2 正面评价潜在主题

| 主题一 | 主题二 | 主题三 |
|---|---|---|
| 不错（76） | 京东（35） | 适合敏感（34） |
| 保湿（49） | 正品（21） | 味道（31） |
| 感觉（46） | 划算（21） | 冬天（21） |
| 效果（42） | 价格（19） | 适合（19） |
| 知道（35） | 喜欢（18） | 舒服（17） |
| 面霜（31） | 满减（15） | 不刺激（14） |
| 滋润（27） | 满意（13） | 不油（13） |
| 皮肤（26） | 快递（10） | 香味（9） |
| 吸收（24） | 便宜（9） | 淡淡的（9） |
| 好用（23） | 推荐（9） | 护肤（7） |

表 3 负面评价潜在主题

| 主题一 | 主题二 | 主题三 |
| --- | --- | --- |
| 油腻（19） | 包装（21） | 囤货（15） |
| 厚重（18） | 不好（19） | 时间（12） |
| 刺痛（16） | 是不是（17） | 物流（12） |
| 痘痘（15） | 假货（15） | 送货（11） |
| 受不了（9） | 真假（13） | 自营（10） |
| 闭口（9） | 塑料（13） | 地方（9） |
| 粉刺（8） | 失望（12） | 断货（9） |
| 难受（7） | 很小（10） | 略难（8） |
| 浮油（5） | 不知（8） | 存货（6） |
| 黏腻（3） | 试用（5） | 缺货（6） |

根据对珂润面霜好评的三个潜在主题的特征词提取，主题一的高频特征词，即不错、保湿、感觉、效果、知道、面霜、滋润、皮肤、吸收、好用。主要反映此款面霜保湿效果好，温和滋润，补水好用的特点；主题二中的高频特征词，即京东、正品、划算、价格、喜欢、满减、满意、快递、便宜、推荐。主要反映京东自营店的价格划算，满减活动，正品等特点；主题三中的高频特征词，即适合敏感、味道、冬天、适合、舒服、不刺激、不油、香味、淡淡的、护肤。主要关注于珂润面霜的功效，适合冬季干燥，缓解敏感，无香精添加等特点。

根据对珂润面霜负评的三个潜在主题的特征词提取，主题一的高频特征词，即油腻、厚重、刺痛、痘痘、受不了、闭口、粉刺、难受、浮油、黏腻。主要反映对于部分消费者来说，面霜质感油腻，不易推开，造成痘痘闭口等情况；主题二的高频特征词，即包装、不好、是不是、假货、真假、塑料、失望、很小、不知、试用、主要反映面霜的塑料包装，分量小，造成消费者不易辨别商品真伪等特点；主题三的高频特征词是囤货、时间、物流、送货、自营、地方、断货、略难、存货、缺货。主要反映京东的物流配送较慢，地方缺货现象等。

综合以上对主题及高频特征词可以看出，珂润面霜的优势可以有以下几个方面：珂润面霜的保湿效果好，温和补水、京东平台的满减活动价格实惠、有正品保证、面霜能缓解季节性干燥、缓解敏感、使用神经酰胺，无添加香精。

相对而言，用户对珂润面霜的负面评价可以总结为以下几个方面：面霜的质感较厚重、上脸不易涂抹、实物的塑料包装、分量较小、消费者不易商品辨别真伪、京东自营的部分地区总是出现断货现象。

根据对京东上珂润面霜的商品评价进行 LDA 主题分析，对珂润面霜这款产品提出以下建议。

1.在保持面霜滋润的同时，改善面霜的厚重质感，从而达到既补水又轻薄的效果感受。

2.增加商品的防伪标识，打消消费者由于商品包装而不易辨别商品真伪的不安全感。

3.物流配送方面，京东自营店应该根据每个地区的销量合理分配各个地区的货物存储，防止出现部分地区一款产品的长时间断货，提高货物配送与服务水平。

# 五、总结和展望

## （一）全文总结

电商网站的发展使得人们越来越依赖于网络购物，而伴随着网购数量的增加，产生了大量的网络交易信息。一方面，用户希望从这些交易信息中得到可靠的商品信息;另一方面，商家希望从这些交易信息中挖掘出有价值的用户反馈。因此，对于电商平台中用户评价的文本挖掘显得很有必要。

