盐 城 师 范 学 院

毕 业 论 文 开 题 报 告

题　　目: 新闻热点爬取与可视化的分析与实现

姓　　名: 蔡同波

二 级 学 院: 数学与统计学院

专　　业: 信息与计算科学

班　　级: 数 15（5）信计

学　　号: 15213526

指 导 教 师: 丁炫凯　　职 称: 副研究员

2019 年 3 月 10 日

一、研究的目的、意义与应用前景等：

研究目的：

本文主要通过对新闻文本数据采集、分词和词频词性统计得到新闻热点词，并对热点数据提供可视化信息和分析，以帮助公众了解重要事件关注情况。

研究意义：

本文通过可视化技术可以直观的计算统计出当前新闻的热点情况，为新闻决策和异常监控提供依据，以减小负面新闻对人们生活的影响，及时掌握事物的发展情况。

应用前景：

随着信息量的不断增加，在同一新闻主题下拥有各种各样的新闻报道，大量的新闻出现在人们的生活中。由于在互联网上发布新闻对个人用户没有太多的限制，一些不法分子常常通过恶意“刷榜”、编造谎言恶意引导舆论方向，这可能会给社会带来不少的不良影响和危机。因此，很有必要检测分析当前新闻中的热门信息情况，为热点聚焦和防控提供一定的参考。

二、研究的内容和拟解决的主要问题：

研究的内容：

本课题拟对社会消费品零售总额的组合预测模型展开研究，具体内容如下：

第一部分，引言。介绍了本课题的研究背景和意义，研究现状、总体任务和结构安排。

第二部分，基于新闻 API 以及 BeautifulSoup 模块的数据抓取。介绍了网络爬虫的概念、采取的爬虫算法和抓取新闻的方式。

第三部分，基于交叉信息熵算法的关键字提取。详细介绍介绍分词前的各种处理、分词算法以及词频词性统计过程，并对抓取的新闻数据进行 JSON 化处理。

第四部分，实验结果及数据可视化。设计并实现数据的可视化呈现数据背后的深层含义，得出研究的最终结果。

第五部分，总结与展望。通过对数据的可视化结果分析、得出研究的结论和对未来展望。

拟解决的主要问题：

通过对新闻数据进行分析，可以得出人们对某个领域的关切程度和社会需要解决的问题，有利于了解当前的舆论焦点，有助于政府了解民意，便于国家对舆论进行正确引导。本文以中国新闻网网站为例，通过爬虫技术抓取社会新闻版块数据，再对采集到的新闻内容进行处理及词频词性分析来计算关键字，最后借助数据可视化技术提供直观的呈现，得出热点所涉及的相关人或事物，以此分析出社会领域关注问题和需要解决的问题情况。

三、研究思路、方法和当前收集的文献：

研究思路：

本文以中国新闻网网站为例，通过爬虫技术抓取社会新闻版块数据，再对采集到的新闻内容进行处理及词频词性分析来计算关键字，最后借助数据可视化技术提供直观的呈现，得出热点所涉及的相关人或事物，以此分析出社会领域关注问题和需要解决的问题情况。

研究方法：

本课题使用文本特征提取法、可视化分析法、案例研究法及文献研究法。

当前收集的文献：

[1] 第 43 次 CNNIC 中国互联网报告发布[J].中国广播,2019(04):48.

[2] 荣晗.基于分布式的网络爬虫系统的研究与实现[D].电子科技大学,2017.

[3] 尚楚涵.互联网舆情信息挖掘技术研究与实现[D].华南理工大学,2013.

[4] 孙健康.大数据背景下数据新闻的可视化研究[D].南昌大学,2017.

[5] 王宝龙.面向新闻领域的文本数据获取系统的设计与实现[D].北京邮电大学,2010.

[6] 巫函.数据挖掘与可视化技术对新闻阅读体验的改善——以腾讯网在巴西世界杯期间的报道为例[J].西部学刊(新闻与传播),2016(07):53-55.

[7] 高云.大数据时代数据新闻的发展现状探析[J].视听,2019(04):184-185.

[8] 张进浩.数理统计在数据分析中的应用[J].金融经济,2017(12):130-131.

[9] 杨凯利,山美娟.基于 Python 的数据可视化[J].现代信息科技,2019,3(05):30-31+34.

[10] Thelwall, M. A web crawler design for data mining[J]. Journal of Information Science, 2001, 27(5):319-325.

[11] Fatemeh Ahmadi-Abkenari,Ali Selamat. An architecture for a focused trend parallel Web crawler with the application of clickstream analysis[J]. Information Sciences,2011,184(1).

[12] Lishan Deng. Innovative Application of Python in Data Crawling —Chinese Version of Movie Recommendation Platform[J]. Journal of Physics: Conference Series,2019,1168(3).

[13] 甘凌博,员婕.数据新闻的交互可视化策略探析——以 2018 年世界杯报道为例[J].新闻传播,2019(03):37-38+41.

