文本复制检测报告单(全文标明引文)

№:ADBD2019R_20190404173348440572976103　　检测时间:2019-04-04 17:33:48

检测文献:　基于文本挖掘的电影短评分析

作者:　苏俊恒

检测范围:　中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
大学生论文联合比对库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
CNKI大成编客-原创作品库
个人比对库

时间范围:　1900-01-01至2019-04-04

检测结果

去除本人已发表文献复制比：13.2%　　跨语言检测结果：0%

去除引用文献复制比：13.2%　　总文字复制比：13.2%

单篇最大文字复制比：4.3%（141220089_沈煜_唐斌_基于优酷的视频访问数据分析）

重复字数：[1899]	总字数：[14351]	单篇最大重复字数：[612]
总段落数：[2]	前部重合字数：[367]	疑似段落最大重合字数：[1432]
疑似段落数：[2]	后部重合字数：[1532]	疑似段落最小重合字数：[467]

指　标：☐疑似剽窃观点　☑疑似剽窃文字表述　☐疑似自我剽窃　☐疑似整体剽窃　☐过度引用

表　格：0　　公　式：2　　疑似文字的图片：0　　脚注与尾注：0

14.5%（1432）　基于文本挖掘的电影短评分析.docx-z_第1部分（总9903字）
10.5%（467）　基于文本挖掘的电影短评分析.docx-z_第2部分（总4448字）

（注释：无问题部分　文字复制部分　引用部分）

指导教师审查结果

指导教师:　崔德鑫

审阅结果:

审阅意见:　指导老师未填写审阅意见

1. 基于文本挖掘的电影短评分析.docx-z_第1部分

总字数：9903

相似文献列表

去除本人已发表文献复制比：14.5%(1432)　文字复制比：14.5%(1432)　疑似剽窃观点：(0)

1	141220089_沈煜_唐斌_基于优酷的视频访问数据分析 沈煜 -《大学生论文联合比对库》- 2018-05-28	6.2%（612） 是否引证：否
2	韦燕丹--201418121229-基于网络爬虫的医疗信息采集技术研究 基于网络爬虫的医疗信息采集技术研究 -《大学生论文联合比对库》- 2018-05-21	6.1%（608） 是否引证：否
3	Scrapy学习 - 博客频道 - CSDN.NET -《网络（http://blog.csdn.net）》- 2017	6.1%（604） 是否引证：否
4	面向安卓系统的Python网络爬虫管理平台设计与实现 蒋文皓 -《大学生论文联合比对库》- 2017-05-30	6.1%（603） 是否引证：否

5	201203111面向网络新闻的MongoDB技术分析与应用 郭一鸣 - 《大学生论文联合比对库》- 2016-06-01	6.1%（603） 是否引证：否
6	面向新闻网站信息的爬虫系统设计与实现 张光宇 - 《大学生论文联合比对库》- 2017-05-24	6.0%（592） 是否引证：否
7	Python爬虫系列之----Scrapy(一)爬虫原理 - fendo - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	6.0%（592） 是否引证：否
8	201203111面向网络新闻的MongoDB技术分析与应用 郭一鸣 - 《大学生论文联合比对库》- 2016-06-08	6.0%（591） 是否引证：否
9	基于网络爬虫的搜索引擎设计与实现 刘超波 - 《大学生论文联合比对库》- 2018-05-05	6.0%（591） 是否引证：否
10	11672944_佘晨喜_互联网涉税信息数据挖掘的研究 佘晨喜 - 《大学生论文联合比对库》- 2017-06-12	6.0%（591） 是否引证：否
11	05+信工+董京蕾 董京蕾 - 《大学生论文联合比对库》- 2018-05-23	5.9%（587） 是否引证：否
12	基于 Scrapy 和 Selenium 的爬虫系统构建 何显杰 - 《大学生论文联合比对库》- 2016-05-18	5.7%（569） 是否引证：否
13	python网络爬虫 刘祥宇 - 《大学生论文联合比对库》- 2017-05-18	5.7%（568） 是否引证：否
14	基于Hack Growth思想的移动应用用户行为/数据分析平台 熊文军 - 《大学生论文联合比对库》- 2016-06-02	5.7%（566） 是否引证：否
15	网页篡改分析系统设计与开发 汪开先 - 《大学生论文联合比对库》- 2015-03-19	5.5%（549） 是否引证：否
16	网页篡改分析系统设计与开发 汪开先 - 《大学生论文联合比对库》- 2015-03-29	5.5%（549） 是否引证：否
17	绿苑酒店客房管理系统的运行与实现 何健 - 《大学生论文联合比对库》- 2015-04-02	5.5%（549） 是否引证：否
18	WK180808209_帅轲_基于网络爬虫算法的用户行为数据获取方法 帅轲 - 《大学生论文联合比对库》- 2018-05-21	5.5%（549） 是否引证：否
19	Scrapy研究探索（三）——Scrapy核心架构与代码运行分析 - lyy14011305的博客 - CSDN博客 -《网络（http://blog.csdn.net）》- 2017	5.3%（529） 是否引证：否
20	123012014104_陈龙伟_林劼_基于Scrapy分布式爬虫的关于知乎网数据的抓取分析 陈龙伟 - 《大学生论文联合比对库》- 2018-04-03	5.3%（521） 是否引证：否
21	3100000190_刘则作_毕业设计最终版 刘则作 - 《大学生论文联合比对库》- 2017-04-05	5.0%（494） 是否引证：否
22	3100000190_刘则作_毕业设计最终版 刘则作 - 《大学生论文联合比对库》- 2017-04-10	5.0%（494） 是否引证：否
23	3100000190_刘则作_毕业设计最终版 刘则作 - 《大学生论文联合比对库》- 2017-04-26	5.0%（494） 是否引证：否
24	微博舆情事件中的意见领袖挖掘 田鲲 - 《大学生论文联合比对库》- 2018-05-28	4.9%（484） 是否引证：否
25	互联网新闻分布式采集管理系统设计与实现 王希超 - 《大学生论文联合比对库》- 2018-06-06	4.9%（484） 是否引证：否
26	文本挖掘 挖掘知识 常青 - 《中国计算机用户》- 2004-06-14	2.0%（201） 是否引证：否
27	155242013045_洪鸿飞_分布式Python爬虫设计与实现 洪鸿飞 - 《大学生论文联合比对库》- 2017-04-05	1.8%（176） 是否引证：否
28	文本挖掘综述 杨霞 - 《大学生论文联合比对库》- 2016-07-10	1.1%（111） 是否引证：否
29	基于电商评论的手机类商品满意度调查 陈丁越 - 《大学生论文联合比对库》- 2018-04-10	1.0%（96） 是否引证：否
30	144081100224_于博_基于统计特征的新闻关键词抽取方法 于博 - 《大学生论文联合比对库》- 2018-05-29	0.7%（73） 是否引证：否
31	政府开放数据相同属性识别和检索方法研究 黄跃萍(导师：赵龙文) - 《华南理工大学博士论文》- 2016-04-13	0.4%（38） 是否引证：否
32	网络电影市场研究——基于网络影评数据	0.4%（37）