本文主要使用的开发语言是 Python，开发环境是 Pycharm 3.0，相比于其他语言，Python 的便捷之处在于其内置有“jieba”中文分词包和 Gensim 库可以用来进行文本的分词处理和 LDA 主题分析。本文的主要任务是设计一款爬取电商平台中“商品评价”信息的爬虫流程，然后对爬取的信息进行文本去重、情感分析、中文分词和 LDA 主题分析，最终得到对商家和用户有用的信息。

为了证明研究的可行性，论文中以京东平台上的一款面霜为演示对象，进行了从数据爬取到文本挖掘的全部流程，并得到相应结论。

## （二）不足之处

本文在设计和实际操作的过程中存在着以下不足：爬虫设计可以优化、文本预处理不够全面、情感分析结果存在误差、产品分析缺乏比对。

爬虫设计方面，研究中设计的爬虫代码不够健壮，缺少错误处理机制，所以若是代码运行过程中出现错误，程序会立即终止，降低爬取的效率。

文本处理方面，一般文本处理包括文本去重、机械语料压缩以及短句删除，而本实验中只对商品评论数据进行了文本去重，没有进行机械语料压缩和短句删除，可能会对之后的结果造成误差。

情感分析方面，研究中采用的是 ROSTCM6 软件进行的情感分析，没有实现全过程的 Python 开发。此外，ROSTCM6 分析得到的结果存在一些误差，有些正面的情感结果会出现在负面情感结果的文件中，从而导致了后续的研究结果存在相应的误差。

研究对象方面，研究中只对京东上的一款商品进行了实证研究，缺乏其他商品的比对，所以说服力会降低。此外，单一实验的成功并不能很好地说明研究中设计的流程对每一个产品都适用，实验次数的增加可能会增加流程设计的可靠性，而且可以在多次的重复试验中发现设计流程存在的不足，对其进行完善。

## （三）研究展望

针对以上不足之处，之后的研究将着手解决：1.系统学习 Python 语言开发爬虫，完善代码的性能和健壮性。2.详细了解文本数据的机械语料加工的方法，尝试对抓取的评论数据进行语料压缩，增加后续实验数据的可靠性。3.学习 Python 的情感分析方法，尝试用 Python 代码对评论数据进行情感分析，降低因

使用 ROSTCM6 软件而产生的实验误差。4.增加不同品牌不同产品以及不同类型的实验数据，不断对数据模型进行补充和完善。

电商平台中商品评价的数据挖掘不仅能为用户提供真实可靠的信息，还能为商家提供产品决策、情报分析以及改善措施。文本挖掘是一个方兴未艾的领域，国外都尚未成熟，国内也处于起步阶段，这里存在着许多有价值的问题值得我们研究。我相信，未来随着自然语言处理、文本信息处理、机器学习等技术的发展，在线评论数据挖掘将会为人们提供更多有价值的信息。