[14] 邓若蕾.大数据时代数据新闻的可视化传播效果探究[J].传播力研究,2019,3(04):32.

[15] 娄立原,周选州.可视化数据新闻实践探析[J].青年记者,2019(02):33-34.

[16] 林文涛,陈伟强,刘杭燕,叶楠.面向热点新闻的爬虫系统设计与实现[J].数字通信世界,2019(01):261-262+132.

[17] Ali Mesbah,Arie van Deursen,Stefan Lenselink. Crawling Ajax-Based Web Applications through Dynamic Analysis of User Interface State Changes[J]. ACM Transactions on the Web (TWEB),2012,6(1).

[18] Shaojun Zhong,Zhijuan Deng. A Web Crawler System Design Based on Distributed Technology[J]. Journal of Networks,2011,6(12).

[19] 鲁义轩.整合与分析:发挥大数据价值的两大关键[J].通信世界,2013(20):48.

四、特色或创新之处：

本文主要利用多线程技术和爬虫算法实现了对中国新闻网的新闻的并行爬取。利用中科院的 ICTCLAS 分词、交叉信息熵算法分别对抓取的 2019 年 3 月 1 日到 31 日期间的社会新闻进行数据处理、词频词性统计以及 JSON 化处理。利用 PyEcharts 技术对数据结果进行分类统计并绘制可视化视图。通过上述功能的实现，本文完成了毕业设计的预期任务，实现了预定目标。

五、研究计划及预期进展：

第一阶段（2018.7.4—2018.12.25）

确定课题，并根据课题查阅和收集资料,确定论文的写作提纲,交指导老师审阅。

第二阶段（2018.12.26—2019.3.10）

按照指导老师审阅后的论文提纲进行开题报告的填写,交指导老师审阅。

第三阶段（2019.3.11—2019.4.8）

完成毕业论文的中期检查报告和外文资料的翻译，参加学院组织的毕业论文中期检查。

第四阶段（2019.4.9—2019.4.26）

根据提纲,进一步收集、整理和分析资料,撰写论文,形成初稿,交指导老师审阅。

第五阶段（2019.4.27—2019.5.15）

根据指导老师的指导意见反复修改、充实、完善，最后形成终稿，提交查重报告，论文提交审查。

第六阶段（2019.5.16-2019.5.24）

根据论文评阅人意见进一步修改论文，准备论文答辩。

毕业论文开题报告评定表

指导教师意见	该课题的选择具有实际意义，能够从生活中来再运用到生活中去，有着明确的目的和良好的运用前景.该选题最终的结论将有利于更好的对新闻热点问题进行分析和预测，为政府各相关部门制定有关政策和进行有效的宏观调控提供合理的建议. 该生选题较为新颖，搜集的文献具有很强的针对性，研究思路清晰、准备充分，研究方法可操作性强.基于上述理由，同意本选题作为毕业论文选题。 指导教师签名： 2019 年 3 月 10 日
答辩小组审核意见	**同意指导老师意见。** 组长签名： 2019 年 3 月 10 日
答辩委员会审核意见	**同意答辩小组意见。** 答辩委员会主任签字（盖章）：_____ 2019 年 3 月 11 日
备注	

盐城师范学院

毕业论文外文资料翻译

学　　院：数学与统计学院

专 业 班 级：数 15（5）信计

学 生 姓 名：蔡同波　　学　号：15213526

指 导 教 师：丁炫凯

外 文 出 处：Journal of Physics: Conference Series,2019,1168(3).

附　　件：1. 外文译文；　　2. 外文原文

指导教师评语：

译文条理清晰，结构完整，语句通顺流畅，专业术语使用精确，符号规范统一，是一篇较好的外文资料翻译。个别专有名词翻译不够准确，仍需继续努力。在翻译过程中，能主动询问老师，查阅资料，求助同学，学习态度值得肯定。

签名：

2019 年 4 月 8 日

1．外文译文

Python 在数据抓取中的创新应用

-中文版电影推荐平台

摘要：计算机硬件和 Internet 的快速发展使网络的信息传输速度和存储容量爆炸式增长。基础技术的发展导致许多为公众服务的多媒体和富文本技术的进步。为了解决用户和公司需要处理和检索大量信息的问题，网络爬虫技术从网页中提取指定的信息然后对其进行分析，从而提供从大量信息中爬行信息的方向性功能。本文介绍了使用 Python 语言从网站上详细提取特定电影信息的 Web 爬虫，并分析了用 Python 语言编写的 Web 爬虫的详细过程。

1 简介

随着互联网技术的发展，这个时代的数据急剧增加。根据 IDC（国际数据公司）的研究报告，到 2020 年全球数据规模将达到 44ZB。因此，数据挖掘是一种有效的解决方案，广泛用于获取，存储，管理和分析大型数据库中的有用数据。