原文内容

北京师范大学珠海分校

本科生毕业论文

论文题目：基于文本挖掘的电影短评分析

学院管理学院
专业信息管理与信息系统
学号 1502020056
学生姓名苏俊恒
指导教师姓名崔德鑫
指导教师单位北师大珠海分校管理学院

2019 年 3 月 14 日

北京师范大学珠海分校学位论文写作声明和使用授权说明

学位论文写作声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。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。本声明的法律结果由本人承担。

论文作者签名： 日期： 年月日

学位论文使用授权说明

本人完全了解北京师范大学珠海分校关于收集、保存、使用学位论文的规定，即：按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务；学校可以采用影印、缩印、数字化或其它复制手段保存论文；在不以赢利为目的的前提下，学校可以将学位论文编入有关数据库,提供网上服务。（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年月日

基于文本挖掘的电影短评分析

摘要

由于电影是根据观影人群的审美心理和消费心理而特意打造的，因此观影人群留下的电影评论是对该电影最直观且最具体的评价。电影评论不仅反映出观众的观影情绪，更反映了观众对于该电影在剧情、演员、价值观和表现手法等方面的看法。因此充分挖掘电影评论背后的信息，可以更加精确地了解观影人群的观影喜好和电影的综合表现。

本研究尝试通过基于Scrapy框架的Python爬虫，爬取猫眼平台上《我不是药神》的电影短评信息，使用文本挖掘方法分析该电影的观影用户结构特点、评论趋势、观影满意度，帮助电影制片商了解大众对于电影的观影感受和评论细节，并为其提供电影改进的参考建议。

关键词：电影短评；文本挖掘；Scrapy；评论观点

SHORT-TERM ANALYSIS OF FILM BASED ON TEXT MINING.

ABSTRACT

Because the film is specially created according to the aesthetic psychology and consumer psychology of the audience, the movie review left by the audience is the most intuitive and specific evaluation of the film. The film review not only reflects the audience's viewing mood, but also reflects the audience's views on the story, actors, values and expression techniques. Therefore, fully exploiting the information behind the film reviews, you can more accurately understand the viewing preferences of the audience and the comprehensive performance of the film.

This study attempts to use the Scrapy framework-based Python crawler to crawl the short commentary information of "Dying to Survive" on the MaoYan Movie, using text mining method to analyze the film's viewing user structure characteristics, commenting trend, viewing satisfaction, and help the film Producer understand the public's perception of the film and the details of the comments, and provide reference suggestions for film improvement.

Key words: Chinese film short review; Text mining; Scrapy; Commentary

目录

1.绪论

1.1 研究背景

2018年上半年全国电影票房达320.21亿元，同比增长18.0%。2017年以来在国家政策引导以及观众“口碑”效应等因素的影响下，电影市场复苏明显。国民随着生活水平地提高，逐渐开始拥有了观影的习惯，电影成为了人们不可或缺的文化娱乐活动，不仅如此，人们还热衷于参与到时下热映电影的激烈讨论当中。

伴随着互联网技术Web2.0的发展，网民不只局限于在网络上浏览网页，更倾向于分享自己对事物的想法。络绎不绝的网民开始在各大电影社区上留下他们自己对于电影剧情、导演、演员、电影特技、电影特效等方面内容的看法，表达自己的观影喜好并分享自己内心的感受为电影产品打造“口碑”。在国内的观影平台，电影社区或APP，如豆瓣电影、Mtime时光网、猫眼电影、淘票票、微博电影等，都开设了相关影评平台让用户分享自己的观影感受，然而用户分享的这些电影评论日积月累，变成了海量的数据资源。本是宝贵的数据资源却面临着缺乏加以挖掘和利用的尴尬境地，这些电影短评信息大多都带着用户自身强烈的情感色彩，且绝大多数的影评都是积极和消极的情绪混合在一起，用户在互联网上产生的数据信息以指数增长的形式快速增长，海量的互联网数据充斥着人们的眼球，这造成了一个我们不得不去解决的难题“互联网信息爆炸”，即互联网中虽然存在着海量的信息可以让人们去分析研究，但人们可以直接从中提取到的有效信息变得越来越少，越来越困难。