# 参考文献

[1]袁志丽.商务部：电子商务交易总额增速是 GDP 的 3.86 倍[EB/OL]. http://intl.ce.cn/specials/zxxx/201505/15/t20150515_5376170.shtml ，2015-05-15.
[2]前瞻产业研究院.近几年中国电子商务规模数据浅析[EB/OL]. http://bg.qianzhan.com/report/detail/459/150708-a39edc6f.html，2015-07-08.
[3]郝建彬. 1995-2015 中国电子商务 20 年发展史话[EB/OL]. http://www.199it.com/archives/369754.html，2015-07-26.
[4]聂林海.我国电子商务发展的特点和趋势[J].中国流通经济，2014，6：97-101.
[5]S M Mudambi,D Schuff. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com[J].MIS Quarterly,2010,34(1):185-200.
[6]苗蕊.在线评论有用性研究综述[J].中国管理信息化，2014，17：126-128.
7]房文敏，张宁，韩雁雁. 在线评论信息挖掘研究综述[J].信息资源管理学报，2016，1：4-11.
[8]汪勇慧.在线评论研究的知识图谱分析[J].情报探索，2015，11：127-132.
[9]岳中刚，王晓亚.在线评论与消费者行为的研究进展与趋势展望[J].软科学，2015，6（29）：90-93.
[10]胡冰, 胡东军，马文超.文本挖掘研究及发展[J].电脑知识与技术，2008，4（4）：792-793.
[11]谌志群.文本趋势挖掘综述[J].情报科学，2010，2（28）：316-320.
[12]袁军鹏，朱东华，李毅，李连宏，黄进.文本挖掘技术研究进展[J].计算机应用研究，2006，2：1-4.
[13]梅馨, 邢桂芬.文本挖掘技术综述[J].江苏大学学报自然科学版，2003，5（24）：72-76.
[14] Lighthouse_王豆豆. 基于电商平台探讨：评价机制的价值与模式创新[EB/OL]. http://www.woshipm.com/evaluating/479089.html，2016-12-03.
[15]孙霄凌.购物网站在线评论系统功能的演变及适应性使用研[D].南京：南京大学，2014.
[16]孙立伟，何国辉，吴礼发.网络爬虫技术的研究[J].电脑知识与技术，2010，6（15）：4112-4115.
[17]于娟，刘强.主题网络爬虫研究综述[J].计算机工程与科学，2015，2（37）：231-237.
[18]周立柱，林玲.聚焦爬虫技术研究综述[J].计算机应用，2005，25（9）：1965-1969.
[19]方星星，鲁磊纪，徐洋.网络舆情监控系统中主题网络爬虫的研究与实现[J].舰船电子工程，2014，9（34）：104-107.
[20]张明杰.基于网络爬虫技术的舆情数据采集系统设计与实现[J].现代计算机，2015，18（4）：72-75.
[21]彭纪奔，吴林，陈贤，黄雷君.基于爬虫技术的网络负面情绪挖掘系统设计与

实现[J].计算机应用与软件，2016，33（10）：9-13.
[22]周中华，张惠然，谢江 .基于 Python 的新浪微博数据爬虫[J].计算机应用，2014，34（11）：3131-3134.
[23]董浩然，谢欢，陈鹏，洪中华，童小华.基于 GIS 主题爬虫的在线房产估价系统与优化[J].地理信息世界，2016，23（2）：107-112.
[24]卢长宝，庄晓燕..餐饮业服务质量在线评论的聚焦维度:基于网络爬虫技术的实证研究[J].天津商业大学学报，2016，36(4)：14-22.
[25]邓宏勇,许吉,张洋,袁敏,施毅. 中医药数据挖掘研究现状分析[J].中国中医药信息杂志，2012，10（19）：21-23.
[26]许高建.基于 Web 的文本挖掘技术研究[J].计算机技术与发展，2007，17（6）：187-190.
[27]胡阿沛，张静，雷孝平，张晓宇.基于文本挖掘的专利技术主题分析研究综述[J].情报杂志，2013，32（12）：88-92.
[28]郭金龙，许鑫.数字人文中的文本挖掘研究[J].大学图书馆学报，2012，3：11-18.
[29]郭金龙，许鑫，陆宇杰.人文社会科学研究中文本挖掘技术应用进展[J].图书情报工作，2012，56（8）：10-17.
[30]王浩畅，赵铁军.生物医学文本挖掘技术的研究与进展[J].中文信息学报，2008，22（3）：89-98.
[31]黄晓斌，赵超 .文本挖掘在网络舆情信息分析中的应用[J].情报科学，2009，1（27）：94-99.
[32]孟雪井，孟祥兰，胡杨洋.基于文本挖掘和百度指数的投资者情绪指数研究[J].宏观经济研究，2016，1：144-153.
[33]吴恒，陈燕翎. 基于 UGC 文本挖掘的游客目的地选择信息研究-以携程蜜月游记为例[J].情报科学，2017，35（1）：101-105.
[34]张玉峰，朱莹.基于 Web 文本挖掘的企业竞争情报获取方法研究[J].情报理论与实践，2006，29（5）：563-566.
[35]熊伟，郭扬杰.酒店顾客在线评论的文本挖掘[J].北京第二外国语学院学报，2013，11：38-47.
[36]肖旻，陈行.基于 Python 语言编程特点及应用之探讨[J].电脑知识与技术，2014，10（34）：8177-8178.
[37]夏火松，李保国.基于 Python 的动态网页评价爬虫算法[J].软件工程，2016，19（2）：43-46.
[38]Richard Lawson.用 Python 写网络爬虫[M].北京：人民邮电出版社，2016：3-38.
[39]金涛.网络爬虫在网页信息提取中的应用研究[J].现代计算机，2012，1：16-18.
[40]杨定中，赵刚，王泰.网络爬虫在 Web 信息搜索与数据挖掘中应用[J].计算机工程与设计，2009，30（24）：5658-5662.
[41]方美玉，郑小林，陈德人，华艺，施艳.商品评论聚焦爬虫算法设计与实现[J].吉林大学学报，2012，42（1）：377-381.
[42]郭涛，黄铭钧.社区网络爬虫的设计与实现[J].智能计算机与应用，2012，2（4）：65-67.

