大学生论文检测系统

文本复制检测报告单(全文标明引文)

№:ADBD2022R_2022052308253320220528171105468196277837　　检测时间:2022-05-28 17:11:05

篇名: 计算机科学与技术1808班-张永远-180508010832-基于Python爬虫的关键词与观看量分析的设计与实现

作者: 张永远

指导教师:

检测机构: 郑州工商学院

提交论文IP: 218.***.***.***

文件名: 计算机科学与技术1808班-张永远-180508010832-基于Python爬虫的关键词与观看量分析的设计与实现.docx

检测系统: 大学生论文检测系统

检测类型: 大学生论文

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
源代码库
CNKI大成编客-原创作品库
机构自建比对库

时间范围: 1900-01-01至2022-05-28

可能已提前检测，检测时间：2022/5/6 10:31:05，检测结果：21.2%

检测结果

去除本人文献复制比：7.5%　　跨语言检测结果：0%

去除引用文献复制比：5.1%　　总文字复制比：7.5%

单篇最大文字复制比：1.1%　（医疗贴吧中广告的提取系统）

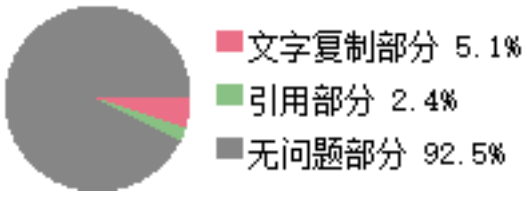
重复字数：[1216]　　总段落数：[2]

总字数：[16107]　　疑似段落数：[2]

单篇最大重复字数：[181]　　前部重合字数：[222]

疑似段落最大重合字数：[1067]　　后部重合字数：[994]

疑似段落最小重合字数：[149]



指标：☐ 疑似剽窃观点　☑ 疑似剽窃文字表述　☐ 疑似整体剽窃　☐ 过度引用

相似表格： 0　　相似公式： 没有公式　　疑似文字的图片： 0

10%(1067)　　10%(1067)　　计算机科学与技术1808班-张永远-180508010832-基于Python爬虫的关键词与观看量分析的设计与实现_第1部分（总10720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:

审阅结果:

审阅意见: 指导老师未填写审阅意见

1. 计算机科学与技术1808班-张永远-180508010832-基于Python爬虫的关键词与观看量分析的设计与实现_第1部分 总字数：10720

相似文献列表

去除本人文献复制比：10%(1067) 文字复制比：10%(1067) 疑似剽窃观点：(0)

1	医疗贴吧中广告的提取系统 张园园(导师：冯洪海) - 《河南大学硕士论文》- 2016-06-01	1.7%（181） 是否引证：是
2	基于大数据的图书阅读分析平台 王利祥 - 《大学生论文联合比对库》- 2020-06-01	1.7%（180） 是否引证：否
3	基于大数据的图书阅读分析平台 王利祥 - 《大学生论文联合比对库》- 2020-06-09	1.7%（180） 是否引证：否
4	201677I1238-王利祥-基于大数据的图书阅读分析平台的设计与实现 王利祥 - 《大学生论文联合比对库》- 2020-06-13	1.7%（180） 是否引证：否
5	现制茶饮大数据报表-凌霞-专接本毕业设计 凌霞 - 《大学生论文联合比对库》- 2021-04-23	1.5%（162） 是否引证：否
6	豆瓣电影评论文本的情感分析及主题提取研究 余洋(导师：石磊) - 《云南财经大学硕士论文》- 2018-09-10	1.4%（149） 是否引证：是
7	182_肖嘉敏_基于Python和ROSTCM6的在线教育评论数据分析系统 晓红 - 《大学生论文联合比对库》- 2019-06-09	1.1%（117） 是否引证：否
8	自组织临界特性的模型分析与开发 王永超(导师：郭辉;王森) - 《宁夏大学硕士论文》- 2019-05-01	1.0%（111） 是否引证：否
9	基于大数据技术的电信客户维系挽的分析与研究 潘娜(导师：徐江峰) - 《郑州大学硕士论文》- 2017-05-01	1.0%（109） 是否引证：否
10	项目管理平台的设计与实现 张成强;卢凤娟; - 《电脑知识与技术》- 2021-05-25	0.9%（100） 是否引证：否
11	基于APICloud的齿轮箱监控平台移动端研究 吴鲁纪;耿福震;杨林杰;李俞峰; - 《电脑知识与技术》- 2019-11-25	0.9%（98） 是否引证：否
12	基于多源数据的铁路出行乘客用户画像及方案推荐研究 吴运腾(导师：杨敏;汪林) - 《东南大学硕士论文》- 2020-05-01	0.9%（95） 是否引证：否
13	铃声音乐生成与管理系统的设计与实现 陈雨诗(导师：骆斌) - 《南京大学硕士论文》- 2019-06-30	0.8%（82） 是否引证：否
14	基于Python的文本情感分析模型设计 曾聪 - 《大学生论文联合比对库》- 2019-05-12	0.7%（78） 是否引证：否
15	基于自然语言的电影评论分析 林媛媛 - 《大学生论文联合比对库》- 2020-03-30	0.7%（78） 是否引证：否
16	3150931022-郑天方-基于python的网络爬虫设计及课程评论分析系统开发-张亚玲 郑天方 - 《大学生论文联合比对库》- 2019-06-03	0.7%（72） 是否引证：否
17	202004100250585254_刘燕_红色电影评论自动采集与分析方法研究	0.7%（72）

