普通高等院校
网络与新媒体专业系列教材

Basics
of Big Data
Analysis

大数据分析基础

罗茜 编著

清華大學出版社
北京

内 容 简 介

本书是传媒专业的研究方法类课程教材，具体介绍了大数据分析的基本原理、主要方法、技术操作、研究应用以及常见工具，系统讲解了从数据收集到分析、挖掘的全套研究流程。本教材聚焦于传播学科与大数据科学的交叉领域，旨在拓宽相关专业学生的学术视野，提供更加丰富和精确的研究工具。

本教材共分 10 章，分别介绍了在计算传播学研究中常用的大数据方法。第 1 章主要介绍了大数据的获取方法，第 2 章至第 10 章分别介绍了文本分析、情感分析、聚类分析、主题模型、机器学习、自动文本分析、社会网络分析、语义网络分析、虚拟仿真等具体的大数据分析方法。

本教材将研究方法与研究案例相结合，内容丰富，难易适中，注重系统性、科学性、实用性、时代性和引导性，既可作为传媒专业及交叉学科教师、研究生、本科生、大中专院校学生的教学、实践与研究资料，又可作为传媒从业者、市场营销人员和社会科学研究者等读者的参考读物。

本书提供课件，请扫描封底二维码获取。

图书在版编目 (CIP) 数据

大数据分析基础 / 罗茜编著 . -- 北京 : 清华大学出版社 , 2024. 10. -- (普通高等院校网络与新媒体专业系列教材). -- ISBN 978-7-302-67413-9

Ⅰ. TP274

中国国家版本馆 CIP 数据核字第 2024N55W53 号

责任编辑：施　猛　王　欢
封面设计：常雪影
版式设计：方加青
责任校对：马遥遥
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：https://www.tup.com.cn，https://www.wqxuetang.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编：**100084
社 总 机：010-83470000　　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市龙大印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张：**21.25　　**字　　数：**441 千字
版　　次：2024 年 10 月第 1 版　　**印　　次：**2024 年 10 月第 1 次印刷
定　　价：69.00 元

产品编号：099467-01

普通高等院校网络与新媒体专业系列教材
编　委　会

序　　言

当今世界，媒介融合趋势日益凸显，移动互联网的快速普及和智能媒体技术的高速迭代，特别是生成式人工智能(artificial intelligence generated content，AIGC)推动着传媒行业快速发展，传媒格局正在发生深刻的变革，催生了新的媒体产业形态和职业需求。面对这一高速腾飞的时代，传统的人文学科与新兴的技术领域在“新文科”的框架下实现了跨界融合，使得面向智能传播时代的网络与新媒体专业人才尤为稀缺，特别是在“新文科”建设和“人工智能+传媒”的教育背景下，数字智能技术的飞速发展使得社会对网络与新媒体专业人才的需求呈几何级增长。

教育部于2012年在本科专业目录中增设了网络与新媒体专业，并从2013年开始每年批准30余所高校设立专业，招生人数和市场需求在急速增长，但网络与新媒体专业的教材建设却相对滞后，教材市场面临巨大的市场需求和严重的供应短缺，亟需体系完备的专业教材。2022年春天，受清华大学出版社的热情邀约，苏州大学传媒学院联合中国科学技术大学、西安交通大学、中国人民大学、北京师范大学等多所网络与新媒体专业实力雄厚的兄弟院校，遴选各校教学经验丰富的一线学者共同组成系列教材编写团队，旨在开发一套系统、全面、实用的教材，为全国高等院校网络与新媒体专业人才培养提供系统化的教学范本和完善的知识体系。

苏州大学于2014年经教育部批准设立网络与新媒体专业，是设置网络与新媒体专业较早的高校。自网络与新媒体专业设立至今，苏州大学持续优化本科生培养方案和课程体系，已经培养了多届优秀的网络与新媒体专业毕业生。

截至2024年初，“普通高等院校网络与新媒体专业系列教材”已签约确认列选22本教材。本系列教材主要分为三个模块，包括教育部网络与新媒体专业建设指南中的绝大多数课程，全面介绍了网络与新媒体领域的核心理论、数字技术和媒体技能。模块一是专业理论课程群，包括新媒体导论、融合新闻学、网络传播学概论、网络舆情概论、传播心理学等课程，这一模块将帮助学生建立起对网络与新媒体专业的基本认知，了解新媒体与传播、社会、心理等领域的关系。模块二是数字技术课程群，包括

数据可视化、大数据分析基础、虚拟现实技术及应用、数字影像非线性编辑等课程，这一模块将帮助学生掌握必备的数据挖掘、数据处理分析以及可视化实现与制作的技术。模块三是媒体技能课程群，包括网络直播艺术、新媒体广告、新媒体产品设计、微电影剧本创作、短视频策划实务等课程，这一模块着重培养学生在新媒体环境下的媒介内容创作能力。

本系列教材凝聚了众多网络与新媒体领域专家学者的智慧与心血，注重理论与实践相结合、教育与应用并重、系统知识与课后习题相呼应，是兼具前瞻性、系统性、知识性和实操性的教学范本。同时，我们充分借鉴了国内外网络与新媒体专业教学实践的先进经验，确保内容的时效性。作为一套面向未来的系列教材，本系列教材不仅注重向学生传授专业知识，更注重培养学生的创新思维和专业实践能力。我们深切希望，通过对本系列教材的学习，学生能够深入理解网络与新媒体的本质与发展规律，熟练掌握相关技术与工具，具备扎实的专业素养和专业技能，在未来的媒体岗位工作中能熟练运用专业技能，提升创新能力，为社会做出贡献。

最后，感谢所有为本系列教材付出辛勤劳动和智慧的专家学者，感谢清华大学出版社的大力支持。希望本系列教材能够为广大传媒学子的学习与成长提供有力的支持，日后能成为普通高等院校网络与新媒体专业的重要教学参考资料，为培养中国高素质网络与新媒体专业人才贡献一份绵薄之力！

2024年5月10日于苏州

前　言

在智能传播时代，人们的日常生活深深嵌入网络之中，几乎所有的社会互动和个体活动都在某种程度上留下了数字踪迹。在线上，从信息搜索、购物消费到媒体互动；在线下，从外出旅行、学习培训到娱乐活动，产生的所有信息都以数据形式被记录并存储，从而形成庞大而复杂的数据生态系统，构成全面而多元的基础信息库。这些行为数据具有丰富性和实时性，正在改变社会科学研究范式，并催生了“计算社会科学”这一研究领域。

大数据时代的崛起，为社会科学研究者提供了新的研究方法和研究途径，研究者能够更加深入地挖掘庞大而又复杂的信息流，揭示其中蕴含的规律，预判未来的发展趋势。从信息采集、处理到知识提取，大数据技术以前所未有的广度、深度以及收集和分析数据的能力，改变了我们获取信息的方式，更为我们理解社会现象提供了全新的认知途径。大数据时代给社会科学研究者带来了前所未有的机遇，提高了理论创新的可能性，但也给社会科学研究者带来了巨大的挑战。数据的庞大、多样和实时性既要求社会科学研究者掌握先进的数据处理和分析技能，又要求社会科学研究者发展新的理论框架和方法论，从而更深刻、全面地理解和解释这些数据。这就对传媒专业的学生培养提出了新的要求。在海量信息时代，传统的研究方法已经不能满足高效处理庞大数据集的需求，传媒专业的学生需要掌握并能运用先进的计算方法，从庞杂的媒体数据中提炼出有价值的信息，以便开展新闻报道、深度分析、广告投放、公关策略优化等实务工作。此外，传媒专业的学生还应能在理论层面理解大数据改变信息生产和传播的过程，从而推动传媒研究趋向于数据驱动，为学科的深度发展创造新的契机。

本书是传媒专业的研究方法类课程教材，一方面侧重技术细节，致力于用浅显和简单的语言介绍复杂的计算方法；另一方面侧重计算方法在新闻传媒研究中的实际应用，以便于读者学习如何将计算方法应用于社会实践和理论创新。本书以大数据分析技术为线索，每章介绍一种方法，包括大数据的获取、文本分析、情感分析、聚类分析、主

题模型、机器学习、自动文本分析、社会网络分析、语义网络分析、虚拟仿真等。除了“第1章 大数据的获取”以外，其他章的编写思路都是先介绍技术原理，再通过数据案例介绍技术操作流程，最后选取若干篇研究案例，介绍相关技术原理在新闻传播学研究中的主要应用。

本教材的知识框架和知识内容基于本人的课堂教案，感谢我的学生蔡文怡、赖咏晴校对教材内容并进行格式调整，感谢我的学生文湘婧、王皎(第1章)、陈辰(第2章)、杨雅坤(第3章)、王昱瑾(第4章)、蔡文怡(第5章)、赖咏晴(第6章)、董晨曦和吴洋洋(第7章)、凌觐如(第8章)、耿珂欣(第9章)、张金(第10章)帮助收集资料并凝聚成初稿。