本文研究如何使用数据捕获技术在大量电影信息数据源中获取有效数据，然后根据我们捕获的数据分析用户的电影偏好，以制作准确的电影推荐。本文介绍了使用 Python 语言进行 Web 爬虫的数据挖掘方法，并从 Web 爬虫的具体介绍，环境配置，数据采集，结果分析以及数据挖掘的前瞻应用等方面进行了解释。

2 文献综述

Web 爬虫是一种程序或脚本，它根据程序[3]设计的某些规则自动抓取网页的特定信息。爬虫技术已经开发了很长时间，并且已经有许多开源库和开源框架可用。在本文中，将使用 Scrapy 的开源框架，它是一个用 Python 编写的 Web 爬虫框架，它使得爬虫的设计和工作变得简单快捷[4]。Scrapy 的架构由八个组件组成：Scrapy Engine，Scheduler，Downloader，Spiders，Pipeline，Downloader Middlewares，Spider Middlewares 和 Scheduler Middlewares。广泛使用的 Web 爬虫程序主要分为以下类型：通用 Web 爬虫程序，聚焦爬虫程序，增量 Web 爬虫

程序和深度 Web 爬虫程序。本文重点介绍了聚焦爬虫。聚焦爬虫在已定义的主题相关页面[5]中定位已爬虫目标网页。此时，可以大大减少爬虫爬行所需的带宽资源和服务器资源。爬虫应用程序存在于生活的各个方面。例如，各种行业数据分析，尤其是在线应用程序倾向于使用网络爬虫。由于行业市场预测可以根据获得的数据分析[6]。同时，网络抓取工具还可以监控网络评论并抢救互联网上的故障单。

3 方法论

3.1 环境准备

首先，您需要安装 Python。然后，您需要配置环境变量。Scrapy 是一个用于抓取网站数据并提取结构数据的应用程序框架。它可以应用于一系列程序，包括数据挖掘，信息处理或存储数据。首选安装 pip 并在 Python 包中设置相关工具是首选。安装 Scrapy 可以使用 pip 完成，在命令提示符下键入“pip install Scrapy”。Scrapy 的结构如下图 1 所示：

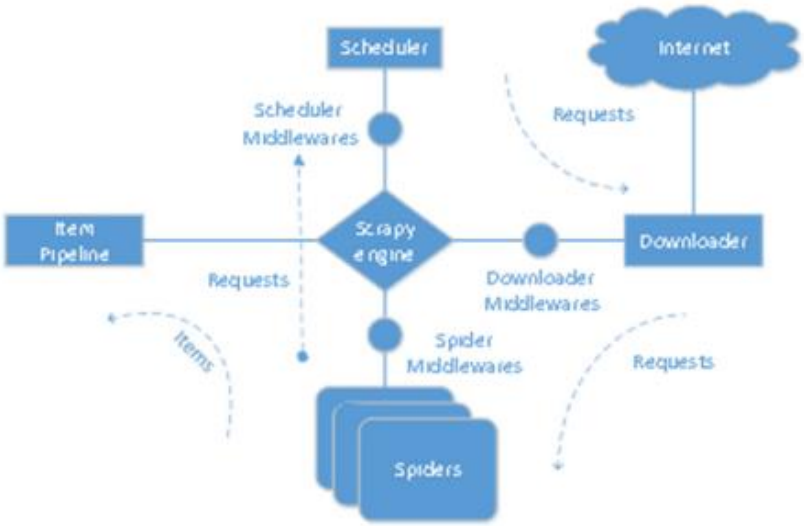

图 1. Scrapy 框架原理

3.2 计划结构

3.2.1 功能模块

3.2.1.1 获取链接

```
def get_html(web_url):
    header = { #head
        "User-Agent":"Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US) AppleWebKit/534.16
(KHTML, like Gecko) Chrome/10.0.648.133 Safari/534.16"}
    html = requests.get(url=web_url, headers=header).text #response
    Soup = BeautifulSoup(html, "lxml")
    data = Soup.find("ol").find_all("li")    # filter and get the information needed
    return data
```

第一步是获取网页的标题信息。request. get()函数根据 URL 连接信息返回响应信息。“. text”将响应对象转换为 str 类型。Beautiful Soup 是一个 Python 库，可以从 HTML 或 XML 文件中提取数据。find_all()函数查找“ol”标签下的所有“li”标签。

3.2.1.2 信息提取和处理

```
def get_info(all_move):
    f = open("E:/py_pro/result/douban2.txt", "a")

    for info in all_move:
        #       rank
        nums = info.find('em')
        num = nums.get_text()

        #       name
        names = info.find("span")
        name = names.get_text()
```

对于排名功能，找到“em”标签可以获得目标电影的排名。对于电影名称，找到“span”标签可以获得电影名称。

```
#       author
    charactors = info.find("p")
    charactor = charactors.get_text().replace(" ", "").replace("\n", "")
    charactor = charactor.replace("\xa0", "").replace("\xee", "").replace("\xf6", "").replace("\u0161",
"").replace(
    "\xf4", "").replace("\xfb", "").replace("\u2027", "").replace("\xe5"
```