1.2 文本挖据研究现状

文本数据挖掘(Text Mining)是指从文本数据中提取具有一定价值的信息和知识的计算机处理技术，他属于信息挖掘的一个研究分支，常用于文本数据信息的知识发现。随着互联网时代的到来，用户可以获得的数据涵盖了从生产资料、商业数据到新闻媒体、娱乐报导等多种类型和形式的文档，形成了一个非常宏大的具有异构性、开放性特点的分布式数据库，而这个数据库中存放的是非结构化的文本数据。文本挖掘利用机器学习，统计方法，算法设计等计算机处理技术，分析大量的非结构化的文本数据（如文本文档、表格数据、网页内容等），抽取或标记相应关键词的概念、词语之间的语义关系，并按照相应内容对文档进行分类从而获取有价值的知识和信息。

文本挖掘的研究最早起源于国外，在20世纪50年代末，H.P.Luhn对文本挖掘领域进行了开创性的研究，将词频统计思想应用于文本数据的自动分类。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，拥有许多研究成果和发明专利并推出了许多著名的文本挖掘工具其中有:IBM的文本智能挖掘机、Autonomy公司的ConceptAgents、TelTech公司的TelTech等。然而国内对于文本挖掘的研究则较晚，且由于中文存在语法结构复杂特殊，语义变化多端等难点，导致针对中文文本挖掘的研究发展较慢。

1.3 研究意义

通过挖掘电影短评信息，可以获得某个电影用户整体的满意程度，辅助网络平台分析电影产品，帮助电影制片商了解大众对于电影的感受，认同度与评论细节，有利于电影制片商创作出高质量的电影。互联网上的观影平台，电影社区或APP，如豆瓣电影、Mtime时光网、猫眼电影、淘宝电影-淘票票、微博电影等只有对所有用户评分综合后再求平均值得出的综合评分来表示这部电影的综合表现，并不能反映出观影群体的具体观影感受，也不能反映出观众对于电影的具体评论细节，无法让电影制片商准确获取到观众具体的审美偏好信息，文本挖掘的方法可以充分地挖掘电影短评中观众背后最真实的感受，与数值化的数据分析相结合可以让电影制片商对电影的综合表现有更加全面的认识，了解到观众更多的评论细节。

根据易观发布的《2017中国电影在线票务市场年度综合分析》中猫眼电影、淘票票、娱票儿所占的市场份额分别位列前三，依次为：26.41%、20.06%、17.49%，因此猫眼电影平台上的影评信息对于该电影的整体评价具有一定的代表意义。

1.4 研究思路

本文的研究思路如图1-1，具体步骤如下:

1、通过基于Scrapy框架的Python爬虫，爬取猫眼电影观影平台上《我不是药神》该电影的电影短评，评论日期，评论时间，用户ID，用户等级，性别，所在城市，评分，用户昵称等相关数据。

2、对获取的评论文本信息进行文本预处理，如分词，去停用词，去除纯数字组合，纯空格组合，特殊字符等文本噪音。

3、针对清理后的评论文本数据，分析该电影的观影用户结构特点、评论观点、评论趋势、观影用户满意度。

4、根据分析结果，反馈用户群体的综合评价和电影改进的参考建议。

图1-1 研究思路

2.网络爬虫

2.1 网络爬虫的概念与反爬虫策略

2.1.1 网络爬虫的概念

网络爬虫又称为“网络蜘蛛”，是通过网站的URL来寻找页面，然后读取该网页的信息查找该网页中的其它URL，然后通过该网页中的其他URL查找下一个网页，循环直到没有其他URL为止并按照某种逻辑来抓取所需的网页数据保存到指定存储系统中的技术。“网络蜘蛛”按照自身系统结构和实现技术的不同，分为以下三种类别:通用网络爬虫(General Purpose Web Crawler)、主题网络爬虫(Topical Web Crawler)、深层网络爬虫(Deep Web Crawler)，在爬虫应用设计中常常融合这三种爬虫技术，借助各自技术的优点从而达到爬虫性能最优、抓取数据最全面的目的。

2.1.2 反爬虫策略

网站服务器面对网络爬虫时，会需要消耗大量的资源去回应网络爬虫的请求，同时这些回应会导致网站数据未经授权就被他人获取，最后这些未经授权的数据的使用用途和目的，是网站管理者不可控制的。基于保护服务器资源，数据安全等方面原因，很多网站会采取反爬虫技术阻止网络爬虫的爬取行为。

常见的反爬虫策略有:

(1) 频率限制：禁止每分钟超过N次访问的IP访问 。

(2) 诱捕：用一些人类看不到的链接，当有IP访问该链接时，封禁对应IP。

(3) 账户登陆：使用账号登陆，才能查看到网页的某一部分重要内容。

(4) 输入验证码：当用户访问次数在一定时间内达到一定的数量时，自动让请求跳到验证界面，需要输入验证码后才能浏览网页内容。

(5) 异步加载：不直接加载整个网页内容，而是等待用户操作后，才会加载网页中相应部分的内容。

2.2 基于Scrapy框架的Python爬虫

2.2.1 Scrapy的架构

Scrapy是一个针对爬取网站提取网页结构性数据的实用型型框架,他可以应用在数据挖掘，信息处理或存储历史数据等一系列的程序中，是一个通用的网络爬虫，由五大组件组成。它们分别是调度器(Scheduler)、下载器(Downloader)、爬虫（Spider）和实体管道(Item Pipeline)、Scrapy引擎(Scrapy Engine)，组件直接之间相互作用形成数据流，Scrapy中的数据流由执行引擎控制，如图2-1所示，具体过程如下:

1、引擎打开一个网站(open a domain)，找到处理该网站的Spider并向该spider请求第一个要爬取的URL。

2、引擎从Spider中获取到第一个要爬取的URL并在调度器(Scheduler)以Request调度。

3、引擎向调度器请求下一个要爬取的URL。

4、调度器返回下一个要爬取的URL给引擎，引擎将URL通过下载中间件(请求(request)方向)转发给下载器(Downloader)。

5、一旦页面下载完毕，下载器生成一个该页面的Response，并将其通过下载中间件(返回(response)方向)发送给引擎。

6、引擎从下载器中接收到Response并通过Spider中间件(输入方向)发送给Spider处理。

7、Spider处理Response并返回爬取到的Item及(跟进的)新的Request给引擎。

8、引擎将(Spider返回的)爬取到的Item给Item Pipeline，将(Spider返回的)Request给调度器。

9、(从第二步)重复直到调度器中没有更多地request，引擎关闭该网站。

图2-1 数据流(Data flow)

2.2.2 目标网页分析

通过PC端访问该电影首页，PC端的用户在短评区域只能浏览热门短评和写短评，无法查看更多的短评,这给本文的爬取任务带来了难题。但是我们尝试着通过Chrome浏览器的模拟器模拟移动端，我们发现了查看全部电影短评的入口，《我不是药神》电影评论的移动端页面如图2-2所示。

图2-2 《我不是药神》电影首页（移动端）

点击查看全部短评。当我们下滑操作的时候,我们可以在Name面板下看到新加载进来的资源。点击资源名称我们可以查看该资源的具体内容,其中包含了该资源的Headers，Preview，Response，Cookies，Timing标签，具体信息如表2-1所示。

表2-1 Name面板信息

标签名称	标签内容
Headers	资源的请求url、响应状态码、请求头和响应头、请求参数等。
Preview	用于资源的预览。
Cookies	显示资源HTTP的Request和Response过程中的Cookies信息。
Timing	显示资源在完整的request生命周期中每个环节所消耗的时间信息。

点击Preview面板，查看该json格式资源的预览信息，可以看到我们所需要爬取的电影短评数据就存放在这个资源当中，因此我们可以通过向该资源发送请求，从而获取该资源的Response信息。

2.2.3猫眼电影反爬虫机制

猫眼电影网采用三种反爬虫机制:限制IP，Ajax异步加载页面和数据限制。

(1) 限制IP

k当同一个IP在短时间内大量地访问猫眼电影平台上同一个页面时，将会被猫眼电影平台地后台服务器屏蔽IP。若依然使用该IP，即便是更换浏览器后重新访问，依然会进入验证界面，且验证页面无法验证，需要更换IP才能重新浏览网页例如：短时间内使用同一IP大量访问，限制页面如图2-4所示。

图2-4 IP限制验证页面

(2) Ajax异步加载页面

目标网页上的评论信息，不是一打开页面就全部加载出来，而是通过下滑操作下滑至页面的最底端，才会将一部分未加载的信息加载到当前页面，Ajax加载如图2-5所示。

图2-5 Ajax异步加载

(3) 数据限制

即便是我们使用下滑操作，下滑页面刷新资源，也不是能查看到所有的数据。当参数Offset=1005时，返回的Jsons数据为空，没有任何数据的加载，相当于每天只能看到下滑至offset=990的评论信息，数据限制页面图2-6所示。

图2-6 数据限制

2.2.4 Python爬虫设计

图2-7 爬虫框架结构

通过Scrapy框架搭建的爬虫结构如上图2-7所示，Spiders目录下存放爬虫脚本文件maoyan_spider.py，该爬虫任务被命名为maoyan，本次爬取任务中没有启动中间件,爬虫框架文件具体说明如表2-2所示。

表2-2 爬虫框架文件信息

脚本文件名称文件功能描述

items.py 定义爬取数据文件

Middlewares.py Spider中间件(Middleware)下载器中间件文件

Pipelines.py Item Pipeline文件

Settings.py Scrapy框架的配置文件

Maoyan_spider.py 爬虫定义文件

Python爬虫代码逻辑如图2-8，具体步骤如下：

图2-8 Python爬虫代码逻辑

1、调用fake_useragent，随机模拟真实的请求头从而模拟浏览器标识，配置原始的url。

2、根据url向服务器发送请求，获取response信息，然后获取定义好的item数据。在获取item之前我们需要明确我们的需求，因为我们只需要爬取我们需要的时间段内的数据。例如我不是药神这部电影2018年7月5号上映，但是其预告片在7月份就发布了，因此在7月5号之前的短评数据，不是完整观影之后的电影短评数据，这一部分的数据就不是我们需要获取的数据。我们需要对当前资源的时间进行判断，当前爬取的时间是否在我们需要的时间段之内（即初始startTime与endTime之间）。startTime<endTime表示资源请求url当前时间在结束时间之前，则爬取完成startTime>=endTime表示资源请求Url当前时间在结束时间之后，是在我们需求的时间段之内，则开始爬取我们已定义好的的item。

3、根据response获取item数据，则若不存在，则输出报错信息，关闭爬虫。INFO: Closing spider (finished)。如果存在所需数据，则提取相应item并更新Url

4、更新url:首先判断url中的offset参数是否小于990。如果offset参数小于990，则offset参数增加15，实现移动端浏览器下滑操作，重复步骤2。