此次书稿得以成书出版，还得益于苏州大学传媒学院给予的大力支持和我的导师清华大学沈阳教授提供的宝贵指导意见，在此表示衷心的感谢。

限于作者水平，书中不足之处在所难免，敬请读者批评指正。反馈邮箱：shim@tup.tsinghua.edu.cn。

罗茜

2023年12月

目　录

第1章　大数据的获取

人类记录社会和自然现象始于远古时代的结绳记事。随着科学技术的发展以及社会的进步，数据的数量持续增长，特别是自18世纪60年代工业革命以来，计算机和互联网的出现催生了存储、分析、查询数据技术，为人们高效处理结构化数据提供了更多可能性。如今，全球数据量呈现爆炸式增长，大数据时代悄然而至。

维基百科将大数据定义为利用常用软件工具来获取、管理和处理数据所耗时间超过可容忍时间的数据集[1]。国际咨询公司IDC定义了大数据的4V特征[2]，即数据规模(volume)大、数据种类(variety)多、数据要求处理速度(velocity)快、数据价值(value)密度低。大数据时代的到来，意味着数据将连续、动态地更新，当前的数据量已超出传统方法和技术手段所能处理的范围。因此，我们需要不断学习新的数据获取方法，以便从庞大的数据集中挖掘出有价值的指标和信息。掌握了大数据的获取方法是在大数据时代中快速提取数据并实现数据价值的关键。

1.1　大数据获取方式

获取大数据是开展大数据分析的前提，那么，我们应该如何获取大数据呢？通常情况下我们可以采取两种方式，分别是网络公开数据获取以及网络爬虫获取。

1.1.1　网络公开数据获取

获取网络公开数据的方法有两种：一是通过搜索引擎查询数据，当我们无法预知数据来源时，使用搜索引擎是最直接的方式；二是通过公开的数据源和数据分享站点获取数据，这种方式通常用于正式研究或获取官方数据。公开数据通常由政府、企业或其他组织提供并向社会公众开放，这些数据可以自由获取和使用。下面介绍一些常用的数据源。

1. 政府公开数据

1) 国内

(1) 中国国家数据 https://data.stats.gov.cn/easyquery.htm?cn=A01

如图1-1所示，中国国家数据作为国内权威的数据平台，集中了各个行业领域的海

[1] Big data.http://en.wikipedia.org/wiki/Big_data.

[2] Benjamin Woo World wide Big Data Technology and Services 2012–2015 Forecast，2012-05.

量数据资源，包括经济、财政、社会、文化、科技等领域。其中绝大部分数据来自政府部门、大型企业和知名机构，并得到严格审核和认证。例如，国内生产总值(GDP)、我国年度粮食产量等数据均可以通过该网站获取。

图1-1　中国国家数据主页

(2) 中国统计年鉴 http://www.stats.gov.cn/sj/ndsj/

中国统计年鉴隶属于国家统计局，是国家统计局政务公开透明的平台。国家统计局承担组织、领导和协调全国统计工作的职能，并将其调查、收集、检测、整理、统计的数据通过中国统计年鉴公开，内容覆盖人口、国民经济核算、价格、固定资产投资等方面以及农业、工业、建筑业等行业。国家统计局每年都会通过统计年鉴收录上一年全国和各省、自治区、直辖市的经济和社会等各方面的大量统计数据，以及历史重要年份和近二十年的全国主要统计数据。中国统计年鉴是我国最全面、最具权威性的综合统计年鉴。

(3) CNNIC http://www.cnnic.net.cn/

CNNIC(China Internet Network Information Center，中国互联网网络中心)，是经国家主管部门批准，于1997年6月3日组建的管理和服务机构，行使国家互联网络信息中心的职责。每年CNNIC都会发布《中国互联网络发展状况统计报告》，发布时间为8月末。

(4) 城市开放数据。城市开放数据并非某个公开数据网站的明确名称，每个省、市及县级政府都会通过网站来公开与该地区相关的数据，我们可以从不同省、市、县级的政府官网中获取数据。例如，通过江苏省人民政府网站(http://www.js.gov.cn/col/col33688/index.html)，我们可以获取江苏省的相关公开数据。

2) 国际

(1) 世界银行 https://data.worldbank.org.cn/

世界银行成立于1945年，是国际三大金融机构之一。世界银行免费提供世界各国的公开发展数据，例如某国的GDP、人口、国际债务等数据。

(2) 世界不平等数据 https://wid.world/zh/data-cn/

该数据库收集了全球不平等指标，数据源自各大组织、机构及个人，涵盖不同国家

和地区的宏观经济、环境、政治和社会领域的不平等指标。

(3) 全球贸易数据 https://comtrade.un.org/

如图1-2所示，全球贸易数据主要提供全球各个国家和地区之间进出口、经济贸易及全球贸易总量等数据。

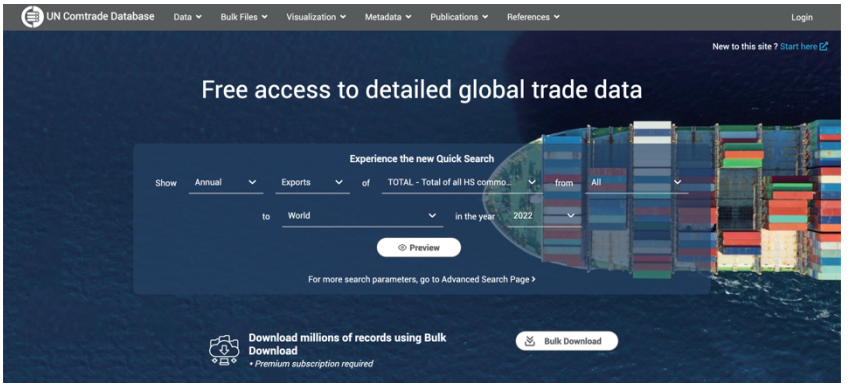

图1-2　全球贸易数据主页

(4) 环球金融数据 https://globalfi nancialdata.com/

环球金融数据提供150多个国家的历史和当前的金融、经济数据，包括利率、汇率、股票市场指数、商品价格，以及来自世界各主要市场的股票、债券和票据的总回报率等长期数据。

(5) WTO数据 https://www.wto.org/english/res_e/res_e.htm

WTO数据自1948年开始统计数据，主要提供货物贸易和服务贸易两方面的数据信息，涵盖贸易流量、商品贸易、服务贸易、国际投资和知识产权等方面。

(6) 联合国数据 http://data.un.org/

如图1-3所示，联合国数据主要提供联合国成员国的重要数据，涵盖各个国家的政治、经济、人口、交通、能源等方面。

图1-3　联合国数据主页

2. 交通出行数据

(1) 高德地图 https://report.amap.com/diagnosis/index.do

高德地图成立于2002年，是中国领先的数字地图内容、导航和位置服务解决方案提供商。通过高德地图，我们可以获取各个城市的交通路况、城市与城市间的交通情况对比等交通出行数据。

(2) 百度迁徙 https://qianxi.baidu.com/

百度迁徙于2014年春运期间由百度推出，旨在通过百度地图定位可视化大数据，汇报国内春节期间人口迁徙情况。如图1-4所示，这一工具能够帮助我们获取我国人口流动、实时航班以及交通状况等相关数据。

图1-4　百度迁徙页面

3. 影视娱乐数据

(1) 酷云收视率 https://www.ky.live/pc.html

酷云收视率主要提供观众观看特定电视频道、电影或电视剧的人数占总观众的比例等数据。

(2) 中国视听大数据(CVB) http://www.cavbd.cn/

“中国视听大数据”是“国家广播电视总局广播电视节目收视综合评价大数据系统”(CVB系统)对外发布信息时使用的名称。中国视听大数据的发布渠道有官方网站、新浪微博、微信公众号、百家号等。中国视听大数据建立并持续完善基于88项核心指标的多维度数据分析体系，为全国各级宣传管理和广播电视主管部门、播出机构、制作机构等用户提供多样化、个性化的数据分析服务。

4. 行研数据

(1) 阿里研究院 http://www.aliresearch.com/cn/index

如图1-5所示，阿里研究院提供有关电商、数字生活等领域的趋势数据报告，这些报告大多与阿里巴巴集团的相关业务有关。

图1-5　阿里研究院页面

(2) 腾讯调研云 https://research.tencent.com/

腾讯调研云是腾讯旗下的平台，主要发布与腾讯相关的数据报告。

(3) 艾瑞网 http://report.iresearch.cn/

艾瑞网是由艾瑞咨询打造的新经济门户站点，公开了艾瑞咨询集团多年来深入互联网及电信相关领域的研究成果，融合更多行业资源。

(4) 艾媒网 https://www.iimedia.cn/#shuju

艾媒网隶属于无线市场调研机构——艾媒市场咨询。如图1-6所示，用户可以搜索房地产、IT互联网、金融、人工智能、新零售、游戏、音乐、教育等各个行业的信息。

图1-6　艾媒网页面

(5) 易观分析 https://www.analysys.cn/

易观分析是易观国际推出的商业信息服务平台，该平台基于新媒体经济发展研究成果，提供结构化产业信息数据，涵盖网上零售、个人移动应用、银行创新业务、新媒体、互动娱乐、运营商新业务、终端等方面。

5. 社交媒体数据

(1) 微博数据中心 http://data.weibo.com/datacenter/recommendapp/?sudaref=www.baidu.com&display=0&retcode=6102