	刘燕 - 《大学生论文联合比对库》- 2020-04-10	是否引证：否
18	202004100250585254_刘燕_红色电影评论自动采集与分析方法研究	0.7%（72）
	刘燕 - 《大学生论文联合比对库》- 2020-04-10	是否引证：否
19	面向公众的“天地图”统计分析系统设计与实现	0.6%（65）
	张健峰;罗朝明;肖炼; - 《测绘》- 2019-06-15	是否引证：否
20	201616611219-刘坤鑫-基于ASP.NET Python党员活动专题网站的设计与实现	0.6%（65）
	刘坤鑫 - 《大学生论文联合比对库》- 2020-05-09	是否引证：否
21	基于JAVA WEB技术的城市智慧停车管理系统的设计与实现	0.6%（61）
	韩辉(导师：周艺华) - 《北京工业大学硕士论文》- 2019-05-01	是否引证：否
22	基于企业微信的黄河水情推送系统	0.6%（60）
	刘欣;廖亮;李亮亮; - 《办公自动化》- 2019-05-15	是否引证：否
23	智能播种机作业监管云平台Web前端的开发与实现	0.6%（60）
	张清博;孙宜田;孙永佳;陈刚; - 《中国农机化学报》- 2019-06-15	是否引证：否
24	基于物联网的液态饲喂远程监控系统	0.6%（60）
	王天波(导师：张军) - 《郑州大学硕士论文》- 2019-05-01	是否引证：否
25	我国对违反“爬虫协议”行为的法律规制研究	0.4%（44）
	曹阳; - 《江苏社会科学》- 2019-05-10	是否引证：是
26	5G无线网络可视化运维方式变革	0.4%（39）
	刘杨; - 《电信科学》- 2020-04-30	是否引证：否
27	基于DDS分布式系统的评估指标体系研究及应用	0.3%（30）
	沈轩(导师：沈军) - 《东南大学硕士论文》- 2018-06-18	是否引证：否

原文内容

本科毕业设计
基于Python爬虫的关键词与观看量分析的设计与实现
院部名称信息工程学院
姓名张永远
学号 180508010832
专业计算机科学与技术
届别 2022届
指导教师刘迎春
2022年4月21日

基于Python爬虫的关键词与观看量分析的设计与实现

摘要：随着现

院部名称	信息工程学院
姓名	张永远
学号	180508010832
专业	计算机科学与技术
届别	2022届
指导教师	刘迎春

在物质生活的逐步提高，人们对生活、知识和思想的探究深度也在提高。知乎是目前国内非常受欢迎的问答社区，用户通过回答别人的问题或者回复回答来分享知识资源或者通过浏览和讨论来汲取所需的资源，目前知乎在提问时会根据问题自动添加话题也可以自行增减和修改话题。本文通过分析从知乎爬虫得到的问题和对应浏览量、话题等数据进行热点分析，分析热点可以分析社会发展、了解时事，还有是可以分析舆论走向。数据分析通过可视化可以直观且清晰的发现规律，本文还通过直接分词问题来对比话题数据分析结果。经调研，热搜、热榜会带来巨大流量，这些都是后台统计数据分析的结果，商人赚的就是信息差，信息差的来源之一就是数据，分析数据可以获得更多的信息。本文主要工作有通过python和一些爬虫所需库来进行数据收集，使用SQLite存储数据有问题及问题主页网站和回答数量、关注量、浏览量。然后处理数据，对于数据内容的修改替换等使之可以完成可视化操作。最后使用flask搭建网站，使用ECharts制作图表、WorldCloud制作词云来实现可视化效果。

关键字：python；爬虫；数据分析；数据可视化

Design and Implementation of Keyword and Viewing Analysis Based on Python Crawler

Abstract: With the gradual improvement of material life now, the depth of people's exploration of life, thought and soul is also increasing. Zhihu is currently a very popular question-and-answer community in China. Users share knowledge resources by answering other people's questions or replying to answers, or absorb required resources by browsing and discussing. At present, Zhihu will automatically add topics according to the questions when asking questions. You can also add, delete, and modify topics by yourself. This article analyzes

the issues obtained from Zhihu crawlers and the corresponding data such as pageviews and topics to analyze hotspots. Analyzing hotspots can analyze social development, understand current affairs, and analyze the trend of public opinion. Data analysis can intuitively and clearly discover rules through visualization. This paper also compares the results of topic data analysis through direct word segmentation. After research, hot searches and hot lists will bring huge traffic. These are the results of background statistical data analysis. what businessmen earn is poor information, and one of the sources of poor information is data, and more information can be obtained by analyzing data. The main work of this paper is to collect data through python and some libraries required by crawlers, and use SQLite to store data. There are questions and the homepage website of the question and the number of answers, attention, and pageviews. Then process the data, modify and replace the content of the data, etc. so that the visualization operation can be completed. Finally, use flask to build a website, use ECharts to make charts, and WorldCloud to make word clouds to achieve visualization.
Keyword: python, crawler, data analysis, data visualization