[43]尹江,尹治本,黄洪.网络爬虫效率瓶颈的分析与解决方案[J].计算机应用，2008，28（5）：1114-1117.
[44]张良均.Python 数据分析与挖掘实战[M].北京：机械工业出版社，2016
[45]曾伟辉，李淼. 基于 Javascript 切片的 AJAX 框架网络爬虫技术研究[J].计算机系统应用，2009，7：169-172.
[46]黄仁，王良伟.基于主题相关概念和网页分块的主题爬虫研究[J].计算机应用研究，2013，30（8）：2377-2381.

# 致谢

大学生涯即将结束，在湖南大学的四年学习生活中，我收获颇多，心智更加成熟，专业知识也更加完备，这些成长离不开老师的关心、同学的帮助与亲人的支持。

首先我要感谢我的论文指导老师江资斌老师，每当我对论文有疑问和困惑的时候，江老师总是十分耐心的为我答疑解惑，不仅为我提供写论文时需要的资料书籍，还在我完成论文初稿的过程中仔细阅读我的论文并提出修改意见。正是如此，我的论文结构才得到逐步的完善;因为江老师悉心的指导，我才完成了这篇毕业论文，在此我要对江老师致以最诚挚的感谢。

其次，我要感谢我的同班同学们，正是大家互相的鼓励与帮助，才让我的大学变得充实和快乐，这四年的生活将是我最美好的回忆。

最后，我要感谢我的父母，本科四年虽然多半的时间都没能和你们在一起，但是你们对我的支持与关心却无处不在，这也是我前进的最大动力。

# 附录 A

**部分评论节选**

| 序号 | 评价 |
|---|---|
| 1 | 宝贝收到挺久了，还没有用。准备用了才想起来拍照评价，买之前做了好久的功课，最后还是决定买珂润的。希望会好用。 |
| 2 | 双十一买的，价钱比平时便宜点。因为前面的没用完，所以这个一直都没用，现在准备用了。，买之前做了好久的功课。希望敏感皮肤能有所改善。 |
| 3 | 小小一罐，冰淇淋般的质地，我是干性皮肤，涂上不油，但是我觉得稍微有点黏腻，保湿不错 |
| 4 | 买这类护肤品，还是只敢京东买。速度又快。比专柜实惠多了。信赖京东 |
| 5 | 物流速度如往常一样给力。买了一罐面霜回来与之前在国内专柜买的对比，外包装盒和瓶子都一般无二，味道质地也一样，除了保质期和批次数字的字体稍稍有区别，其他都一致。 |
| 6 | 买了一天就送货到位了。速度直接肯定。包装盒有一个压痕，但是里面有空气塑料袋，东西没有损伤。 |
| 7 | 不错，买了第四瓶了！这时候感觉面霜有点不够滋润了，考虑加点油油！坐标山东临沂，大北方！ |
| 8 | 神速，上午下单，下午到。老婆用了，感觉不错。活动也比较给力。希望以后有套装以及更多活动就好了 |
| 9 | 喜欢，以后还会来的，适合我的肤质，而且包装跟专卖店一样 |
| 10 | 有点贵啊，打开有点像牛奶冰激凌 |
| 11 | 去年在京东买的，当时用着不觉得好，就搁置了，今年过了年又拿来用，效果竟然超好于是打算再买，可可惜京东一直无货！昨天看到又上架，果断下单!!! |
| 12 | 用着还可以吧好小的一瓶晚上用挺好的早上用有点油看个人肤质吧 |
| 13 | 用了半盒来评价～之前一直用着科颜氏的高保湿。冬天的时候皮肤有些敏感，会痒和刺痛，后来干到起皮。上一盒用完了就想试试珂润。上脸感觉是没有科颜氏的好推，手感有一点点油厚。但是！早上涂了以后，一整天都没有起皮！让我惊讶的是到了晚上回家摸着还是润润滴～哈哈哈意外收获～虽然单克价格比科颜氏高一点点，但决定再入手一盒～ |
| 14 | 快递师傅人特别好特别负责辛苦啦很保湿特别好用 |
| 15 | *，之前用神仙水剥离的皮肤太薄了，以至于秋冬季会泛红，但素，涂了珂润面霜之后再就没有泛红过，不油，甚至有一丢丢哑光效果。慕斯质感，油皮冬天用再适合不过了。强力推荐，花王授权，上午下单下午送达。*简直精吸！喜翻儿喜翻儿惹！请国产野鸡通通去世厚！ |
| 16 | 一直都用这款面霜，用起来挺好 |
| 17 | 小小的一瓶，涂起来挺舒服的，没有什么味道也不是很油，看了一下生产日期也是最近的。用一段时间再过来追评，京东快递一如既往的快包装也真是一如既往的简陋啊? |
| 18 | 特意用过再来评价的~首先，正品无疑。珂润浸润保湿面霜都被推烂了，实在 |