如图1-7所示，微博数据中心主要提供新浪微博平台的分析数据，包括各个账号的微博发布情况和粉丝情况等。这些数据可用于粉丝分析、评论内容分析、粉丝互动分析以及相关行业账号分析等。

图1-7　微博数据中心页面

(2) 清博智能 http://www.gsdata.cn/

清博智能是我国新媒体大数据平台，专注于舆论大数据和人工智能服务，可提供新闻门户网站、论坛、微信、微博、贴吧等社交平台的舆论数据。

6. 指数数据

(1) 百度指数 http://index.baidu.com/v2/index.html#/

百度指数依托于百度海量网民的行为数据，是一个数据分享平台。该平台可使用网页搜索和新闻搜索的海量数据，分析不同关键词在过去一段时间内的“用户关注度”和“媒体关注度”。

(2) 谷歌趋势 https://trends.google.com/trends/

谷歌趋势是Google公司建立的一个基于用户搜索行为的数据平台，如图1-8所示。该平台通过分析谷歌搜索引擎每天数十亿的搜索数据，提供不同时期某一关键词或者话题在谷歌搜索引擎中的搜索频率及相关统计数据。这些数据可用于市场研究、受众分析和产品营销方向确定等。许多学术研究所需的数据都来源于谷歌趋势。

图1-8　谷歌趋势页面

7. 汇总数据

(1) 大数据导航 http://hao.199it.com/

大数据导航汇集了国内外多个大数据网站和工具资源站的数据源和数据库。它定期更新并分享高质量的大数据资讯，涵盖政府数据、企业数据、社会数据等。

(2) 谷歌公共数据浏览器 https://datasetsearch.research.google.com/

谷歌公共数据浏览器是谷歌公司推出的一款数据可视化工具，它收集来自世界银行、欧盟统计局、美国劳工统计局和美国人口普查局等多个数据提供方的数据。

1.1.2　网络爬虫获取

网络爬虫(web crawler)又称为网络蜘蛛(web spider)或Web信息采集器，它是一种自动下载网页的计算机程序或自动化脚本，是搜索引擎的重要组成部分[1]。通常情况下，网络爬虫从一个称为“种子集”的URL集合开始运行，首先将这些URL全部存入一个有序的待爬行队列中，然后按照一定的顺序从队列中提取URL并下载所指向的页面，最后分析页面内容，提取新的URL并存入待爬行队列中。重复以上过程，直至待爬行队列为空或满足某个爬行终止条件，这一过程称为网络爬行(web crawling)[2]。如果将数据分析比作炼石油，爬虫的主要目的就是挖石油，以便进行下一步的处理工作。

1. 网络爬虫的作用

网络爬虫的价值其实就是数据的价值。网络爬虫常被用于获取一手网络数据，它免去了人工复制和粘贴的烦琐，实现了数据获取的快速和高效。在传统的人工获取数据的过程中，用户需要手动打开浏览器，提交请求，复制有价值的数据，粘贴并保存。网络爬虫作为循环的自动化程序，只需模拟浏览器发送请求，就能自动提取有价值的数据，然后将数据存入数据库或者文件中。这种方式缩短了数据获取时间，提升了数据获取的持续性，扩大了数据获取的容量。

2. 网络爬虫的基础知识

1) URL

URI(uniform resource identifier)是用于标识资源的标准标识符[3]。URL(uniform resource locator，统一资源定位器)是URI的一种具体形式。URL不仅能标识一个资源，还能指明如何定位这个资源。比如，https://www.Google.com/就是一个URL，用户通过

---

[1] Abukausar M，Dhaka V，Kumar S S. Web crawler：a review[J]. International Journal of Computer Applications，2013，63(2)：31-36.

[2] 秦雅琴，马玲玲. 网络爬虫技术在交通信息获取中的应用综述[J]. 武汉理工大学学报，2020，44(3)：456-461.

[3] Berners Lee T，Fielding R，Masinter L. Uniform Resource Identifiers(URI)：Generic Syntax.RFC2396，August 1998. http://www.ietf.org/rfc/rfc2396.txt.

该URL可以进入谷歌搜索页面。

2) HTML

HTML(hyper text markup language，超文本标记语言)[1]是一种用于创建网页的标准标记语言。用户可以使用HTML来建立自己的Web站点，HTML在浏览器上运行，并由浏览器解析。

(1) HTML的基本结构。HTML的基本结构如图1-9所示。

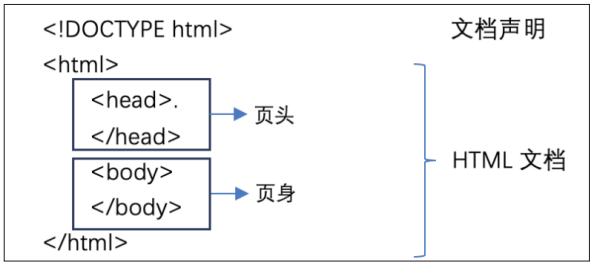

图1-9 HTML的基本结构

以下为HTML的基本示例。

```
<!DOCTYPE html>
<!-- doc===document文档  type===类型  html==文件类型
    作用：声明文档类型：HTML页面文件
-->
<html>
<!-- 超文本标记语言：所有的标签都需要放在html标签内部-->
<head>
    <!-- 头部：网页的头部 -->
    <meta charset="UTF-8">
        <!-- 定义字符编码格式  H5遵循的编码格式：utf-8 -->
    <title>我的第一个页面</title>
        <!-- 标题：网页的标题 -->
</head>
<body>
    <!-- 主体：网页的主体 -->
        <h1>这里是标题</h1>
        <p>网页的文字内容</p>
<!-- <h1> 与 </h1> 之间的文本被显示为标题
    <p> 与 </p> 之间的文本被显示为段落 -->
</body>
</html>
```

(2) HTML标签。在HTML中，一个网页通常以<html>开始，以</html>结束。HTML标签是由元素组成的，它们主要用于标记内容的不同模块，以及为这些模块赋予

[1] 张月琳，姚卓英，陈滢. Internet网络中的WWW系统及HTML语言[J]. 东南大学学报，1996.

含义。HTML标签通常使用尖括号包围，例如<html>和</html>，这两个标签表示一个HTML文件的开始和结束。

HTML标签有两种形式，一种是成对出现的标签，另一种是自闭合标签。无论是哪种标签，都不应包含空格。尽管不是所有开始标签都需要相应的结束标签，但最好还是两者都提供，这有助于提高网页的可读性和可维护性。如果开始标签和结束标签之间没有内容，那么可以将它们写成自闭合标签，例如
。下面介绍常用的HTML标签和相关属性。

① ：段落标签，用于定义网页中的文本段落，段落之间可自动换行。属性align可用于定义段落中文本的水平对齐方式。

②
：换行标签，用于在行与行之间创建换行，它是一个自闭合标签。

③ <h1> 到 <h6>：标题标签，用于定义不同级别的标题。<h1>表示一级标题，级别最高，字号最大；<h6>表示六级标题，级别最低，字号最小。

④ <blockquote>：块引用标签，可用于包含块级元素，而不仅仅是纯文本。

⑤ <img>：图片标签，用于插入图像。例如，属性src用于指定图像源文件；alt用于设置图像加载失败时的替代文本；width和height用于设置图像的宽度和高度；border用于指定图像边框样式；align用于设置图像在垂直方向上的对齐方式。

⑥ <a>：超链接标签，用于创建超链接。例如，属性href用于指定链接目标地址；target用于指定链接如何在浏览器中打开；_self表示在当前页面打开；_blank表示在新的空白窗口中打开。

示例：

(3) HTML元素。HTML文档由一系列元素(elements)组成，这些元素可以用来包围不同部分的内容，以特定的方式呈现或工作。简单来说，元素=起始标记(begin tag)+元素属性+元素内容+结束标记(end tag)。起始标签包含元素名称，被尖括号包围，表示元素从这里开始生效。结束标签与起始标签类似，只是在元素名称前加上斜杠，表示元素的结束。

(4) HTML属性。HTML元素通常拥有属性，属性=属性值+属性名。属性提供有关元素的额外信息，这些信息不会在实际内容中显示出来。属性始终以名称和值的形式出现，例如name="value"。这些属性一般出现在元素的开始标签中，它们可以用于提供附加信息。属性可以分为全局属性(适用于所有元素)和局部属性(仅适用于特定元素)两种类型。

(5) HTML DOM。HTML DOM(文档对象模型)定义了访问和操作HTML文档的标准方法。在HTML DOM中，一切事物都被视为节点，DOM将HTML文档表示为节点树的形式，以树结构表示HTML文档，如图1-10所示。