为了理清作者信息，有许多标点符号干扰数据处理，因此需要更换。更换后，信息可以整齐排列。在评论部分，我们需要确定是否有评论。然后系统可以继续进行文档处理或直接输出“无备注”。

```
def Load_data(fh_train):
    fh_train = open(fh_train)
    trainSet = np.zeros((944, 1683))
    dict1 = { }
    for lines in fh_train:
        user, item, score, _ = lines.strip().split('\t')
        dict1.setdefault(user, {})
        dict1[user][item] = float(score)
        trainSet[int(user)][int(item)] = dict1[user][item]
    trainSet = np.delete(trainSet, 0, 1)
    trainSet = np.delete(trainSet, 0, 0)
    fh_train.close()
    return np.mat(trainSet)
fh_train = 'E:/py_pro/ml-100k/u1.base'
trainSet = Load_data(fh_train)
np.savetxt('trainSet.csv', trainSet, delimiter = ',')
```

3.2.1.3 用矩阵分析分数信息

3.3 数据处理方法

在大量网页信息中，程序需要处理无效信息以便找到诸如等级，电影名称，导演和评论之类的目标信息。本文能够筛选已爬虫的目标站点信息。过滤器将在网页的
标记中获取电影信息，然后进行处理。网页中存在大量非法信息，网页爬虫无法和谐地处理，例如评论中的标点符号。这类非法信息已在项目中被取代，从而实现统一的信息处理。由于数据在采集后是连续的，因此需要将其切割以便以后处理。在该程序中，用户，项目和分数被分成单个单元用于数据切片。处理获取的数据时，数据库中的密钥和值应该对应，注释的值需要与二维用户和项目一致。

4. 应用和功能的讨论

4.1 矩阵处理

在数据处理过程中，本文采用矩阵方法进行处理和分析。首先，该程序创建一个零矩阵，然后将用户，电影和注释等信息导入矩阵。该矩阵解释了用户和电影之间的关系。不同的分数代表不同的含义。例如，5 分意味着电影值得推荐。如下图 2 所示：

5.00	3.00	4.00	3.00	3.00	0.00	4.00	1.00	5.00	0.00	2.00	0.00
4.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4.00	0.00	0.00	0.00	0.00	0.00	2.00	4.00	4.00	0.00	0.00	4.00
0.00	0.00	0.00	5.00	0.00	0.00	0.00	0.00	5.00	0.00	3.00	5.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	4.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.00	0.00	4.00	5.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.00	0.00	0.00	2.00	0.00
0.00	0.00	0.00	5.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3.00	0.00	0.00	5.00	1.00	0.00	2.00	4.00	3.00	0.00	0.00	5.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5.00	0.00	0.00	5.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00	5.00
0.00	0.00	0.00	0.00	0.00	0.00	4.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	3.00	0.00	5.00	0.00	5.00	5.00	0.00	0.00	0.00

图 2.数据处理矩阵

4.2 运行结果

本文中的项目程序将从电影页面中获取电影的排名，名称，作者和评论。然后程序使用 XPath 创建一个函数，以一定的顺序将获取的电影数据存储在. txt 文件中。如图 3 所示：

1、肖申克的救赎
导演:弗兰克·德拉邦特FrankDarabont主演:蒂姆·罗宾斯TimRobbins/...1994/美国/犯罪剧情
希望让人自由。
9.6

1.The Shawshank Redemption
Director: FrankDarabont Actor: TimRobbins/... 1994/American/Criminal
Hope is a good thing, maybe the best of things, and no good thing ever dies.
9.6

图 3.数据存储的格式

4.3 分析过程中遇到的问题，以及如何处理

在本文中，遇到了难以识别标点符号等非法字符字符的问题。因此，使用可识别和处理的字符替换非法字符的方法。作为值的评级信息需要与两个密钥相关联。通过使用这个二维密钥对，程序可以保证准确和正确的一致关系。某些网站的验证过程是一个挑战，如验证码，拖动图片和选择图片。而创新的验证方法越来越难以处理。某些网站的防火墙会限制来自同一 IP 的请求的次数。但通常 IP 限制不是针对网络爬虫设计的，而是用于处理 DOS 攻击。代理 IP 可以请求更多，但它仍然受到限制。这个问题绝对无法解决。在某些情况下，程序会部分收到 403 错误，因为服务器没有响应恶意攻击。这里可以使用伪造的标题来帮助程

序伪装成 Web 浏览器。伪装过程包括从网站上找到标题并将其复制到程序中。

5. 结论

数据挖掘在数据爆炸式增长的情况下起着非常重要的作用[7]。本文首先介绍了数据挖掘的发展，并详细介绍了数据挖掘的理论、流程和应用。同时，本文主要讨论网络爬虫作为数据挖掘的重要组成部分。此外，本文还详细介绍了 Web 爬虫及其相关架构的开发。该项目使用 Python 语言构建基于网络爬虫抓取的电影排名网站信息的电影推荐系统，并进行测试和数据分析。这个基于网络爬虫的电影推荐项目将极大地方便选择电影的观众，并将作为领导数据挖掘技术服务的良好范例。