目录

1 绪论

1.1 研究目的和意义

1.1.1 Python爬虫研究目的和意义

我国一直注重经济发展，但对文化产业的支持也在增加。 知乎是我国最受欢迎的知识问答社区。 用户通过共享信息资源，了解自己需要的资源。在付费内容领域，知乎月活跃付费用户数已超过250万，总内容数超过300万，年访问人次超过30亿[1]。知乎的核心作用是了解用户提出的问题匹配其他用户推荐其回答问题。 在知乎上，用户对提出的问题进行标记其所属话题，系统也会通过标记找到相关问题推荐给用户。网络爬虫的出现对于提升搜索的覆盖率和精准率有着很重要的意义[2]。

1.1.2 数据可视化研究的目的和意义

如今，无论是哪个行业，良好的数据分析都非常重要。***信息提取是指从特定的信息流中将人们感兴趣的信息过滤出来，在本文中的信息提取可以转化为文本的分类问题[3]。***数据分析只是对特定数据的准确分析。我们使用适当的统计分析技术来分析收集的大量数据，提取信息得出结论，并对数据进行更详细的检查和总结。数据分析的目的是集中、提炼和改进隐藏在大量无组织数据中的信息，以发现研究成分的内在规律。

事实上，数据分析可以帮助人们做出关于他们能做什么的决定。数据分析也是规划和收集数据、分析目的地并将其转化为信息的过程。数据分析具体可以：分类、预测分析、关联规则和推荐系统、数据缩减和降维、数据探索和可视化。

总的来说数据分析的意义就是告诉你过去发生了什么，这些现状为什么会发生，以及未来会发生什么。

各种个性化推荐让人们束缚在信息茧房中，本文可以提供一种以宏观角度分析信息热点的功能，让人们不只是被迫接受算法推荐的信息，更全面以多角度看待问题。大数据时代人们需要更强的整理信息和多维度分析信息的能力，人没有交流讨论就没有思想碰撞就不能有进步，知乎可以说部分代替了旺盛时期的贴吧、论坛。有一个自动收集整理信息并将信息进行可视化分析的工具具有非常重要的意义。

1.2 国内外文献综述

1.2.1 国内文献综述

本文选择 Python 作为此系统的开发语言。Python开发环境定义了列表、字典、元组等许多高级数据类型。这些是本文中提取数据所需的。与其他编程语言相比，您可以简化很多代码并使其更具可读性，包括作为 Python 语言特色之一的正则表达式。使编程更容易的强大功能。 ***Python 语言是可以面向过程式或面向对象的编程语言，易于学习。 Python 编程环境还提供了一种交互式编程模式，方便用户在开发应用程序时监控和查看应用程序的内容[4][5]。***

爬虫是获取、检索数据一种方式，能按照一定规则自动抓取某个网站或者万维网信息的程序；现实环境中大部分网络访问都是由爬虫造成的。本文从爬虫和数据处理分析两部分来展开。

网络爬虫又被称为网络蜘蛛，是一段可以自动抓取 Web 上信息的程序或脚本[4]。网络爬虫按照系统结构和实现技术实际应用中通常是将系统几种爬虫技术相互结合[6]。

网络爬虫，英文名为 Network Spider，故可以简单理解为网络中爬行的爬虫，它的本质是一组计算机程序[7]，要完成对网页的下载、搜索要按指定要求获取相关数据。并且这个过程无需用户干预而循环执行。通过请求页面上的 HTML 文档来识别对特定页面的访问。检查现有站点，不断在站点之间移动，自动创建文件并在内部存储它们网络数据库。当网络爬虫进入超链接时，它会搜索信息并接收其他超链接。该 URL 主要基于HTML 结构，不依赖于用户干预[3]。***首先向待爬取的网站发起请求，如果目标网站的服务器响应正常，会得到响应，然后再通过合理技术手段解析目标网页内容，最后选择爬取、***保存所需要的文本数据[8]。

数据分析是一种常用的统计方法，其主要功能是多维和描述性的。一些几何技术显示了不同数据之间的关系，并帮助您提取统计信息以更简洁地解释此数据中的重要信息。其他用于收集数据以找出谁是同质的，以便更好地理解数据。

数据分析可以处理大量数据并识别该数据中最有用的部分。近年来该领域的成功可能主要归功于制图技术的改进。这些图可以通过直接分析数据来揭示难以察觉的联系。更重要的是，它与现象分布无关，与经典的统计方法相反。数据分析是数学和计算机科学相结合的结果。数据分析是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程[9]。

通过分词技术，可以粗略地看出用户普遍看重的方面。在词云图中，词汇越大，说明该词汇在文本中出现次数越多，越能代表更多的用户[10]。

1.2.2 国外文献综述

Yu L, Li Y, Zeng Q, et al在《Summary of web crawler technology research》文章中阐述了网络爬虫在搜索引擎上扮演着重要的角色，尤其是在提取网页时。Web爬虫最重要的作用是在Internet的大数据中爬行，查找有效的信息，并将所需的信息数据存储到本地数据库中，是穿越超链接和索引的计算机程序[11]。

1.3 研究的主要内容和方法

熟悉网页页面结构，正则表达式、bs4网页解析提取页面元素，运用爬虫库requests框架、数据库增删改查，能根据需求，处理常见的反爬，抓取数据。

用Flask进行网页搭建、ECharts对数据进行可视化处理，WordCloud进行绘制词语图，了解聚类分析等分析方法。

熟悉使用HTML web开发。能对常见数据载体格式进行数据的解析。

1.4 系统技术介绍

1.4.1 Python爬虫介绍

首先分析网页，网站通常由三部分组成：超文本标记语言（HTML）、层叠样式表（CSS）和活动脚本（JScript）。

获取待***爬取网页的链接，我们还需要对每一个 url 做去重检测。***

解析 html 网页源代码获得文本信息，长时间或持续的爬取一个网站可能会触发网站的反爬虫机制，因此需要对爬虫技术做伪装，采用设置（user-agent），***可将登录的 cookie 信息通过 user-agent 一同存在请求中发给浏览器。***