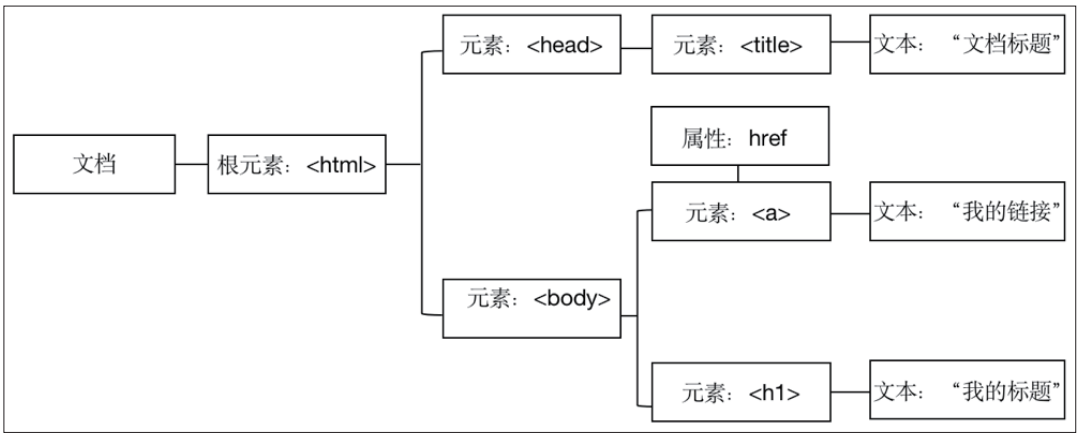

图1-10　HTML DOM tree

(6) HTML DOM节点树。根据W3C的HTML DOM标准，HTML文档中的所有内容都是节点。除文档节点外的每个节点都有父节点，大部分元素都有子节点。分享同一个父节点的节点是同辈，即兄弟节点。

整个文档是一个文档节点，每个HTML元素是元素节点，HTML元素内的文本是文本节点，每个HTML属性是属性节点，注释是注释节点。HTML DOM将HTML文档视作树结构，这种结构被称为节点树。所有HTML元素(节点)均可被创建、修改或删除。

节点树中的节点之间存在层级关系。父节点(parent)、子节点(child)和同胞节点(sibling)等术语用于描述这些关系。父节点拥有子节点。同级的子节点被称为同胞(兄弟或姐妹)。在节点树中，顶端节点被称为根(root)。除了根节点之外，每个节点都有父节点，一个节点可拥有任意数量的子节点。节点树中各节点之间的关系如图1-11所示。

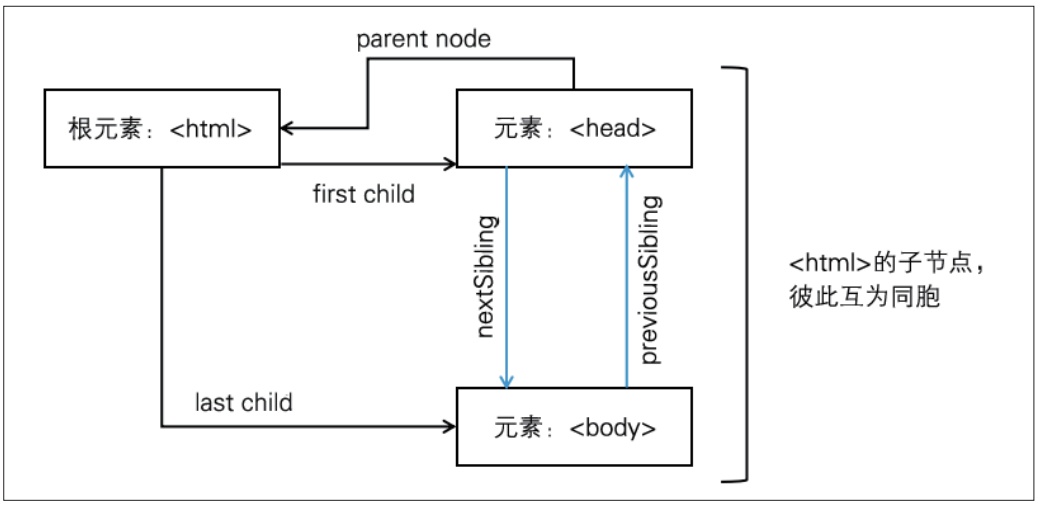

图1-11　节点树中各节点之间的关系

3) 正则表达式

正则表达式(regular expression)是一种文本模式，又称为规则表达式，它是由普通字符(例如，字母a到z)和特殊字符(称为“元字符”)组成的。正则表达式用于描述、匹配符合某种语法规则的文本字符串，通常用于检索、替换那些符合某个模式(规则)的文本[1]。正则表达式是一种强大的工具，它使用单个字符串来描述、匹配符合某种语法规则的字符串集合。正则表达式主要有以下4个用途。

(1) 搜索。正则表达式可用于在文档、源代码、日志等文本数据中查找匹配特定模式的文本。例如，用户可以使用正则表达式来搜索日志文件，以查找特定日期的日志条目。

(2) 验证。正则表达式可用于测试字符串是否符合特定的模式或格式，用户可使用正则表达式来验证输入的数据。例如，用户可以使用正则表达式验证电子邮件地址、电话号码、日期或密码是否符合预期的格式。

(3) 替换文本。正则表达式可用于将文本中的特定模式替换为其他内容，从而提高数据清洗和格式化的工作效率。例如，用户可以使用正则表达式来替换文本中的所有URL链接，或将电话号码的一部分替换为隐藏字符等。

(4) 提取。正则表达式可以基于模式匹配从字符串中提取子字符串，从文本中提取特定信息。例如，用户可利用正则表达式在网页内容中提取所有链接或抓取特定标签中的数据。

对于静态文本的简单搜索和替换操作，用户使用传统的文本查找和替换方法即可，而正则表达式的真正优势在于其具有灵活性，它能够处理动态文本，允许用户定义复杂的模式，以适应各种不同的情况，从而使文本处理更加强大和高效。

3. 主要爬虫工具

在新闻传播领域，针对社交媒体的研究不胜枚举，例如，对社交媒体上话题讨论度和影响力等的分析。通常，研究者需要对社交媒体上的原始数据进行挖掘和分析。数据是数据分析工作的核心，研究者要获取数据，通常需要使用不同类型的爬虫工具。目前，常见的爬虫工具大致可以分为爬虫软件和爬虫程序两类。爬虫软件包括GooSeeker、八爪鱼采集器、火车采集器等，爬虫程序主要使用Python、Java、R等编程语言开发。

1) 爬虫软件

(1) GooSeeker(集搜客)。GooSeeker是一款用于网页信息和数据爬取的数据软件，它可以在语义标注和结构化转换的基础上，实现网页信息和数据的抓取。相较于其他爬虫软件，即使在免费的情况下，GooSeeker也能执行几乎所有爬虫任务。作为一款简单易用的网页信息抓取软件，GooSeeker可以提取网页中的文字、图表、超链接等多种元

[1] 张长富，黄中敏. javascript动态网页编程实例手册[M].北京：海洋出版社，2005.

素，还提供数据挖掘攻略、行业资讯等功能。

GooSeeker的功能主要分布在客户端和官方网站上，如图1-12所示。GooSeeker的客户端采用浏览器布局，被形象地命名为“爬虫浏览器”。用户可以借助其内置的MS谋数台与DS打数台功能，通过可视化点击和确定采集规则等方式，对目标数据进行采集。除了客户端，GooSeeker的官方网站还提供一系列辅助功能，本章第1.2节将详细介绍GooSeeker在网络爬虫方面的应用。

图1-12　GooSeeker网页版页面

(2) 八爪鱼采集器。八爪鱼采集器是一款全网通用的互联网数据采集器[1]，它可以模拟用户浏览网页的行为，通过简单的页面点选，生成自动化采集流程，将网页数据转化为结构化数据，并存储到Excel或数据库中。八爪鱼采集系统的核心是一个自主研发的分布式云计算平台，该平台能在很短的时间内，从各种不同的网站或网页中轻松获取大量规范化数据。它能帮助那些需要从网页获取信息的用户实现数据的自动化采集、编辑和规范化，从而使用户摆脱对人工搜索和收集数据的依赖，降低信息获取成本，提高效率。综合来看，八爪鱼采集器是一款较为流行的爬虫软件，即使用户不懂编程，也能轻松抓取数据。八爪鱼采集器页面如图1-13所示。

图1-13　八爪鱼采集器页面

[1]　吴涛. 巧用八爪鱼采集器开展政务公开审计[J]. 审计月刊，2019(11)：32-33.