2．外文原文

Innovative Application of Python in Data Crawling —Chinese Version of Movie Recommendation Platform

Abstract: The rapid development of computer hardware and Internet makes the information transmission speed and storage capacity of the network explode. The development of basic technology has led to the advance of many multimedia and rich text technologies that serve the public. In order to solve the problem that users and companies need to process and retrieve massive amounts of information, web crawler technology extracts the specified information from web pages and then analyzes it, providing a directional function of crawling information from large amounts of information. This article explains the web crawler that uses Python language to extract specific movie information from the website in detail and analyzes the detailed process of web crawler written in Python language.

1. Introduction

With the development of Internet technology, data has explosively increased in this era. According to the research report of IDC (International Data Corporation), the global data size will have reached 44ZB by 2020. Therefore Data Mining is an effective solution, which is widely used for gaining, storing, managing, and analyzing useful data in large databases [1].

This paper studies how to use data capture technology to get valid data in massive movie information data source, and then analyze users' movie preferences based on data that we captured to make accurate movie recommendations. This article introduces the data mining method of web crawler using Python language, and explain it from the specific introduction of web crawler, environment configuration, data acquisition, result analysis, and prospective application of data mining [2].

2. Literature Review

A web crawler is a program or script that automatically crawls specific information of web page in accordance with certain rules designed by the programme [3]. The crawler technology has been in development for a long time, and there are already many open source libraries and open source frameworks available. In this article, the open source framework of Scrapy, which is a web crawler framework written in Python that makes the design and work of crawlers simple and fast, will be used [4]. The Scrapy's architecture consists of eight components: Scrapy Engine, Scheduler, Downloader, Spiders, Pipeline, Downloader Middlewares, Spider Middlewares, and Scheduler Middlewares. Widely-used web crawlers are mainly divided into the following types: general purpose web crawlers, focused crawlers, incremental web crawlers, and deep web crawlers. This article focused on the focused crawler. The focused crawler locates the crawled target webpage in a pre-defined topic-related page [5]. At this time, the bandwidth resources and server resources required for the crawler crawling can be greatly reduced. Crawler applications exist in every aspect of life. For example, various industry data analysis, especially online applications tend to use web crawlers. Since the industry market forecast can be made according to the acquired data analysis [6]. At the same time, web crawlers can also monitor network comments and rob tickets on the internet.

3. Methodology

3.1. Environmental Preparation

Firstly, you need to install Python. Then you need to configure the environment variables. Scrapy is an application framework written to crawl website data and extract structural data. It can be applied to a series of programs including data mining, information processing or storing data. Installing pip and setting up related tools in the Python Package in advance are preferred. Installing Scrapy can be done using pip, typing “pip install Scrapy” at the command prompt. The structure of Scrapy is shown in Figure 1 below.

Figure 1. Principle of Scrapy framework

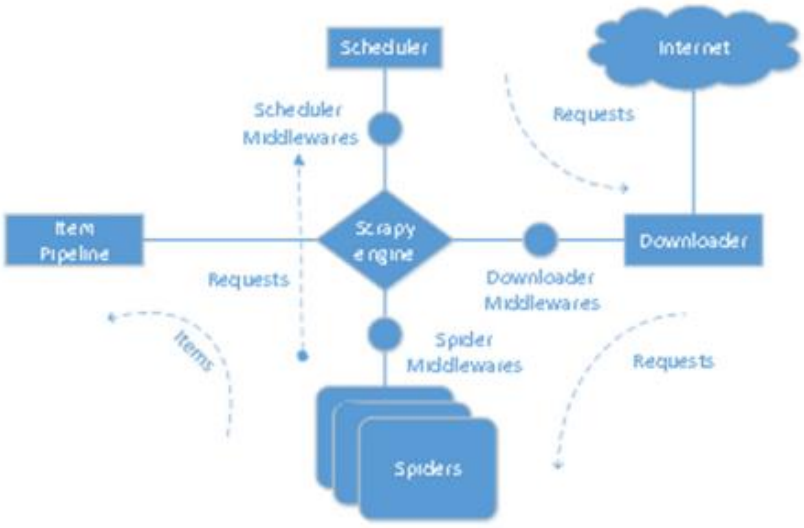

Figure 1. Principle of Scrapy framework

3. 2. Structure of Program

3. 2. 1. Functional Module

3. 2. 1. 1. Gaining of the Source Web

```
def get_html(web_url):
    header = { #head
        "User-Agent":"Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US) AppleWebKit/534.16
(KHTML, like Gecko) Chrome/10.0.648.133 Safari/534.16"}
    html = requests.get(url=web_url, headers=header).text #response
    Soup = BeautifulSoup(html, "lxml")
    data = Soup.find("ol").find_all("li")    # filter and get the information needed
    return data
```

The first step is to get the header information of the webpage. The request.get() function returns response information based on the URL connection information. ".text" converts the response object to str type. Beautiful Soup is a Python library that can extract data from HTML or XML files. The find_all() function looks for all the "li" tags under the “ol” tag.