HTML是整个网站的结构，相当于整个网站的框架。 “<”和“>”符号是HTML标签，标签是成对的。外观在 CSS 中定义。JScript 代表一个动作。 在 JScript 中可以找到交互式内容和各种特殊效果，它描述了网站的不同功能。

当使用人体作为类比时，HTML 是人体骨架，它定义了嘴、眼睛、耳朵等的位置。 CSS是一个人的外貌信息，比如嘴巴长什么样，眼睛是双眼皮还是单眼皮，眼睛是大是小，皮肤是黑是白。 JScript 代表人类技能，例如跳舞、唱歌和演奏乐器。

要抓取网页，您首先需要分析您的网页设计。许多网站现在使用一种称为 Ajax（异步加载）的技术。 这意味着当你打开一个网站时，你会首先看到上面的一些，其余的会慢慢加载。 因此，您可以查看许多可以滑动的网页，并且某些网页可以在您导航时缓慢加载大量信息。 这种页面的优点是页面加载速度非常快（因为您不必一次加载所有内容）。但是，这项技术不适合爬虫，此时需要花一点力气。知乎不滑动只出现五个回答，由于此次爬虫不爬取评论内容，所以这个问题不需要解决。

爬虫就是把网页中需要的数据通过指定标签和属性的方式提取出来，然后通过正则表达更准确的提取出标签内的所需数据。

1.4.2 Flask框架介绍

Flask相对于Django而言是轻量级的Web框架。和Django不同，Flask轻巧、简洁，通过定制第三方扩展来实现具体功能。可定制性，通过扩展增加其功能，这是Flask最重要的特点。Flask的两个主要核心应用是Werkzeug和模板引擎Jinja。能快速搭建网站，能接受用户传递的整数或字符串，也能把数据传递给网页。

1.4.3 Echarts图表介绍

Enterprise Charts 简称为 ECharts，这是一个使用JavaScript 实现的开源可视化库[12]，涵盖各行业图表。

一个纯JavaScript图表库。ECharts，缩写来自Enterprise Charts，商业级数据图表，一个纯Javascript的图表库，可以流畅的运行在PC和移动设备上，兼容当前绝大部分浏览器。具有各种各样的图表，丰富的动态效果。在官网有详细的教程和非常多的示例，能在官网实时渲染修改代码。

1.4.4 WordCloud词云介绍

词云是一种可视化文本展示方式，它是由文本中提取的数据组成彩色图像。词云图的核心价值在于以关键词的大小的可视化表达来传达大量文本数据背后的有价值的信息。Wordcloud是一个生成词云的Python包，可以以词语为单位直观和艺术的展示文本。

1.5 系统开发的平台和运行环境

1.5.1 系统开发环境平台

运行环境是Windows10，开发本系统是在PyCharm下进行，PyCharm是主要用于python开发的集成环境。

1.5.2 运行所需包

在爬虫阶段有：bs4用于网页解析，获取数据、re正则表达式，进行文字匹配、urllib制定URL，获取网页数据和获取网页错误信息、sqlite3进行SQLLite数据库操作、time 在本系统用于设置等待时间。

在数据可视化部分在Flask里导入有flask包、sqlite3，在网页中导入echarts.min.js文件，在WordCloud部分导入有wordcloud，中文分词工具jieba，matplotlib：一个综合库，主要用于绘图，numpy矩阵运算，sqlite3。

2 爬虫数据可视化系统分析

2.1 可行性分析

2.1.1 经济可行性分析

在生产和生活中获取更多的信息有非常大的帮助，而把信息以一种简单且直观的方式展现更是如虎添翼。

人工时间成本升高，为减轻重复性劳动负担，爬虫数据量之大不能使用人工方式代替。开发、维护此系统的成本较少，因此在经济上是可行的。

2.1.2 技术可行性分析

本系统采用Python语言使用爬虫技术，可视化分析采用flask框架、Echarts制图技术和Word Cloud词云技术。后台数据库采用轻量化关系数据库管理系统SQLite，本系统在技术方面是可行。

2.1.3 操作可行性分析

本系统用户的操作只是在网页来观看信息展示，展示信息获取处理是在后台处理，用户不能干预。网页操作简单，风格、效果展示简介易懂，因此本系统在用户操作上也是可行的。

2.2 需求分析

2.2.1 功能需求分析

本系统主要是实现爬虫及数据可视化系统

图2-1 总体思路概括

图2-2 可视化概述图

2.2.2 性能需求分析

1）爬虫对于性能要求不高。

2）本文要爬虫的内容不是很巨大，对于数据库选择SQLite足够满足使用。

3）Flask是轻量型WEB框架，Echarts是百度开源框架，能够在绝大多数计算机上流畅运行，WordCloud如果设置生成图片的dpi高些会导致生成速度不高，不过可以在后端提前生成好，对用户体验不影响。