(3) 火车采集器。火车采集器(locoy spider)是一款网页抓取工具，它专门用于采集网站信息，包括文字、图片等内容，并支持多线程操作[1]。火车采集器被广泛应用于各大主流文章系统和论坛系统，是目前用户数量最多的互联网数据抓取、处理、分析和挖掘软件之一。它不受限于网页和内容的类型，同时支持分布式采集，因此具有较高的效率。然而，与其他工具相比，火车采集器的使用门槛较高，需要用户具备一定的网页和HTTP协议等方面的知识，深入了解工具操作流程，因此用户需要一些时间来熟悉操作方法。火车采集器页面如图1-14所示。

图1-14　火车采集器页面

2) 爬虫程序

(1) Python。Python是一种面向对象、解释型、通用、开源的脚本编程语言。它是目前最受欢迎的编程语言之一，广泛应用于Web开发、数据分析、人工智能、科学计算、桌面应用、游戏开发等多个领域[2]。Python是一种动态语言，因此更适合初学者，相较于Java、C、C++等其他编程语言，Python更简单，更容易上手。此外，Python具有很高的语言兼容性，代码相对简洁，因此经常被用作网络爬虫的主要工具。Python页面如图1-15所示，本章第1.3节将详细介绍Python在网络爬虫方面的应用。

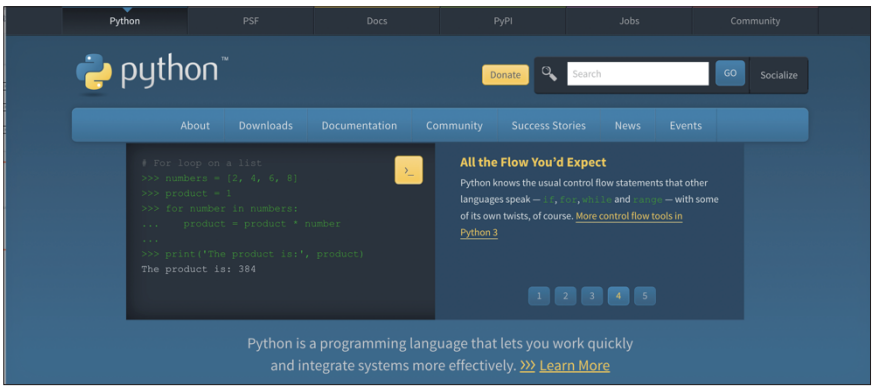

图1-15　Python页面

[1]　火车采集器. 信息数据采集论坛[EB/OL]. [2014-04-10]. http://bbs.locoy.com/.

[2]　钱程，阳小兰，朱福喜. 基于Python的网络爬虫技术[J]. 黑龙江科技信息，2016 (36) .

(2) Java。Java是世界上应用最广泛的编程语言之一，由Sun Microsystems在20世纪90年代开发。作为一种通用型语言，Java主要应用于网站后台开发、Android应用程序开发、大数据开发和客户端程序开发等。从Web应用程序到移动应用程序再到批处理应用程序，Java几乎涉及软件开发的每个领域[1]。在网络数据爬取方面，Java拥有众多高质量的爬虫库，但与其他工具相比，它的应用成本较高且较为复杂，因此越来越少的人将Java作为网络爬虫的首选工具。Java页面如图1-16所示。

图1-16　Java页面

(3) R。R是一种开源编程语言，被广泛用作统计软件和数据分析工具，是一种高级的统计、计算和可视化语言[2]。R语言是基于统计数据而创建的，因此它在数据分析师、数据科学家和统计学家群体中广受欢迎，它是仅次于Python的第二大数据科学编程语言。目前，R主要用于数据分析、绘图、数据挖掘和矩阵计算等领域。在网络数据爬取方面，丰富的模块和优雅直观的图表功能是其一大优势。R有两种获取数据的方式，一种方式是使用RCurl包和XML包，首先获取网页代码，然后解析HTML代码；另一种方式是使用rvest包，这种方式更加方便快捷。R页面如图1-17所示。

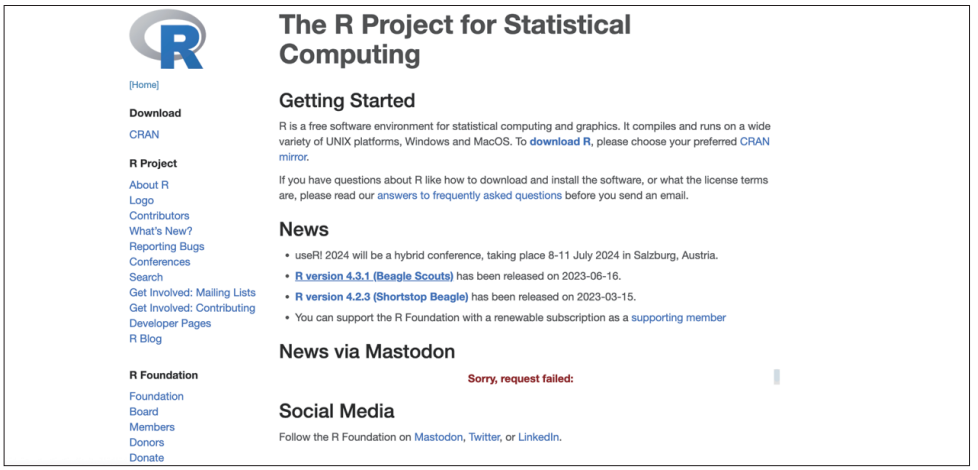

图1-17　R页面

[1] 冯键. Internet上开发软件的编程语言——Java编程语言[J]. 科技进步与对策，2001，18(7)：142-143.

[2] 王斌会. 多元统计分析及R语言建模[M]. 广州：暨南大学出版社，2010.

1.2　使用GooSeeker获取网络数据

网络爬虫是解决数据获取问题的有力工具，然而，数据爬取过程往往涉及编程，对用户的计算机技能水平有一定要求，导致普通用户难以在短时间内掌握爬虫方法。GooSeeker降低了数据获取的门槛，且简单易用，是众多不擅长编写复杂代码但渴望获取数据的用户的首选入门工具，特别适合初学者。

1.2.1　GooSeeker简介

GooSeeker的开发始于2007年，它是国内最早的网络爬虫工具之一。GooSeeker是一款基于云计算架构的网页数据提取工具包[1]，可以根据用户设定的规则自动抓取网页数据，包括文本、图片、表格、超链接等多种网页元素，并能按照一定的结构输出提取结果文件(通常为XML文件)。

GooSeeker的操作非常简单，软件内置丰富的模板资源，用户可一键抓取所需数据。相较于其他更专业的爬虫软件，GooSeeker更像一个具备数据采集功能的浏览器。在具体操作中，用户通过简单的鼠标拖拽操作就能生成爬虫程序，无须具备编程知识，只要了解和熟悉相关操作方法即可。

1.2.2　GooSeeker的优势与应用

1. GooSeeker的优势

GooSeeker融合了实用性和易用性，拥有强大的功能且免费，有独立的网络爬虫浏览器，用户可以免费爬取数十个网站数据，也可以付费请技术人员帮忙设置规则。在操作层面， GooSeeker主要具有两大优势。

(1) 直观点选，海量采集。使用GooSeeker，用户通过鼠标点选就能轻松进行数据采集，打破了专业技术背景的限制。GooSeeker具备强大的并发抓取功能，适用于处理大规模数据的场景。无论是处理动态网页还是静态网页、选择ajax还是html、抓取文本还是图片，GooSeeker都能应对自如，无须依赖其他软件，实现一站式采集。

(2) 文本分词和标签化。GooSeeker能自动进行文本分词，构建特征词库，并将文本标签化，从而生成特征词对应表。这个过程有助于多维度的量化计算和分析。用户可以利用这一功能迅速掌握行业动态，深入理解政策，从而把握市场机遇。

---

[1]　刘蓓琳，张琪. 基于购买决策过程的电子商务用户画像应用研究[J]. 商业经济研究，2017(24)：49-51.

2. GooSeeker的应用

近年来，GooSeeker已成功将互联网内容结构化和语义化技术应用于金融、保险、电信运营、电信设备制造、零售、电商、旅游、教育等各个行业。GooSeeker围绕自身核心产品，由一系列软件组件为各行业提供大数据解决方案，在不同领域有不同的应用。

(1) 内容聚合。GooSeeker将各个领域的信息汇聚起来并自动分类管理，从而形成行业垂直信息聚合平台。例如，金融和财经信息的汇总和管理。

(2) 市场情报与竞争分析。GooSeeker能分析零售市场中的竞争要素，包括定价、货架布局、促销活动、库存管理、品牌影响等，这有助于用户从电商网站提供的数据中获取竞争信息。

(3) 消费者洞察和品牌分析。GooSeeker通过聚合和挖掘消费者互动信息，研究消费者对产品的期望、产品与市场的契合度、品牌态度、品牌感知、品牌差距等，有助于更深入地展现品牌在市场中的表现。

(4) 商机发掘。将GooSeeker用于商圈分析，可确定最佳开店地点；将GooSeeker用于企业画像，可挖掘B2B销售机会；将GooSeeker用于需求分析，可确定潜在用户。GooSeeker通过提供多样化的数据解决方案，帮助用户解决核心问题。

1.2.3 GooSeeker的操作步骤

1. 下载与安装

用户可以从官方网站下载安装包，下载地址为 https://www.gooseeker.com/。如图1-18所示，选择“下载爬虫”，按提示操作即可。软件安装完成后，新用户需要在GooSeeker网站上注册账号，以便之后登录软件。

图1-18 GooSeeker下载页面

2. 制作与采集

(1) 打开MS谋数台。

(2) 输入目标网站的网址，按规则命名主题。

① 在MS谋数台的网址栏输入想要爬虫抓取的网页网址，按“Enter”键进行加载，用户可以在MS谋数台下方的浏览器窗口看到页面。

② 页面加载完后，在右边“工作台”中的“命名主题”下方的“主题名”栏，输入自定义的主题名，单击旁边的“查看”按钮，测试主题名是否已被占用。若提示“该名可以使用”，则命名成功。