| | |
|---|---|
| | 是很火。我也是在皮肤不稳定的时候入的。它的使用感很好，并且没有味道。上脸很容易推开，本人混油皮，完全可以接受，一点都不黏腻。春夏也能使用。我早晚都用，在不熬夜不乱用其它护肤品的情况下，基本没怎么再爆痘。所以它的维稳效果是特别棒的。所以推荐有需要的小伙伴入手~ |
| 19 | 京东快递很快，包装专业结实。快递小哥服务代码态度非常好。很有礼貌。这款宝贝现在搞活动买非常划算，也是正品，大品牌值得信任。推荐大家购买 |
| 20 | 我，我买的是滋润型的水，本来自己很偏，干性的皮肤用起来还是觉得很油的，换清爽型的会好些里面确实不含酒精，和防腐剂的，涂到眼睛那里也不会觉得有辣辣的感觉，很少很少的量，就能够使脸比较舒服了。 |
| 21 | 开心的收到了，只是看着好小的一罐啊，用了几次了，吸收还不错，清爽滋润，非常好，淡淡的清香！快递不错，双十一也还可以了，包装也还可以，京东的活动比较实惠吧，就是有时候会没货！ |
| 22 | 帮朋友买的，珂润的面霜用过很不错，试试洗发水和身体乳霜。京东物流很快，是正品满意。 |
| 23 | 这个真的慢滋润，冬天用比多元面霜感觉还好用些。也没有什么刺激，挺喜欢的。护肤品只是一部分，要想皮肤好平时的生活习惯也很重要。希望能变更美更好的自己。 |
| 24 | 双 11 买的，比较实惠。日期也比较新鲜。相信京东的品质。 |
| 25 | 趁有活动买来囤货的，长草这个面霜很久了！感谢快递小哥送货上门，今天上午下的单，这会儿就收到货了，京东的速度没话说 |
| 26 | 好小一瓶，没有味道，擦起来挺舒服。本人两颊起皮，痒。敏感干皮。擦了两天不痒了，挺好的。超出预期 |
| 27 | 真爱面霜珂润舒敏的原理是添加神经酰胺功能成分，相当于皮肤角质层里面的水泥，敏感皮缺乏神经酰胺导致皮肤耐受力脆弱。这个面霜含神经酰胺成分最多，质地也是厚重乳霜质地（无矿物油无番石榴）但是好吸收。干敏皮必入的。 |
| 28 | 这款经常卖断货啊用了一段时间了。抹在脸上舒服，会持续回购的 |
| 29 | 关注这个品牌很久了，听说是干皮敏感肌的救星，买来试用 |
| 30 | 158 两瓶立减 100，超级划算。天冷一直用这款，保湿防敏，好用。和专柜买的一样。 |
| 31 | 双十一之前满减加用券以酒吧一瓶买到，很便宜了。用起来很舒服，薄薄一层，抹了，也不油，前两天不小心掉了，洒了半瓶多，心痛死了。 |
| 32 | 一直在用，感觉很保湿，很滋润。无限回购中。 |
| 33 | 包装挺好，美妆分不出真假，看起来还不错。 |
| 34 | 一直都用的这款面霜，活动价很超值，一下囤了 4 瓶 |
| 35 | 东西不错，蛮滋润的，就是不便宜。搞活动也小贵。 |
| 36 | 面霜保湿补水能力强，总体值得推荐 |
| 37 | 香味正好淡淡的好闻期待效果 |
| 38 | 皮肤特别需要保湿，就买来用用啦 |
| 39 | 和专柜买的一样效果，保湿程度一般般 |
| 40 | 清爽不油腻，吸收挺快，没什么味道，朋友推荐给我的。 |