5、如果offset参数大于990，即表示当前时间的电影短评已全部获取完毕，需要更新url中的时间参数startTime。

6、更新时间参数startTime:为了减少重复爬取数据和保证数据的连贯性，我们需要判断startTime（当前请求资源url的时间参数），cmtime（当前爬取的最后一条评论的评论时间）。如果cmTime和startTime相等，则startTime = startTime-datetime.timedelta(days=1)，即startTime更新为前一天（牺牲少部分的数据以便爬虫继续爬取我们所需的item数据）。如果不相等，startTime=cmTIme即startTime更新为我们当前爬取的最后一条电影短评的评论时间（cmTime）保证数据的连贯性。

7、url参数更新完毕后，重复步骤2，直到爬取完成。

我们将接收到的item，通过Pipelines保存到数据库中,数据库表设计如图2-9所示。

图2-9 数据库表设计

3.电影短评分析

图3-1 电影短评数据文件

经数据库导出的电影短评数据文件为DyingToSurvive.csv，大小为8.87M，如图3-1所示。其中包含了91720个用户，猫眼等级0-6，涵盖1105个城市地区，电影评分从0-5，评论日期从2018-12-18到2018-07-06，共92650条的电影短评数据。

3.1文本预处理

(1) 中文分词

中文分词是指将一系列汉字分成单个单词。分词一个按照某些规定将连续的单词序列重新组合成单词序列的过程。众所周知，语法和语义结构在不同的语言中也各不相同。在英语文本中，空格被用作单词之间的自然分隔符，而中文则不使用空格作为自然分隔符，只有在句子和段落之间使用明显的分隔符。虽然英语文本也存在着短语定界的问题，但是中文的语法和语义比英文更加复杂，导致中文分词难度更大。本文借助2018年12月份开源的pkuseg工具包，来帮助我们进行中文分词。pkuseg是由北京大学语言计算与机器学习研究组研制推出的一套基于统计的中文分词工具包，在各大测试集上pkuseg都具有较高的分词准确率，相比于其他的分词工具包，该工具包能够更好地帮助我们准确地进行中文分词，建立良好的文本挖掘基础，pkuseg分词结果如表3-1所示。

表3-1 pkuseg分词结果

电影短评内容 pkuseg分词结果

很感人，很现实 ['很', '感人', ' ', ', '很', '现实']

很现实值得看 ['很', '现实', '值得', '看']

徐峥又一部有代表意义的作品，拿下了金马奖最佳男演员。 ['徐峥', '又', '一', '部', '有', '代表', '意义', '的', '作品', ' ', ', '拿下', '了', '金马奖', '最', '佳', '男', '演员', '。']

(2) 去停用词

所谓“停用词”，表示我们在处理文本的时候，其中与文本内容中的情感信息，或文本主题信息关系性较弱的词语，所以在进行我们的筛选过滤之后，剩下的词语更利于我们进行主题分析，或者情感分析。例如:经过分词之后，我们生成了n个中文单词，在这n个中文单词之间，夹杂着数字，半全角符号，语气词等，这些词语和符号并不是我们需要的，因此我们需要将其去除，最后得到对接下来进行的文本分析真正有意义的m个词语。而在中文文本分析中，常用的停用词表有哈工大停用词表、四

川大学机器智能实验室停用词库、百度停用词表，及网络上的开源停，本文选用的是哈工大停用词表、四川大学机器智能实验室停用词库、百度停用词表这三个停用词表集合去重后的综合停词表，去停用词文本预处理如表3-2所示。

表3-2 去停用词

电影短评内容分词并去停用词后
好看，引人深思的影片好看引人深思影片
徐峥又一部有代表意义的作品，
拿下了金马奖最佳男演员。 徐峥一部代表意义作品
拿下金马奖最佳男演员
还是值得一看的值得一看
好看(。·ω·。)/♡ 好看

3.2电影评论用户结构特点

爬取到的用户数据中，只有用户等级，用户城市这两个维度可以让我们分析电影评论用户结构特点，其中用户城市为部分为省市级，部分为乡镇县区太过于细化，将其统一划分为省市级需要耗费大量的人力时间，因此我们只从用户等级这一角度探讨用户的结构特点。

猫眼电影平台上《我不是药神》的观影用户等级如图3-2所示，用户等级为2的比例最大，约占51.48%，其实是用户等级为1，约占23.32%。等级1-2的用户一共占猫眼平台上该电影全体观影用户的74.8%,而等级为0，新注册的用户只占0.01%。

图3-2 猫眼等级分布

图3-3 猫眼会员级别规则

该电影在猫眼电影平台上的观影人群大多是由来自于用户等级为1-2的用户，根据图3-3猫眼会员级别规则中的猫眼会员成长值规则说明，我们可以合理地推断，填写评论的用户基本上都是猫眼电影平台的正常用户，不是为了刷影评而新注册的“水军”。等级为1的用户没有观影或者写影评的习惯，因为用户注册时通过填写资料拥有了一定的成长值，再通过每日登陆和留下至少两部电影影评就能到达用户等级2，而等级2的用户则是有留下两篇以上，十篇以下电影短评，或是有多日登陆且多次观影行为，正在渐渐养成自己的观影习惯或是写影评习惯的用户，从猫眼电影平台上《我不是药神》的观影用户等级分布说明，在猫眼平台上，该电影吸引了没有经常观影或者写影评习惯的用户选择观影并在网络平台上留下自己的观影感受，主动地推广这部电影的“口碑”。

3.3电影评论趋势

图3-4 日评论趋势分布图

从图3-4日评论趋势分布图中，我们可以看到自2018.07.06《我不是药神》上映，其电影的日评论一直保持在较高水平，截至2018.12.18,平均日评论数达到558条。在2018.08.12这天达到该电影最高日评论数1228条，随后在2018.09.07前后，评论热度开始下降。2018.10.19日该电影的评论热度下降趋势趋于平稳，开始保持在相对较低的水平，于2018.12.18达到最低的日评论数27条。