(3) 新建整理箱。

① 单击右方的工作台中的“创建规则”按钮，再单击“新建”按钮，在弹出的窗口中输入想要命名的整理箱，例如命名为“列表”。

② 在整理箱中添加抓取内容。右击整理箱名称选择“添加—包含”，例如先添加“xxx”；继续添加则右击“xxx”，选择“添加—其后”，添加“yyy”等名称。

需要注意的是，整理箱中必须有一个是关键内容，用户需要选择一个抓取内容，将其设为关键内容，例如把“xxx”勾选为“关键内容”。

(4) 进行内容映射。

① 用户在浏览器窗口中单击想要获取的内容，例如单击“xxx”区域，这时MS谋数台会自动定位到HTML节点的位置(DIV节点)。

② 展开节点，找到“xxx”对应的text标签。

③ 右击这个text标签，选择“内容映射—xxx”。

④ 采用相同步骤，完成“yyy”的内容映射。

(5) 使用样例复制。如果用户需要抓取相同结构的数据，例如微博评论，更便捷的方法是使用样例复制，以下是抓取微博评论的示例步骤。

① 用户单击整理箱名称，即“列表”。

② 勾选右侧方的“启用”，开启样例复制功能。

③ 分别找到第一条评论和第二条评论对应的节点。

④ 右击第一条评论对应节点，选择“样例复制映射—第一个”。

⑤ 右击第二条评论对应节点，选择“样例复制映射—第二个”。在此过程中，用户可以单击右侧的“测试”按钮对当前的规则进行测试，检验结果是不是自己想要抓取的内容。

(6) 创建记号线索。在抓取数据过程中，用户可能需要对网页进行翻页。要解决翻页问题，需要创建一个“记号线索”。

① 单击右方工作台中的“爬虫路线”。

② 单击“新建”并勾选“记号线索”，创建记号线索。

③ 勾选“连续翻页”，这表示在执行抓取任务时，爬虫可以在同一个DS打数台窗口内抓取完成当前页面，之后直接跳到下一个页面进行抓取。

④ 由于翻页后将继续使用当前规则进行抓取，用户不需要更改目标主题名。

⑤ 右击网页上的“下一页”，找到定位节点，选择“翻页映射—作为翻页区—线索1”进行线索定位映射。

⑥ 右击网页上的“下一页”，找到定位节点，选择对应的text节点创建线索映射，右击text标签，选择“翻页映射—作为翻页记号”。

(7) 保存规则。在MS谋数台右侧单击“存规则”，这样用户就可以使用创建好的规则进行数据抓取。如果用户要搜索已保存的规则，可以在MS谋数台的“搜规则”工作台中输入已创建的规则名称进行搜索。

3. 数据抓取

(1) 打开DS打数台，在搜索框中输入所要使用的规则主题名。

(2) 右击主题名，在弹出的菜单中选择“统计线索”，可以看到有多少个线索等待抓取，线索就是网址。

(3) 单击“单搜”，DS打数台开始自动进行数据抓取工作，并会将结果以XML的格式存储下来。

(4) 如提示“没有线索了，可添加新线索或者激活已有的线索”，则说明线索已经采集完一遍。这时如果用户需要再次采集，可以右击主题名，选择“线索管理—激活所有线索”；如果要采集其他相同结果的网页，则选择“添加”，再把多个网址复制添加，就可以进行“批量采集”。

(5) 在DS打数台中，单击“爬虫群—启动”。

(6) 选择“会员中心—规则管理—我的规则”，单击“导入数据”，用户可以直接导入XML数据。

(7) 导出数据，页面显示导出成功后，即可下载。


案例：使用GooSeeker抓取豆瓣电影数据

本节提供案例：使用GooSeeker抓取豆瓣电影数据，请读者扫描二维码获取。

1.3 使用Python获取网络数据

Python广泛应用于新闻传播学领域，它可以协助从业者更好地理解受众兴趣、优化内容制作、评估内容效果以及把握舆论态势。在新闻传播学研究中，研究者可以通过Python从各种数据源中提取、整理和分析数据，从更广阔的视角了解新闻传播现象。在大数据获取的过程中，使用Python来获取相关数据成为一个极其重要的环节。

1.3.1　Python和PyCharm介绍

Python是一种功能强大、具有解释性和交互性、面向对象的第四代计算机编程语言。它是由荷兰人Guido van Rossum(吉多·范罗苏姆)于20世纪80年代末设计开发的，Guido van Rossum于2005年加入Google，继续领导和参与Python语言每一个版本的设计和开发工作[1]。Python这个称呼源自Guido van Rossum钟爱的电视剧*Monty Pyhton's Flying Circus*(《蒙提·派森的飞行马戏团》)。

Python代码相对复杂，且编写效率不高，通常需要借助PyCharm这个工具。PyCharm是一种专门为Python语言提供支持的集成开发环境(integrated development environment，IDE)，由JetBrains (一家捷克的软件开发公司)开发，它拥有一套强大的工具集，提供调试、语法高亮、项目管理、代码跳转、智能提示、自动完成等众多功能[2]，能够帮助开发者更好地理解和编写Python代码，显著提高编程效率。

Python和PyCharm相辅相成，为研究者获取、清洗、整理数据等提供了便利。Python作为一种代码解释器，可以将Python代码翻译成计算机能够明确理解的指令，因此，通过PyCharm编写Python程序时，需要Python解释器的支持。

1.3.2　Python的优势和适用领域

Python以其简单易学、可跨平台操作、拥有丰富的第三方库和庞大而活跃的生态系统等特点，赢得了众多研究者和科研工作者的喜爱，应用领域不断扩大。接下来，本节将简要介绍Python的优势和适用领域。

1. Python的优势

(1) 简单易学。Python采用清晰简洁的语法，易于理解和学习。初学者可以从基础的“hello world”程序开始，逐步深入学习相关语言特性和字符串操作等内容。

(2) 跨平台操作。Python可以在多个操作系统上运行，包括Windows、macOS等。这意味着用户可以在不同的系统平台上开发和运行Python程序，增强了灵活性和可移植性。

(3) 丰富的库和生态系统。Python拥有庞大的社区支持，提供大量的第三方库和工具，用户能够快速解决各种问题、完成各种任务。例如，Pandas库用于数据分析和处理，提供高效的数据结构和数据分析工具；Requests库用于发送HTTP请求；Matplotlib库用于数据可视化等。这些库大大提高了开发效率。

(4) 应用广泛。Python在人工智能、机器学习等领域非常流行，它提供了丰富的第三方库和框架支持。同时，Python广泛应用于Web开发、网络爬虫、数据分析等任务，

---

[1]　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.

[2]　郭建军，林丽君，何泽仁，等. 基于Python语言的按键脚本开发工具[J]. 科技创新导报，2019，16(23)：140-141.

具备广阔的发展前景。

2. Python的应用领域

Python是一种高级编程语言，具有简单易学、可读性强、通用性强的特点，广泛应用于商业、Web开发、人工智能、传播学等领域。

(1) 在商业领域，Python可以基于NumPy、Pandas等第三方库进行数据处理和分析，帮助企业从大量数据中提取有价值的信息，为商业合作提供数据预测和决策支持。例如，在电商在线评论中引入文本情感分析，有助于用户判断产品评论的情感倾向，以情感倾向为基础建立情感指数，从总体、店铺、月度等维度展开分析，用户能够更细致地了解电商在线评论中的情感倾向[1]。

(2) 在Web开发领域，Python的Web框架(如Django和Flask)可使Web应用程序的开发更加快速、高效。许多企业使用Python来构建电子商务网站、社交媒体平台、内容管理系统等应用。

(3) 在人工智能领域，Python具备的TensorFlow、Keras和PyTorch等工具，可用于机器学习和深度学习。企业可以借助这些工具来开发和部署各种应用，包括推荐系统、图像识别和自然语言处理等。

(4) 在传播学研究中，Python可以用于数据清洗、文本分析、情感分析、社交媒体数据分析等任务。例如，Python的Pandas库提供高效的数据处理工具，可用于清洗和整理数据；自然语言处理库支持文本分析和情感分析，有助于研究者从文本数据中提取关键信息和理解言论。Python还可用于从社交媒体平台(如微博和Twitter)获取海量数据，并加以分析。

在新媒体时代，Python作为重要的大数据获取技术，为研究者提供了极大的便利。因此，了解和掌握基本的Python知识对于相关领域的研究者和从业者来说至关重要。

1.3.3 Python和PyCharm的下载安装

1. Python的下载安装

(1) 安装Python。下载地址为https://www.python.org/。Python官网下载界面如图1-19所示，用户安装成功后，界面会显示“setup was successful”。

(2) 验证安装环境变量是否配置成功。用户可以使用快捷键Win+R，在弹出的窗口中输入cmd，单击“确定”，如图1-20所示。在命令提示符窗口中输入Python，按Enter键，随后出现如图1-21所示的页面，说明Python环境变量配置成功。由于版本不同，解释器版本内容显示可能会有所差异，用户可根据自己安装的版本来验证。

---

[1] 刘玉林，菅利荣. 基于文本情感分析的电商在线评论数据挖掘[J]. 统计与信息论坛，2018，33(12)：119-124.