3. 2. 1. 2. Information extraction and Processing

```
def get_info(all_move):
    f = open("E:/py_pro/result/douban2.txt", "a")

    for info in all_move:
        #       rank
        nums = info.find('em')
        num = nums.get_text()

        #       name
        names = info.find("span")
        name = names.get_text()
```

For the ranking function, finding the "em" tag can get the rank of the target movie. For the movie name, finding the "span" tag can gain the movie name.

```
#       author
    charactors = info.find("p")
    charactor = charactors.get_text().replace(" ", "").replace("\n", "")
    charactor = charactor.replace("\xa0", "").replace("\xee", "").replace("\xf6", "").replace("\u0161",
"").replace(
    "\xf4", "").replace("\xfb", "").replace("\u2027", "").replace("\xe5"
```

To sort out author information, there are a lot of punctuation marks that interfere with data processing, so that needs to be replaced. After replacement, the information can be arranged more neat.

```
# score
scores = info.find_all("span", {"class": "rating_num"})
score = scores[0].get_text()

remarks = info.find_all("span", {"class": "inq"})
if remarks:    # remark
    remark = remarks[0].get_text().replace("\u22ef", "")
else:
    remark = "No remarks"
print(remarks)
```

In the comments section, we need to make sure whether there are comments. Then the system can proceed with document processing or directly output "No remarks".

3. 2. 1. 3. Score information analyzing with matrix

```
def Load_data(fh_train):
    fh_train = open(fh_train)
    trainSet = np.zeros((944, 1683))
    dict1 = { }
    for lines in fh_train:
        user, item, score, _ = lines.strip().split('\t')
        dict1.setdefault(user, {})
        dict1[user][item] = float(score)
        trainSet[int(user)][int(item)] = dict1[user][item]
    trainSet = np.delete(trainSet, 0, 1)
    trainSet = np.delete(trainSet, 0, 0)
    fh_train.close()
    return np.mat(trainSet)
fh_train = 'E:/py_pro/ml-100k/u1.base'
trainSet = Load_data(fh_train)
np.savetxt('trainSet.csv', trainSet, delimiter = ',')
```

3. 3. Data processing method

In a large amount of web page information, the program needs to process the invalid information in order to find target information such as rank, movie name, director and comment. This article are able to filter the target site information that is crawled. The filter will get the movie information in the tag of the web pages and then process it. There are a large number of illegal information in the webpage that the web crawler cannot handle harmoniously, such as punctuation in the comment. Illegal information of this kind has been replaced in the project, so as to achieve unified information processing. Since the data is continuous after the acquisition, it needs to be cut for later processing. In the program, the users, items and scores are divided into single units for data slicing. When processing the acquired data, the key and value in the database should be corresponding, and the value of the comment needs to be in accordance with the two-dimensional pair of user and item.

4. Discussion of Application and Function

4. 1. Matrix Processing

During the data processing, this paper uses a matrix mothod for processing and analyzing. Firstly, this program creates a zero matrix and then import information such as users, movies, and comments into the matrix. This matrix explains the relationship between the user and the movie. Different scores represent diffeerent meanings. For example, a 5-point means that the movie is worthwhile to recommend. As shown in Figure 2 below.

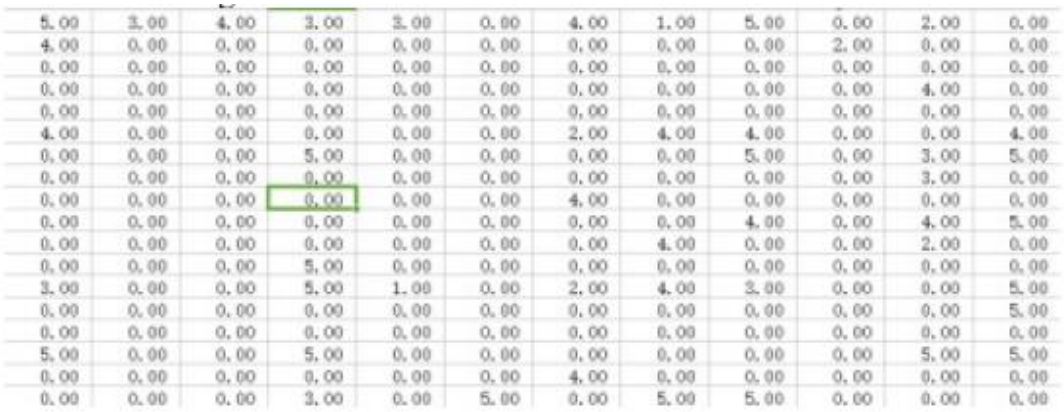

Figure 2. The matrix of Data Processing

4. 2. Operation result

The project program in this paper will grab the movie's ranking, name, author, and comment from the movie page. Then the program uses XPath to create a function that stores the acquired movie data in a .txt file in certain order. As shown in Figure 3