2.2.3 系统的流程

分析提取网页链接，分析网页具体代码所对应要爬取信息编写对应代码，保存到数据库，运行可视化系统即可打开网页查看四个主页面和若干子页面。

3 系统总体设计

3.1 爬虫功能设计

知乎问题爬虫的具体项目有：问题、话题、浏览量、关注量、回答数量。如图3-1所示。

图3-1 爬虫的具体项目结构图

3.2 数据库设计

本系统使用了两个数据库，一个是处理话题精华问题的链接和问题题目，另一个是爬虫的数据。由于本项目对于数据库要求不高，每个数据库有一张表。

图3-2 数据流图

3.3 数据可视化设计

3.3.1 flask框架设计

在主页下有数据展示、图表、词云和团队链接，在数据展示和图表下有具体话题的链接，词云页面是直接显示所有内容。

图3-3 可视化系统Flask框架

3.3.2 Echarts图表设计

设计关于回答数量和浏览量的散点图，回答量和浏览量大部分都是数量少，因此点局部集中要能够对局部放大缩小。

3.3.3 Wordcloud词云设计

由于知乎话题结构就是像一个有根树，所以选择背景图为一颗树的图片。

4 系统的详细设计与实现

4.1 系统前台界面的实现

系统的界面设计主要包含了系统的主页以及各功能界面的设计与实现。

4.1.1 主页功能界面的实现

运行flask后点击下方出现网站即可在浏览器打开主界面，系统界面简洁明了，操作简单，设计首先在所有网页顶部均有通往各个主页面的导航条，在主页主要是展示其他主页面的简单数据和链接。

图4-1 主页界面

4.1.2 数据展示页面的实现

进入首先是选择具体话题的页面，选择后是以列表的形式显示爬取过的信息。

图4-2 数据展示选择部分

图4-3 数据展示部分子页面

4.1.3 Echarts图表界面的实现

1)与数据展示页面类似先进入选择具体问题页面，多了一个全部数据的选项。选择后才可进入查看图表。

图4-4 Echarts图表展示

2) 图表页面可以用鼠标滚轮和拖动或者下方和右方拖动条对图表进行局部缩放查看。在产业话题中回答量和浏览量的正相关性是整体最明显的，同时也有回答量很多但浏览量不多的问题，可见对于单个问题回答量和浏览量没有绝对相关性。

图4-5 Echarts图表子页面产业话题

图4-6 Echarts图表子页面产业话题放大

3) 全部数据就是把以上七个话题数据库的叠加。

图4-7 Echarts图表子页面全部数据

图4-8 Echarts图表子页面全部数据放大

在三千浏览量以下虽然回答量和浏览量没有绝对关系，但随着浏览量的增加回答数量比例也在相应增加。

大部分问题都是只有很小的阅读量和浏览量，最大浏览量达七千万之多，相应的回答数量也有四千七：什么叫降维打击？三体（书籍），「形而上」话题,降维打击。

4.1.4 WordCloud词云界面的实现

词云页面采取了直接全部显示后台生成的所有话题的词云。

图4-9 根话题词云

图4-10 学科话题词云

学科话题下问题的关于大学、考研比例很高，并且对比其他话题的问题量此话题的问题量也很高，由此可推断知乎用户比例较大的学历和年龄结构是大学及以上。

4.2 后台功能的实现

4.2.1 爬虫主要功能的实现

判断你爬取的数据是不是静态数据，查看你网页的源码。了解基本的HTML格式，使用浏览器的F12功能键打开开发者工具定位到要爬取的内容相关语句，并记录下来在接下来的爬虫过程中要用到。由于大部分网站都有反扒措施，我们也需要一些手段来应对，包括模拟浏览器头部信息，向服务器发送消息，表示告诉服务去，我们是什么类型的机器、浏览器（本质上是告诉服务器，我们可以接受什么类型的文件内容）。具体方法就是在要访问的页面按F12，然后点击Network→刷新页面→马上点击开发者工具的红色点stop，鼠标定位到时间轴最左面→点击name下面一行然后出现Headers就是浏览器发送给服务器的头部内容，在headers最下面"User-Agent"是浏览器标识，我们要加入urllib里模拟正常浏览器访问，否则是直接发送给浏览器python版本号。“爬虫协议”*另一个是Allow或Disallow值，用于设置特定搜索引擎所能访问或禁止访问的具体内容[13]。*

图4-11 headers头部信息

还有需要导入time模块添加等待时间来防止被反爬。经过尝试设置3秒最合适。

第二步：发送网络请求

导入发送请求的模块requests，打开PyCharm，点击文件菜单→设置→项目→Python解释器，再点击右侧出现的加号，在弹出页面输入要添加的模块，最后点击安装包等待提示安装成功即可。

图4-12 安装第三方包requests

可以按以上步骤安装bs4用于网页解析，获取数据，re包用于正则表达式，进行文字匹配，urllib包用于制定URL，获取网

页数据，xlwt包是进行excel操作的，使用SQLLite数据库操作用sqlite3包。

爬虫的基本原理

请求网站的过程分为两个部分：

请求（Request）：向用户显示的所有网页都必须向服务器执行这一步。

响应（response）：服务器收到用户的请求后，验证其有效性无误后将响应的内容发送给用户。用户收到服务器响应的内容并在网站上给我们显示熟悉的内容。

有两种方法可以请求网站。

GET：最常用的方法，大多数网站使用此本文的主要工作包括以下几个方方法，响应率高。

POST：相比GET方法，它有更多的以表单的形式上传变量的功能，所以你不仅可以请求信息，还可以编辑它，一般用于用户登录。

所有在源码中的数据请求方式都是GET， POST 的请求获取数据的方式不同于 GET，POST 请求数据必须构建请求头才可以。

因此，在创建爬虫之前，您必须首先决定将请求发送到何处以及如何发送。

用Beautiful Soup分析网站

您已经可以从requests库中获取网页的源代码。 下一步是从源代码中查找并提取数据。 Beautiful Soup 是一个 Python 库，其主要功能是从网页中检索数据。 Beautiful Soup已经转移到 bs4 库。 这意味着您需要在导入 BeautifulSoup 之前设置 bs4 库。再使用正则表达抽离出所需信息，最后使用sqllite保存信息。由于时间关系本次只获取根话题即其六个子话题里的问题。