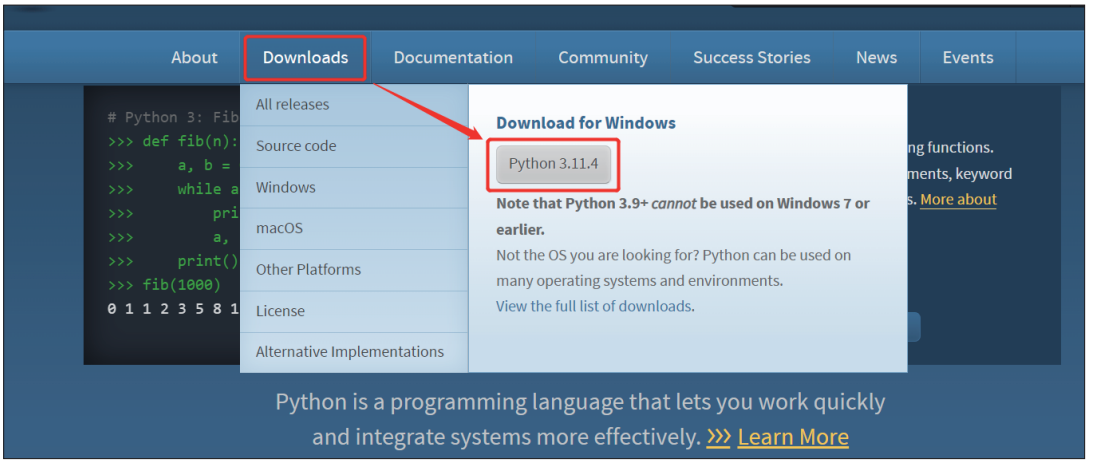

图1-19　Python官网下载界面

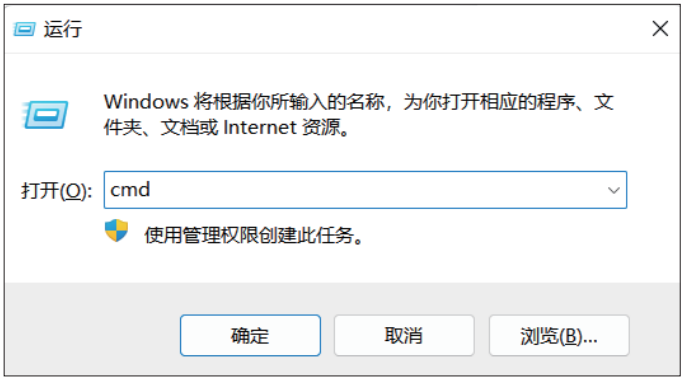

图1-20　使用快捷键Win+R后弹出的窗口

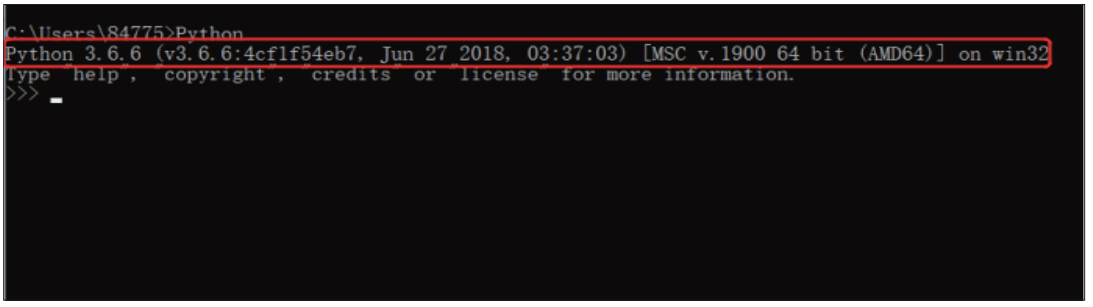

图1-21　Python验证窗口

2. PyCharm的下载安装

(1) 安装PyCharm。用户需要安装PyCharm来打开相关代码，下载地址为www.jetbrains.com/zh-cn/pycharm/download。Pycharm安装界面如图1-22所示。

(2) 根据实时界面设置安装环境。PyCharm在运行中会出现“No module named ××××”的提示，或在其右上方窗口显示红色波浪线，以此告知用户未安装某软件包，如图1-23所示。这时，需要用户手动安装Python库。手动安装这个库主要有四种方法：直接

在PyCharm里单击“安装软件包×××”，或在PyCharm的file—setting—Python解释器里搜索安装；在cmd里安装；通过Anaconda Prompt安装；直接复制其他用户装好的库。

图1-22　PyCharm安装界面

图1-23　未解析的引用

(3) 如图1-24所示，单击“运行”，即可获取数据。

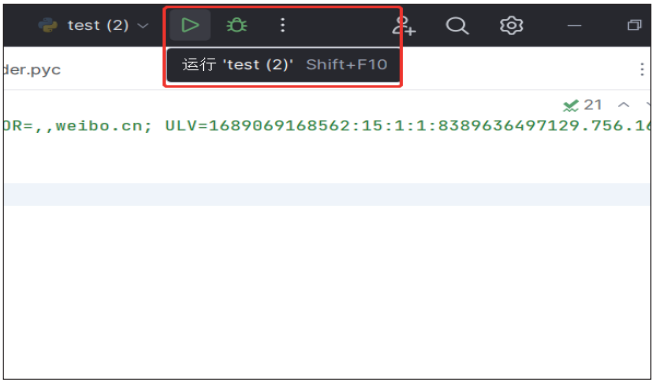

图1-24　运行结果页面

1.3.4　Python和PyCharm的使用步骤

1. Python的使用步骤

用户完成Python开发环境的安装后，就可以着手开发Python程序。下面以编写Python程序“hello world”为例介绍Python的应用方法。“hello world”是一个基本程序，用于打印输出“hello world”到屏幕上，以下为具体步骤。

(1) 如图1-25所示，打开命令提示符程序。

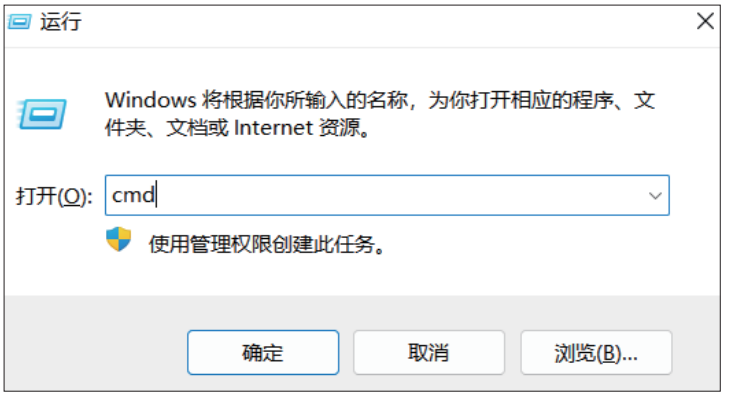

图1-25　打开命令提示符程序的界面

(2) 输入print(“hello world”)，按Enter键。print代表“打印、输出”，代码含义为引号内的内容。用户按Enter键后，如图1-26所示，当窗口显示“hello world”，表明该Python程序已编写完成。

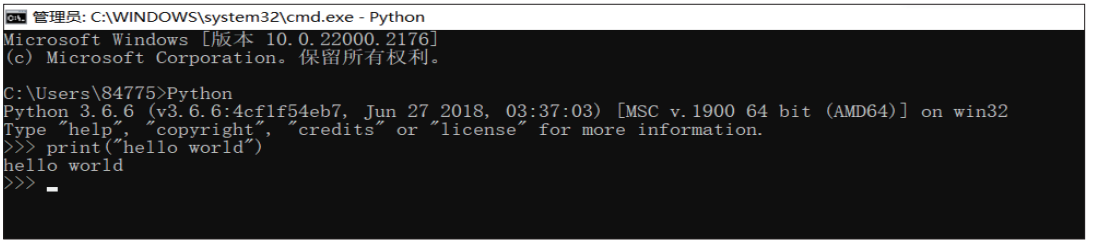

图1-26　打印输出“hello world”