3.4观影用户满意度

3.4.1 观影用户评分分析

图3-5 用户评分图

由上图3-5用户评分图中可知，72.67%的观影用户给了5分，14.89%的观影用户给了4.5分，8.26的观影用户给了4分，因此可见几乎据大多数的用户，都给了4分以上的高评分。说明该电影得到了广大观影人群的喜欢和赞扬，因此我们试通过用户评分来判断用户的情感倾向，用户评分为0-2.5的电影短评为消极评论，用户评分为3-5的电影短评为积极评论。

图3-6 观影用户性别情感倾向

图3-6 观影用户性别情感倾向图中，观影用户性别情感倾向图中,Y轴中的1代表了男性用户，0代表了女性用户，红色为好评的占比，黑色为差评的占比，数字标签为具体占比值。从图中，我们可以看到，男性用户和女性用户的好评率分别为98%和97%，说明该电影在猫眼平台上，无论是男性还是女性，这部电影都深受他们的喜爱。

3.4.2 观影内容偏好分析

我们分别对情感极性为积极和消极的评论内容分别进行观点分析，通过提取关键词信息，了解该电影具体在哪一方面的内容受到了用户的偏爱，哪一方面的内容不能被观影用户所认可。

在提取关键词信息中，我们使用TF-IDF算法来评估消极评论和积极评论中某个字词对于该评论内容的重要程度。

指　标
疑似剽窃文字表述

1. 随着互联网时代的到来，用户可以获得的数据涵盖了从生产资料、商业数据到新闻媒体、娱乐报导等多种类型和形式的文档，形成了一个非常宏大的具有异构性、开放性特点的分布式数据库，而这个数据库中存放的是非结构化的文本数据。
2. 年代末，H.P.Luhn对文本挖掘领域进行了开创性的研究，将词频统计思想应用于文本数据的自动分类。
3. 著名的文本挖掘工具其中有:IBM的文本智能挖掘机、Autonomy公司的ConceptAgents、TelTech公司的TelTech等。然而国内对于文本挖掘的
4. 网页中相应部分的内容。

2.2 基于Scrapy框架的Python爬虫

2.2.1 Scrapy的架构

Scrapy是一个针对爬取网站提取网页结构性数据的实用型型框架,他可以应用在数据挖掘，信息处理或存储历史数据等一系列的程序中，是一个通用的网络爬虫，

5. 3.1文本预处理

(1) 中文分词

中文分词是指将一系列汉字分成单个单词。分词一个按照某些规定将连续的单词序列重新组合成单词序列的过程。

2. 基于文本挖掘的电影短评分析.docx-z_第2部分

总字数：4448

相似文献列表

去除本人已发表文献复制比：10.5%(467)　　文字复制比：10.5%(467)　　疑似剽窃观点：(0)

1	基于电商评论的手机类商品满意度调查 陈丁越 - 《大学生论文联合比对库》- 2018-04-10	10.5%（467） 是否引证：否

原文内容

其中TF-IDF中的TF指的是该词语在该文档中出现的频率，即某个词在文档中的出现次数除以该文档的总词数。对于在该文档里的词语ti：tfi,j=ni,jknk,j 式1ni,j指的是某个词在该文档dj中的出现次数，而knk,j则是在该文档dj中所有词语的出现总次数。

IDF指的是逆向文件频率，该值越小代表该词越普遍出现。对于在该文档里的某个词语ti，其idfi是由文档的总数除以包含该词语ti文档的总数，经过对数计算后所得:idfi=log|D||j:ti∈dj| 式2

|D|指总文档数目|j:ti∈dj|表示出现该词的文档数目，通常需要表示为1+|j:ti∈dj|，以避免该词在语料库中未出现而导致分母为零。

最终TF-IDF值则为：tfidfi,j=tfi,j×idfi 式3

本文将短评内容分别按照用户评分划分为积极评论内容，消极评论内容，然后分别计算积极评论内容，消极评论内容中的每一个词汇的TF-IDF权值，最终按降序排列截取前100个词汇，作为评论的关键词。鉴于篇幅有限，最终我们选取了排名前30的积极评论内容关键词，消极评论内容关键词作关键词及TD-IDF权重作为关键词的部分展示，如表3-3，表3-4所示。

(1) 积极评论内容关键词及TD-IDF权重：

表3-3 积极评论内容关键词及TD-IDF权重

积极评论内容关键词 TD-IDF值积极评论内容关键词 TD-IDF值

好看 0.8096126704319562 感触 0.049522499897738596

电影 0.3556190791980433 剧情 0.04917029588659145

不错 0.31431849673174256 搞笑 0.04287501710604861

感人 0.29875484038197425 影片 0.04264876488745888

感动 0.13964691359140405 推荐 0.03945319922848491

现实 0.111169676259658 完美 0.0379253706900391

值得一看 0.10458308392589323 特别 0.03782432436327655

值得 0.07639758330760672 超级 0.03543818446058903

徐峥 0.07198057389768742 人性 0.03214011510932758

良心 0.07120845624727493 泪点 0.031325527282311354

真实 0.06885450871292034 意义 0.030130619020294788

喜欢 0.05255241110831077 演员 0.02843675729700217

很棒 0.05182618900905006 希望 0.02806093129387496

演技 0.05057979004610327 题材 0.02801694203630819

社会 0.05015043813623769 国产电影 0.025396108237159664

从表3-3积极评论内容关键词及TD-IDF权重中，我们可以看到在积极的情感评论内容中，除了排在较前的积极情感词之外，现实，值得，感动，徐峥，演技，搞笑，社会，剧情，演员，国产电影等关键词说明观影用户被对这部电影的情节，剧情，反映的社会现状，和演员的演技表示认同。