核心代码：

```
# 正则表达
# 问题自带话题
rehuati = re.compile('<meta content="(.*?)".*"/>')
# 题目
rewenti = re.compile('title">(.*)</h1>') # <h1 class="QuestionHeader-title">马斯克称「人类如果不多生孩子，文明将会崩溃」，如何看待其言论？</h1>
…
…
if cuowu != 404 and cuowu != 410: # 判断页面是否为404
 # 问题
 for popover in soup.select("h1[class='QuestionHeader-title']"):
 data = [] # 保存一个问题的所有信息
 # 先保存链接id
 data.append(str(urlnumb))
 wenti = re.findall(rewenti, str(popover))[0].replace('"', "'") # findall返回列表猜测：文字添加数据库是需要加引号，与问题中的引号干扰
 # print(wenti) # sqlite3.OperationalError: near "很人渣": syntax error
 data.append(wenti)
 for popover in soup.select("meta[name='keywords']"):
 # print(popover) # 测试
 huati = re.findall(rehuati, str(popover))[0] # re.findall()第二个参数需要字符串类型
 # print(huati)
 data.append(huati)
 # print(data)
…
…
if len(data) == 6:
 for index in range(len(data)):
 if index == 1 or index == 2: # type(data(index)) != type(float)
 data[index] = '"' + data[index] + '"' # 非数字插入数据库时需要有双引号或单引号
 sql = '''
 insert into zhihu(
 link,wenti,huati,guanzhu,liulan,huida)
 values(%s)
 ''' % ",".join(data) # 把列表元素用逗号分割，组成字符串。
```

指　　标
疑似剽窃文字表述

1. Design and Implementation of Keyword and Viewing Analysis Based on

2. 数据的准确分析。我们使用适当的统计分析技术来分析收集的大量数据，提取信息得出结论，并对数据进行更详细的检查和总结。数据分析

3. 框架。和Django不同，Flask轻巧、简洁，通过定制第三方扩展来实现具体功能。

4. ECharts，缩写来自Enterprise Charts，商业级数据图表，一个纯Javascript的图表库，可以流畅的运行在PC和移动设备上，兼容当前绝大部分浏览器。具有各种各样的图表，

5. 系统的详细设计与实现
   4.1 系统前台界面的实现
   系统的界面设计主要包含了系统的

6. 源代码。 下一步是从源代码中查找并提取数据。 Beautiful Soup 是一个 Python 库，其主要功能是从网页中检索数据。

%是把后面填到%s位置.join链接
爬虫效果展示：
图4-13 爬取过程
图4-14 根话题数据展示
图4-15 学科话题数据展示
4.2.2 爬虫编写遇到的问题及对策
在把数据保存到数据库时报以下错误
图4-16 保存数据出错提示
经过仔细观察后发现，数据库在存储数字时不需要加引号，而在回答量比较多的情况下知乎数据在第三位数前有逗号，如果不把逗号替换数据库就会以为传递了7个数据
图4-17 保存数据错误原因分析
解决办法：把回答量的逗号替换成空字符。
一共出现两次这种错误，在判断是否页面无内容的两个分支上各出现一次，原因就是未把链接序号转为字符串。
图4-18 格式错误报错1
出现以下错误：
图4-19 格式错误报错2

判断列表长度正常后执行：

```
data[index] = '"' + data[index] + '"' # 非数字插入数据库时需要有双引号或单引号
 sql = '''
 insert into zhihu(
 link,wenti,huati,guanzhu,liulan,huida)
 values(%s)
 ''' % ",".join(data) # 把列表元素用逗号分割，组成字符串。%是把后面填到%s位置.join链接
 # connection.execute("INSERT INTO UTILISATEURS (FULLNAME, EMAIL, CIN, ADDRESS, PHONE, RIB) VALUES
(?,?,?,?,?,?) ", (str(fullname), str(email), str(cin), str(address), str(phonenumber), str(ribnumber)))
else:
 sql = '''
 insert into zhihu(link,wenti,huati,guanzhu,liulan,huida)
 values(%s)
 ''' % data
```

解决办法：

```
else:
 sql = '''
 insert into zhihu(link,wenti,huati,guanzhu,liulan,huida)
 values(%s,NULL,NULL,NULL,NULL,NULL)
 ''' % data[0]
```

如果爬取过程由于某些原因中断，由于表已经建立，重新执行会报错

用异常机制：

```
try:
 sql = '''
 create table zhihu
 (
 id integer primary key autoincrement,
 link int,
 wenti varchar,
 huati varchar,
 guanzhu int,
 liulan int,
 huida int
 )
 ''' # 创建数据表，autoincrement自增长
…
…
 except:
 print('已经存在表')
```