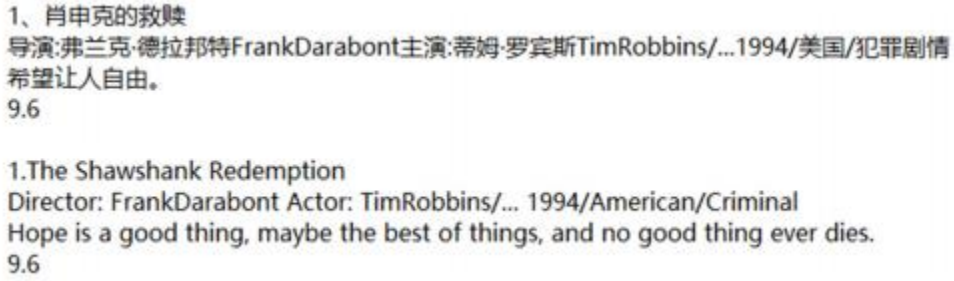

Figure 3. The Format of Data storage

4. 3. Analyze the problems encountered in the process, and how to deal with

In this paper, the problem of difficulty in identifying illegal character characters such as punctuation marks was encountered. The method of replacing illegal characters

with recognizable and processed characters is used therefore. A rating information as a value needs to be associated with two keys. By using this two dimensional key pair, the program can guarantee accurate and correct congruent relationship. The verification process of some websites is a challenge such as verification code, dragging a picture and selecting pictures. And the innovative verification method is getting more and more difficult to deal with. The firewall of some website would limit the times of requests from the same IP. But usually the IP restriction is not designed to aim at the web crawler, it is used deal with DOS attacks. Agent IP which can request more, but it is still been limited. This problem cannot be solved absolutely. Under some circumstance, the program would partly receive 403 error because the server does not respond to malicious attacks. Forged headers can be used here to which would help the program to camouflage as a web browser. Camouflage process includes finding the header from the website and copying it into the program.

5. Conclusion

Data mining plays a very important role in the case of explosive growth of data [7]. This paper begins with the introduction of the development of, and introduces the theory, flow path and application of data mining in detail. Meanwhile, this paper mainly talk about the web crawler as an important part of data mining. What's more, this paper also describes in detail the development of web crawlers and their related architecture. This project uses the Python language to build a movie recommendation system based on movie ranking website information crawled by web crawler, and conducts testing and data analysis. This web crawler-based film recommendation project will greatly facilitate the audience who selecting movies and will work as a good example to lead the data mining technology servicing for our life. In the future, data mining will incorporate various data types such as graphics data, video image data, and sound data [8]. Data mining can excavate and analyze various data in life to provide more convenient support for all aspects of life.

盐城师范学院毕业论文中期检查表

数学与统计 **学院** 信息与计算科学 **专业** 数15(5)信计 **班**

<table>
<tr><td rowspan="5">学生自查</td><td>毕业论文题目</td><td colspan="5">新闻热点爬取与可视化的研究与实现</td></tr>
<tr><td>姓名</td><td>蔡同波</td><td>学号</td><td>15213526</td><td>指导教师</td><td>丁炫凯</td></tr>
<tr><td>根据工作进度安排应完成的任务</td><td colspan="5">1.确定研究课题，根据课题收集资料，确定论文写作提纲；
2.完成毕业论文开题报告；
3.完成相关文献资料的收集、整理和分析工作；
4.撰写论文，形成初稿并交给指导老师审阅。</td></tr>
<tr><td>工作完成情　况</td><td colspan="5">简述：
我在指导老师的帮助下，认真地查阅资料后，明确了研究方向和写作提纲，并根据自己的实际情况拟定了课题研究计划，在得到指导老师的审阅通过后开始了论文初稿的写作工作。初稿已经完成，并交指导老师审阅。
进度：☑按期完成　□基本按期完成　□已拖期</td></tr>
<tr><td>未按期的主要原因及解决办法</td><td colspan="5"></td></tr>
<tr><td rowspan="4">指导老师检查</td><td>工作进度</td><td colspan="5">☑较快　□正常　□较慢</td></tr>
<tr><td>工作质量</td><td colspan="5">☑较好　□一般　□较差</td></tr>
<tr><td colspan="6">具体意见：
1.英文摘要用词不够准确，尚须调整；
2.论文格式较规范，但部分内容还需调整，比如：小标题需进一步概括；
3.中文引言语言不够精练，需指出研究工作的意义；
4.对新闻热点实例分析太过空泛，理论深度不够。
指导老师（签名）：
2019 年 4 月 8 日</td></tr>
<tr><td colspan="6">答辩小组意见：
同意指导老师意见。
组长（签名）：
2019 年 4 月 8 日</td></tr>
</table>