| | |
|---|---|
| 41 | 霜很好用一点都没有不舒服的感觉吸收挺好的 |
| 42 | 很小，但是满满的霜，还没用，应该可以吧？ |
| 43 | 趁双十二优惠帮我女儿买的，应该不错吧。 |
| 44 | 这个霜其他都没什么问题，就是这个霜用了 2 天，把表面用了，今天早上起来擦脸的时候就发现有个洞，我也不知道怎么回事？亲们，请问你们有这种情况吗？这种情况是否正常？ |
| 45 | 刚才，刚才，是吧那个是乳液和，水的顺序弄错了，乳液的话，我再说脸上非常的滋润，很好吸收，里面确实是无添加和没有防腐剂的，也没有酒精，夏天对于我偏干的皮肤来说非常适用，每次只要，即 1 到 2，就已经足足的够了 |
| 46 | 包装很不错，送货很快，冬天必须使用这个，效果很好很滋润，买了好多回了 |
| 47 | 用了几天来评价，保湿效果很好，也没有过敏，快递很快，本来想买韩妆的，找的过程中看见这个品牌，后来看着不错，就从京东超市里找到了现货的，抱着试试的态度买的，用了以后还是蛮惊喜，很好用最主要保湿效果真的很好，推荐一下。 |
| 48 | 敏感性皮肤，一到换季时皮肤就会过敏红肿。无意中在网上看到了关于珂润的评价，买来使用后果真没有任何不适感，不油腻，保湿效果很好。 |
| 49 | 真不错，同事推荐的。 |
| 50 | 在网上看了如何辨别真假，说是面霜如果颠簸不会有气泡，打开果然是这样，希望是正品，用着没有什么不良，滋润效果可以。二次购买，自营送货速度快。 |
| 51 | 之前在日本买过，但是送老妈了，这次再买个自己用，效果还未知 |
| 52 | 包装完好，快递给力，拧开瓶盖凑近闻才有一股淡淡的清凉药香味，试着擦了一点真的是清爽不油腻，一点也不粘，本身皮肤偏油，所以保不保湿感觉不出来，不过口碑这么好，保湿效果应该是不错的。 |
| 53 | 之前买过一瓶已经快用完了，孕期用着安心，天热了要减少用量，不然会有点粘。保质期很长，给你们看一下成分。 |
| 54 | 一直在京东买珂润家的东西，比超市里便宜一些，送货快，没有味道不刺激，喜欢。 |
| 55 | 我是敏感体质，一直用珂润洗发水。过去在天津恒隆买，现发现京东有后，而且活动更优惠！实在太意外惊喜了！ |
| 56 | 这段时间皮肤反复过敏，换了好几种产品，最近皮肤又过敏，最终选择珂润救急买的，晚上使用后确实还不错，早上起床发现过敏起的小红疹都消了，会坚持使用，希望皮肤变好 |
| 57 | 第二瓶了，真的很好用，滋润又不腻，对于我这种干性敏感肌肤简直是太适合了，就是小小一罐用的太快～ |
| 58 | 空调房里用了一下，保湿效果很好，完胜雅诗兰黛红石榴。以后就用这个做空调房保湿了，性价比高！ |
| 59 | 速度杠杠，上午下单，下午就到，这就是京东速度。还没使用，不知道效果是否和以前一样。 |
| 60 | 珂润用过好几个了，使用感很好，效果也还不错，价格活动时候还可以～信任京东，所以买了好几个^0^ |