在运行到518403259时出现以下错误

打开此链接的页面：

图4-20 网站提示链接内容被删除

这种情况属于没有考虑到，解决办法：在if和else中间加上elif语句，同时修改if语句

```
if cuowu != 404 and cuowu != 410: # 判断页面是否为404
elif cuowu == 410:
 print("410该内容以删除")
 data = [str(urlnumb)]
else:
 print("404链接无内容")
 data = [str(urlnumb)]
```

由于未考虑冷门问题没有人回答的现象出现以下错误

图4-21 没有用户回答

解决办法：在数据库的if-else中间加上elif

```
elif len(data) == 5: # 没有回答的情况
 for index in range(len(data)):
 if index == 1 or index == 2: # type(data(index)) != type(float)
 data[index] = '"' + data[index] + '"' # 非数字插入数据库时需要有双引号或单引号
 sql = '''
 insert into zhihu(
 link,wenti,huati,guanzhu,liulan,huida)
 values(%s,NULL)
```

''' % ",".join(data)

如果直接按照链接递增的方式来获取数据：一共爬取两万五千多条数据，只有两条是有内容的

图4-22 按链接递增爬取结果1

图4-23 按链接递增爬取结果2

不具有分析效果。

问题编号有九位和八位，总数量能达到十亿多，就算知乎问题数量有千万之多，爬取有用链接编号概率不高，这也是一种反爬虫的手段。虽然知乎问题编号大致上是根据时间排序，但不是连续的，所以只能换个思路。

图4-24 根话题主页

知乎每个话题下都有一些知乎自己筛选的一千个精华问题，其中包含文章和问题，由于问题没有话题标签，筛选掉文章大概只能获取到总数量一半的问题大概五百，不同话题差别很大。

虽然获取的数据量不大，但总归是一种可行的方法。

在写入数据库时有以下问题

图4-25 存入数据库报错

猜测：文字添加数据库是需要加引号，与问题中的引号干扰

wenti = re.findall(rewenti, str(popover))[0].replace('"', "'") # findall返回列表猜测：文字添加数据库是需要加引号，与问题中的引号干扰

使用replace把双引号替换成单引号解决问题

4.2.3 Flask主要功能实现

使用flask框架搭建网站，在flask内编写处理数据传递到网页，编写需要展示的网页，在网页内使用Echarts制作图表

Fladk核心代码：

```
# echarts图表
@app.route('/score/<name>') # 通过访问路径，获取用户的字符串参数
def scores(name):
 datalist = []
 names = name + '.db'
 con = sqlite3.connect(names)
 cur = con.cursor()
 sql = "select * from zhihu"
 data = cur.execute(sql)
 for item in data:
 datalist.append(item)
 cur.close()
 con.close()
 name = name.replace("已获取知乎", "")
 huidaliulan = []
 for list in datalist:
 lis = []
 lis.append(int(str(list[6]).replace("None", "0"))) # 部分数据回答量为None 不适用数值转换会报 'int' object has no attribute 'replace'
 lis.append(list[5])
 huidaliulan.append(lis)
 print(huidaliulan)
 return render_template("tubiao.html", data=huidaliulan, name=name)
```

4.2.4 Echarts图表功能及实现

核心代码：

```
series: [
 {
 type: 'scatter',
 data: {{data}},
 dimensions: ['x', 'y'],
 symbolSize: 8,
 itemStyle: {
 opacity: 0.8
 },
…
…
```

4.2.5 WordCloud词云功能及实现

分词作为一种基本技术，目前由很多较成熟的标记化工具可以使用，例如Stanford Tokenizer、OpenNLPTokenizer、jieba以及哈工大LTP等[14]。本文使用jieba。

停用词来源于网络上整理好的停用词表，。删除了停用词的分词词表相对来说更为简洁[15]。

Wordcloud核心代码：

…
# 绘制图片
fig = plt.figure(1)
plt.imshow(wc)
plt.axis('off') # 是否显示坐标轴
# plt.show() # 显示生成的词云图片
# 输出词云图片到文件
plt.savefig(r'./static/assets/img/问题生活、艺术、文化与活动话题.jpg', dpi=800)
4.2.6 数据可视化编写过程存在的问题及对策
在编写展示数据网页时出现
图4-26 网页效果错误
分析原因由于访问路径改变导致图片和css路径错误
图4-27 分析效果错误原因
解决办法：批量替换
图4-28 解决效果错误
加个/指向当前站点根目录的 static文件夹中。
图4-29 解决效果错误网页示例
4.3 数据库的连接
4.3.1 爬虫部分数据库连接
预处理系统会把知乎精华问题网页的数据处理提取出链接和问题题目，放入预处理数据库，然后爬虫主程序会把预处理数据库中的链接依次读取进行爬虫操作，爬取的数据存入新的爬虫数据库。
4.3.2 可视化部分与数据库连接
可视化系统只涉及对数据库的读取，Echats对数据库中的回答量和阅读量读取，词云是对数据库中的问题或话题读取，经对比提取那种显示结果相差不大。
5 系统测试
5.1 测试系统的目的
系统测试是运行程序以发现错误的过程，而成功的测试是发现以前未被发现的错误的过程。
5.2 测试方法
对爬虫数据可视化系统进行测试。对与爬虫运行测试，flask运行测试，词云生成图片测试，和Echarts图表显示测试。由于编写过程就是在不断的报错中修改的，所以最后的测试未出现错误。
表5-1 爬虫数据可视化系统的测试用例