观众认为这部电影是一部不错好看的电影，是一部值得被推荐与观看的电影，是一部反映现实社会现状的电影，是一部感动人心引人深思的电影，也是一部演技在线的国产喜剧，这部由徐峥主演的国产喜剧，故事情节贴近人们现实的生活，揭示社会现实，备受观众好评，在欢声笑语中将社会中最现实的药价问题娓娓道来，让每个人都在电影中看到了自己的影子。

(2) 消极评论内容关键词及TD-IDF权重：

表3-4 消极评论内容关键词及TD-IDF权重

消极评论内容关键词 TD-IDF值消极评论内容关键词 TD-IDF值

一般 0.3485007252613728 笑点 0.036565499738988844

电影 0.31264962384238576 看不懂 0.033342744381971934

剧情 0.14304602795873697 观影 0.03125197614767902

电影院 0.08335686095492983 票房 0.029084660142263404

垃圾 0.08050257942508816 感动 0.028092442007222024

评分 0.07118733351662469 退票 0.027649526706351207

三观 0.06822629463691976 影片 0.026227551698956457

差评 0.06452735248056854 不怎么样 0.026058764711586902

无聊 0.0525021519209104 煽情 0.026028275654210148

喜剧 0.04912034288657251 不好 0.025509025427356963

泪点 0.047320058485750265 笑点 0.036565499738988844

烂片 0.045149888261353 票价 0.023339921067380354

难看 0.0430671008488665 导演 0.022041015669062615

影院 0.042285300385255485 故事 0.022023658021122706
失望 0.03668482956403023 太吵 0.019755411608555234

在表3-4消极评论内容关键词及TD-IDF权重，除了的消极情感词之外，还有剧情，电影院，评分，票房，煽情，导演，故事，影评，票价评等关键词。我们总结为这几大类的内容，分别为情节内容，观影环境，票价，电影价值观，电影类型，电影的影评及评分。我们通过关键词字符串匹配短评内容的方式，细致地了解观影用户对该电影不满的具体原因。观影用户认为，该电影的情节内容过于循规蹈矩，与电影《达拉斯买家俱乐部》在故事情节上有几分相似，让有看过类似影片的观众感到无聊和乏味，电影只是展示了病人为生命而作抗争，却没有客观地提到药价和药品专利的复杂性，片面地将正版药品供应商设置成“坏人”，误导了人们认为贩卖和购买仿制药是道德的，电影的类型是喜剧但是却没有太多的笑点，沉重的电影气氛让奔着喜剧电影来的观影用户感到失望，在观影之后，认为电影“口碑”及评分也存在虚高的情况。

4.电影改进建议

通过基于Scrapy框架的Python爬虫，我们爬取了猫眼平台上《我不是药神》的电影短评信息，并对观影人群进行了统计分析，对电影短评内容进行了评论趋势，满意度分析。

根据分析结果，我们可以相当直观得看出该电影得到了广大观影用户的喜欢和赞扬，无论是男性还是女性的观众，他们都一致地认为这部电影很符合他们的审美需求，并在网络平台上留下了自己的观影感受，自发地去推广这部电影的“口碑”。自电影上映以来，其热度一直保持在较高水平，日均评论数达到500多条，尽管随着影片的拍片期截至，但是影片的评论的热度并没有出现断崖式的下降，而是螺旋式地下降，依然在发挥余热。人们对于该电影的主要关注点集中在剧情，演员，影片类型，影片背后反映的真实生活与价值观。观众认为这部电影是一部不错好看的电影，是一部值得被推荐与观看的电影,是一部反映现实社会现状的电影，是一部感动人心引人深思的电影，也是一部演技在线的国产喜剧。但是也存在部分观众认为该电影结局采用大团结式的美好结局，在情节上和国外同类的电影在剧情模式上有大部分的雷同从而显得老套，电影的气氛和喜剧电影一贯的风格不同且在电影没有客观地提到药价和药品专利的复杂性，观影之后感到失望，认为电影“口碑”及评分也存在虚高的情况。

为此，本文建议电影制片商可以在电影的创作上提高电影的深度，剧情发展模式有所创新，故事情节突出喜剧元素，保持并巩固该电影创作上的已具有的优势，改善自身的不足，进一步迎合观众的审美偏好。该影片的优势在于导演选用了出色的演员，他们精湛的演技让观众眼前一亮，该电影的创作剧本真实地反映了人们的生活现状，讲述一个走私违禁药品以求自救却与法规相悖的故事，成功地引起观众的共鸣与深思从而赢得了观众的认可。然而可以在电影剧本的创作中添加更多的喜剧元素来与喜剧的内核形成更加强烈的对比，设计更具喜剧元素情节和桥段来展现现实生活中的喜与悲，爱和恨。在保证展现戏剧张力的同时辨证地对探讨药价高昂背后的原因，对我国医疗体制建设中存在的问题进行深入地剖析，药品从研发生产到销售使用需要各个社会部门厂家机构的参与，其中涉及到诸多环节，环环相扣缺一不可，不能简单地将发行高价药的正版药品供应商放在人民利益的对立面。在剧情模式上不一定要采用与国外类似电影的模式，可以因地制宜地结合我国医疗体系制度创作出属于自己的独特模式。相信改进后的电影将更加迎合人们的审美偏好，不久的将来电影制作商会制作出下一部更加优秀的影片。