注：1.中期检查要讲求实效，主要是找问题，找差距。对中期检查不合格的学生提出警告。

2.此表一式二份，一份反馈学生，一份交二级学院备案。

3.在相应地方填写或打√。

盐城师范学院
毕业论文答辩记录表

学生姓名	蔡同波	学 院	数学与统计学院	专 业	信息与计算科学
班 级	数 15（5）信计	学 号	15213526	指导教师	丁炫凯
课 题 名 称	新闻热点爬取与可视化的研究与实现				
答辩小组组长	姜海波	答辩小组成员	姜海波，崔蓉蓉，张进兵		
答辩地点	主楼 A415			记录人	张进兵

答辩中提出的主要问题及学生回答问题的简要情况：

问：研究本课题的目的和意义是什么?

答：目的：本文主要通过对新闻文本数据采集、分词和词频词性统计得到新闻热点词，并对热点数据提供可视化信息和分析，以帮助公众了解重要事件关注情况。

意义：本文通过可视化技术可以直观的计算统计出当前新闻的热点情况，为新闻决策和异常监控提供依据，以减小负面新闻对人们生活的影响，及时掌握事物的发展情况。

问：简述本文的主要工作及主要方法.

答：主要工作：本文主要通过对新闻文本数据采集、分词和词频词性统计得到新闻热点词，并对热点数据提供可视化信息和分析，以帮助公众了解重要事件关注情况。

主要方法：本课题使用文本特征提取法、可视化分析法、案例研究法及文献研究法。

问：谈谈本文的创新之处.

答：本文主要利用多线程技术和爬虫算法实现了对中国新闻网的新闻的并行爬取。利用中科院的 ICTCLAS 分词、交叉信息熵算法分别对抓取的 2019 年 3 月 1 日到 31 日期间的社会新闻进行数据处理、词频词性统计以及 JSON 化处理。利用 PyEcharts 技术对数据结果进行分类统计并绘制可视化视图。

答辩小组组长签字：

记录人：

答辩时间：2019 年 5 月 25 日

盐 城 师 范 学 院

毕 业 论 文 成 绩 评 定 表

数学与统计学院 信息与计算科学专业 班级 数 15(5)信计 姓名 蔡同波 学号 15213526

课题名称	新闻热点爬取与可视化的研究与实现

指导教师对毕业论文的评语：

论文通过搜集目前已有的对新闻热点进行预测的一些方法，归纳总结创新，运用数据可视化技术对现有的新闻数据进行分析和预测，最后得出结论。文章条理清晰，结构完整，语言通畅流利，格式整齐，说明作者具备了不错的论文写作能力。文章中分析过程严谨，专业术语使用准确，符号规范统一，参考文献编号齐全，标注清楚，说明作者写作态度认真，文章突破了传统新闻报道方式陈旧，内容抽象化，语言机关化公文化，流于表面，难以让受众真正理解和思考数字的纵深意义。论文揭示事件发展的方向和趋势，提高了预测的精确度，为政府决策和宏观调控提供了切实可行的参考。文章只做了定量分析，没有进行定性和定量的组合分析，这在很大程度上降低了研究的深度，说明作者的归纳探究的能力有待进一步提高。但是，作者写作态度认真，证明严谨，论文写作积极主动。总体来说，该文不失为一篇合格的本科毕业论文。

得分：_____ 指导教师签字：_____

2019 年 5 月 18 日

评阅人对毕业论文的评语：

该论文运用大数据可视化的相关理论知识，对中国新闻网社会新闻版块的新闻热点情况进行了分析预测。该论文选题具有实际意义，立意明确、思路清晰、内容完整、逻辑性较强，对个人、社会和政府有关部门决策和舆论防控有一定的参考价值。论文表明该同学查阅了较多的文献资料，具备一定的文献综述和资料整理能力，已掌握了研究论文写作的一般方法。若能对课题做更深层次的探讨，文章将更有深度。建议按期参加毕业论文答辩。

得分：_____ 评阅人签字：_____

2019 年 5 月 20 日

答辩小组意见：

在答辩过程中，该同学介绍了论文的写作思路并回答了答辩小组的提问，答辩内容条理清晰，语言表达流畅，回答问题有理有据，论述的逻辑性较强，主要问题回答准确，比较有见地，论文中部分结果具有一定的深度和独创性。问题回答比较全面完整，能够运用所学知识解决实际问题。答辩效果较好。

经答辩小组表决，同意蔡同波同学通过论文答辩。

得分：_____ 答辩小组组长签字：_____

2019 年 5 月 25 日

综合评定成绩（等第）：_____

答辩委员会主任签字（盖章）：_____

2019 年 5 月 25 日