序号	测试点	操作步骤	期望结果	实际结果
1	爬虫运行测试	运行爬虫程序	运行成功，不报错，显示提示信息。	
2	flask运行测试	运行flask程序	运行成功，在浏览器打开不报错。	
3	词云生成图片测试	运行词云程序	运行成功，无报错，生成词云图。	
4	Echarts图表显示测试	打开图表网页查看	打开成功，正常显示图表。	

序号测试点操作步骤期望结果实际结果
1 爬虫运行测试运行爬虫程序运行成功，不报错，显示提示信息。
2 flask运行测试运行flask程序运行成功，在浏览器打开不报错。
3 词云生成图片测试运行词云程序
运行成功，无报错，生成词云图。
4 Echarts图表显示测试打开图表网页查看打开成功，正常显示图表。
结论
爬虫是获取、检索数据一种方式，能按照一定规则自动抓取某个网站或者万维网信息的程序；现实环境中大部分网络访问都是由爬虫造成的。本文从爬虫和数据处理分析两部分来展开。数据分析是从样本到总体意义上的推断、是精简过的，通过对原始数据的简单而直接的提取，虽然我们看不出这些迹象，但这也只是部分，而不是整体。数据分析的某些部分，超出了它的语言学范围，在某种意义上，他们指导我们观察或分析有价值的方向。数据分析是一个比推理的过程更大的环节。
本文设计并实现了Python网络爬虫，完成了对各指定知乎话题精华问题的爬取，提取了问题的题目、话题、被浏览量、回答量等数据。使用此数据做出了观察浏览量与回答数量的散点图，并根据数据的话题做出了词云图。从知乎网站爬取了大量的数据。使用BeautifulSoup、正则表达、SQL处理、保存数据。基于多种第三方工具实现数据的可视化处理，锻炼了自己的专业技能，提高了学习能力，通过分析数据提升了数据敏感度，提高了多维度看待问题的能力。
后续工作的展望
由于时间关系爬取话题数量较少，后续爬取所有话题数据量高后要把这些连接起来做一个完全自动化的爬虫和数据展示分析，还有要做数据展示的排序，界面的美化，图表还要增加一个话题与关注者或者回答量的关系更精准的反应话题热度。
参考文献
[1]. 中国经济网. 2021新知青年大会开幕知乎将继续加大对创作者支持[J].
[2]. 潘娜. 基于大数据技术的电信客户维系挽留的分析与研究[D]. 河南:郑州大学,2017.
[3]. 张园园. 医疗贴吧中广告的提取系统[D]. 2016. DOI:10.7666/d.D01052856.
[4]. Y. Daniel Liang. Python 语言程序设计[M]. 成都：机械工业出版社，2013：30-33.

[5]. Jennifer Campbell. 利用 python 进行数据分析[M]. 成都：机械工业出版社，2012：18-21.
[6]. 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.
[7]. 封俊.基于 Hadoop 的分布式搜索引擎研究与实现[D].[硕士学位论文].太原:太原理工学,2010
[8]. 余洋. 豆瓣电影评论文本的情感分析及主题提取研究[D]. 云南:云南财经大学,2018.
[9]. 陶皖主编. 云计算与大数据[M].西安电子科技大学出版社，2017.01：第44页.
[10]. 杨磊磊. 大数据视角下非结构化文本数据的顾客满意度研究[D]. 2017. DOI:10.7666/d.D01198010.
[11]. Yu L, Li Y, Zeng Q, et al. Summary of web crawler technology research[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1449(1): 012036.
[12]. 崔蓬 .ECharts 在数据可视化中的应用 [J]. 软件工程 ,2019（6）:42-46.
[13]. 曹阳.我国对违反“爬虫协议”行为的法律规制研究[J].江苏社会科学,2019(03):159-167.DOI:10.13858/j.cnki.cn32-1312/c.2019.03.021.
[14]. 阮泽楠. 音乐社交平台用户情绪特征研究[D].浙江理工大学,2019.
[15]. 张瑾. 知乎“抄袭”话题评论的情感分析[D].云南财经大学,2018.

致谢

四年的求学生涯，我走得虽然有点辛苦但是收获颇丰。在老师，朋友的全力支持下，也在不断地学习以及进步，在此论文即将付梓之际，我也在反思自身，在这大学四年的光阴里，是否有虚度，是否有努力，是否有进步。

我的论文能够顺利进行离不开我的论文老师对我的指导。对此我非常感谢指导老师能够在繁忙的教学工作中抽出时间来对我的论文进行审查和修改。以及所有教过我的老师们，你们严格而细致，一丝不苟的做法是在我今后的工作或者学习中的榜样；您们循循善诱的教导和不拘一格的思路，能够让我无论是学习或是生活中都受益匪浅。

同时，也非常感谢在我困难时给予我帮助与陪伴的小伙伴们，你们的帮助与鼓励成为我坚持下来的力量。对此，最应该感谢的是在我身后默默支持我的父母，无以报答父母的养育之恩，只希望你们永远健康快乐便是我最大的心愿！在这论文即将完成之际，我的心情十分激动，我的导师以及我的朋友成为我在我开始进入课题到最终论文得以完成对我给予了莫大关注，在这里，也同样请接受我真诚的感谢！