5.总结

此次基于文本挖掘的电影短评分析中，我们使用了基于Scrapy框架的Python爬虫，爬取了猫眼平台上《我不是药神》的中文电影短评数据，通过统计分析来分析观影用户的结构特点，电影评论趋势，观影用户的满意度，使用TF-IDF算法分析评论细节。在分析结果中，我们得到了猫眼平台上关于《我不是药神》该电影的观影用户等级的分布，观影人群的评论观点，观影感受和内容的偏好，但是在其中，也存在不少的还需要注意的问题。我们默认所有爬取的短评数据都是人们发自内心真实填写的，但是其中有很多短评都是滥竽充数，也存在部分的恶意评论内容，本文尚未对该类型的劣质影评作出有效的识别和处理，面对较难归类的城市信息，还没有找到准确的处理方法去分析用户结构中的用户城市维度，同时对于用户情感的划分也不是十分地准确，存在部分用户给了低分，但是评论内容属于积极评论地情况，但是研究思路与方法在方向上的正确的。需要解决以上问题，则需要重新起一篇论文进行研究概述了，迫于时间有限，篇幅有限的情况下，暂且不做研究。

参考文献

[1]涂小琴. 基于Python 爬虫的电影评论情感倾向性分析[D]. 云南师范大学,2017.
[2]蔡光波. 面向主题的多线程网络爬虫的设计与实现[D].西北民族大学,2017.
[3]唐守忠. 文本挖掘关键技术研究[D]. 北京林业大学,2013
[4]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(31):792-793.
[5]殷复莲. 中国电影评价体系及评价方法研究[D]. 中国传媒大学,2017.
[6]杨郁琪. 基于文本挖掘的用户满意度影响因素研究[D].中北大学,2018.
[7]曹奇敏. 网络信息文本挖掘若干问题研究[D].北京理工大学,2015.
[1]李晓笛. Web文本挖掘技术研究及应用[D].北京交通大学,2015.
[9]威云霞. 中文文本挖掘技术的研究与应用[D].西安电子科技大学,2014.
[10]陈晨. 面向Web文本挖掘的主题网络爬虫研究[D].电子科技大学,2017.
[11]张丽. 文本挖掘中关键词与文本摘要自动提取研究[D].青岛理工大学,2018.
[12]李梅. 文本挖掘中若干关键技术研究[D].西北农林科技大学,2016.
[13]康东. 中文文本挖掘基本理论与应用[D].苏州大学,2014.
[14]沈静. 浅析中文分词方法[J].漳州职业技术学院学报,2016,18(03):45-48.
[15]王迪. 文本挖掘中的中文实体关系抽取[D].北京邮电大学,2013.
[16]张彦. web中文文本的数据挖掘技术研究[D].山东大学,2011.
[17]徐德. 关于互联网文本数据挖掘的一些关键技术研究[D].电子科技大学,2011.
[18]何慧. WEB文本挖掘中关键问题的研究[D].北京邮电大学,2009.
[19]刘宁. 客户评价挖掘算法研究与实现[D].吉林大学,2009.
[20]Xu Sun, Houfeng Wang, Wenjie Li. Fast Online Training with Frequency-Adaptive Learning Rates for Chinese Word Segmentation and New Word Detection. ACL. 253–262. 2012

致谢

首先感谢北京师范大学珠海分校，在这里我不仅学会了许多专业相关的知识、开拓了自己的视野，还收获了令人难忘的友情。特别感谢崔德鑫老师，在我迷茫的时候总是能给予指引，无论是在学习中还是在生活中，最后由您作为指导老师让我感到非常幸运。同时也感谢四年间教导过我的所有老师，教授我宝贵的知识。最后感谢所有的朋友，有你们的陪伴大学生活变得丰富多彩，以后也将成为难忘的记忆。

指　标
疑似剽窃文字表述

1. 里的词语ti：tfi,j=ni,jknk,j 式1ni,j指的是某个词在该文档dj中的出现次数，而knk,j则是在该文档dj中所有词语的出现总次数。
   IDF指的是逆向文件频率
2. 对数计算后所得:idfi=log|D||j:ti∈dj| 式2
   |D|指总文档数目|j:ti∈dj|表示出现该词的文档数目，通常需要表示为1+|j:ti∈dj|，以避免该词在语料库中未出现而导致分母为零。
   最终TF-IDF值则为：tfidfi,j=tfi,j×idfi 式3
   本文将
3. 解决以上问题，则需要重新起一篇论文进行研究概述了，迫于时间有限，篇幅有限的情况下，暂且不做研究。

致谢
首先感谢北京师范大学珠海分校，在这里我不仅学会了许多专业相关的知识、开拓了自己的视野，还收获了令人难忘的友情。特别感谢崔德鑫老师，在我迷茫的时候总是能给予指引，无论是在学习中还是在生活中，最后由您作为指导老师让我感到非常幸运。同时也感谢四年间教导过我的所有老师，教授我宝贵的知识。最后感谢所有的朋友，有你们的陪伴大学生活变得丰富多彩，以后也将成为难忘的记忆。

说明：1.总文字复制比：被检测论文总重合字数在总字数中所占的比例

2.去除引用文献复制比：去除系统识别为引用的文献后，计算出来的重合字数在总字数中所占的比例

3.去除本人已发表文献复制比：去除作者本人已发表文献后，计算出来的重合字数在总字数中所占的比例

4.单篇最大文字复制比：被检测文献与所有相似文献比对后，重合字数占总字数的比例最大的那一篇文献的文字复制比

5.指标是由系统根据《学术论文不端行为的界定标准》自动生成的

6.红色文字表示文字复制部分;绿色文字表示引用部分;棕灰色文字表示作者本人已发表文献部分

7.本报告单仅对您所选择比对资源范围内检测结果负责