注意：输入代码中使用的符号时，应切换至英文输入状态，否则会出现如图1-27所示的错误。

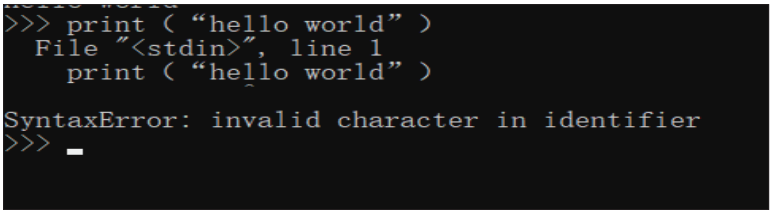

图1-27　错误示例

2. PyCharm的使用步骤

(1) 如图1-28所示，在左侧的“Project Files”中找到Python安装路径文件夹，右击“Python File”，创建文件并按Enter键。

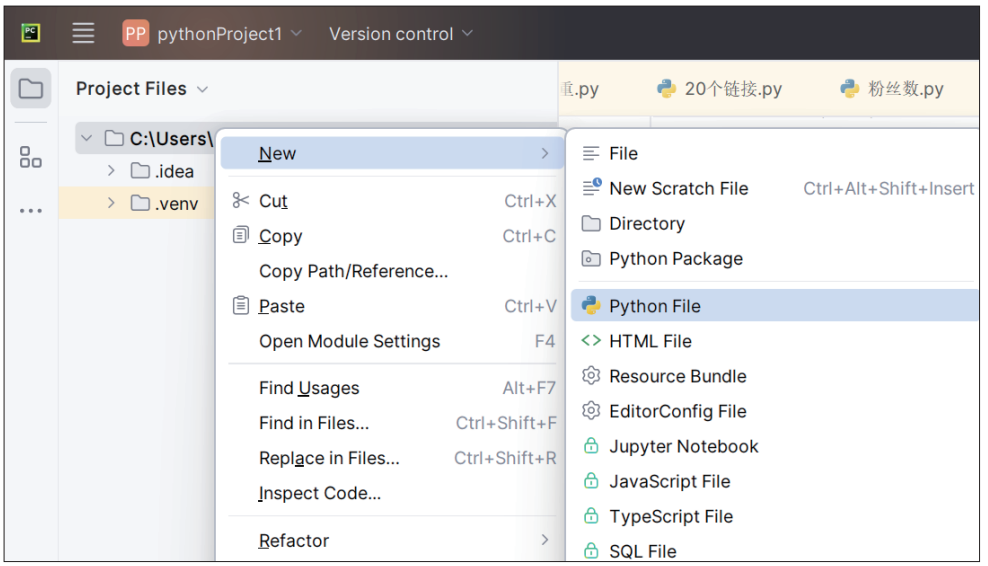

图1-28　创建文件

(2) 如图1-29所示，编写代码，以上文的“hello world”为例。

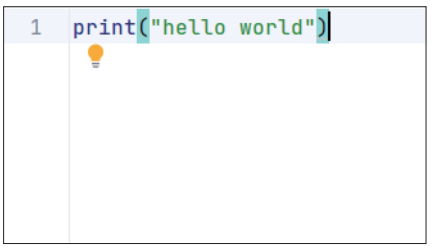

图1-29　编写代码

(3) 如图1-30所示，单击右键，选择“运行‘Python test’(U)”，开始运行代码。

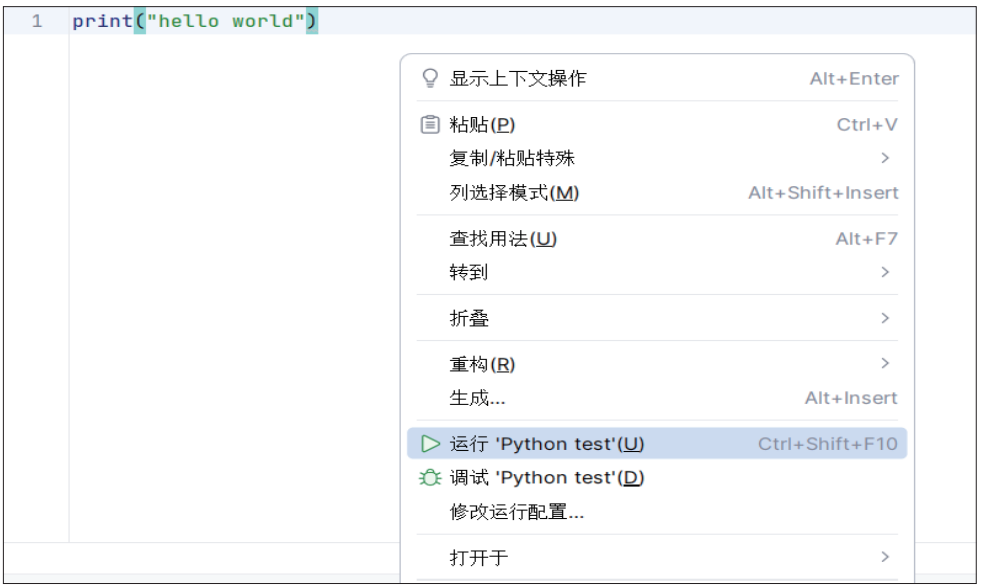

图1-30　运行代码

(4) 如图1-31所示，运行窗口显示的相关内容表示程序执行成功。

图1-31　运行结束

1.3.5　使用Python爬取网络数据的步骤

在Python编程中，要完成特定任务通常需要经历一系列步骤，包括明确编程目标、设计算法或逻辑、编写代码、测试运行以及后续的优化和维护。本节将通过一个简单的网页数据爬取示例来演示Python编程的相关步骤。

1. 导入模块

用户需要导入相关模块(module)，这些模块可以帮助用户快速实现某些功能。可以将模块看作一个大工具包，每个工具包中都有各种工具供用户使用，进而实现各种功能。

在Python中，用户要使用模块，需要使用import语句将模块导入到当前的代码中。导入的语法为：[from 模块名] import [模块|类|函数|][as 别名]。这里需要注意的是，“[]”在语法中表示“可选”的意思，即from内容和as内容可以省略。

在该案例中，需要导入urllib模块。urllib模块是Python标准库的一部分，无须额外安装即可使用，它提供URL处理功能。但仅仅通过import urllib是解决不了问题的，因为urllib是一个标准库模块，它还包含一系列子模块，例如urllib.request、urllib.parse、urllib.error、urllib.robotparser等，用于处理URL请求、URL解析、URL错误和机器人协议等。用户可以通过使用urllib.request来进一步打开URL并发送HTTP请求，获取URL返回的内容。

有读者会好奇：这里的“.”有什么作用？可以省去吗？答案是不可以。用户需要通过“.”来确定层级关系，例如urllib.request表明用户用到的是urllib的request子模块。

接下来就可以编写第一部分代码，导入用户所需要的模块。

```
import urllib.request
```

2. 将信息传入URL

在第一步中，用户导入了urllib.request模块，该模块提供一个简单的接口来打开URL并发送HTTP请求。在urllib.request子模块中，有一系列函数和类可供使用，其中常用的函数之一就是urlopen()。urlopen()是urllib.request子模块中的一个函数，用于打开URL并发送HTTP请求。接下来，用户需要将具体信息导入URL，即通过使用urllib.request.urlopen()函数打开一个URL。

在本例中，URL是“https://new.qq.com/”，即腾讯新闻网站。用户可以设置一个response变量，并通过“=”进行赋值。变量是在程序运行时，储存计算结果或表示值的抽象概念。简单来说，变量在程序运行时用于记录数据。每个变量都有存储的值(内容)，称为变量值。每个变量都有名称，称为变量名，也就是变量本身。将变量值(内容)赋予用户自定义的变量名，需要用到“=”。“=”表示赋值，即将等号右侧的值赋予左侧的变量。变量的定义格式为“变量名=变量值”。

接下来，用户可以进行第二步代码写作，将信息传入URL。

```
response=urllib.request.urlopen("https: //new.qq.com/")
```

3. 使用函数抓取网页内容

打开URL后，即可读取网页内容。在Python中，读取网页内容的下一步是使用read()函数。之前用户使用urllib.request.urlopen()函数打开了URL，这个函数返回一个HTTP Response对象。用户可以使用read()函数来读取这个响应对象的内容。用户通过调用response.read()获取网页内容(这里的“.”是Python中的点运算符，用于调用对象方法或访问对象属性)，并通过decode('utf-8')将内容转换为UTF-8编码的字符串。

这里需要大致说明函数和UTF-8的基础内容。

(1) 函数。函数是指组织好的、可重复使用的、用来实现特定功能的代码段。使用函数可以提高程序的复用性，减少重复性代码，提高开发效率。Python中的函数主要有三类，即内置函数、标准库函数和自定义函数。内置函数是由Python解释器提供的，无须导入任何模块即可直接使用，例如print()、len()、read()等。标准库函数是指模块中的函数，Python标准库中包含许多模块，每个模块都提供多个函数和类供用户使用，例如上文中的urllib.request.urlopen()。自定义函数是指Python允许用户自定义的函数，通过def关键字，用户可以在程序中定义函数，以便调用函数执行特定的任务。

(2) UTF-8(unicode transformation format-8-bit)。这是一种用于表示unicode字符的可变长度字符编码方式。unicode是一种字符集，它包含世界上几乎所有的字符，包括各种语言字符、符号、表情等。

在这里，将“response.read().decode('utf-8')”赋值给自定义变量content。

```
content=response.read().decode('utf-8')
```

4. 打印输出

完成以上步骤后，打印输出内容即可。

```
print(content)
```

以下是上述过程的完整代码。

```
import urllib.request
response=urllib.request.urlopen("https: //new.qq.com/")
content=response.read().decode('utf-8')
print(content)
```

到这里，一个基础的代码就编写完成了。代码编写完成后，用户需要完成程序运行、调试、更新与维护等后续操作。

本节提供案例：使用Python获取新浪微博数据，请读者扫描二维码获取。

案例：使用Python获取新浪微博数据

本章小结

本章主要探讨了大数据获取的重要性及方法。随着大数据时代的到来，数据从简单的处理对象开始转变为一种基础性资源，获取大数据成为利用数据资源的首要步骤。本章介绍了常见的两种数据获取方法，即网络公开数据获取及网络爬虫获取。

第1.1节简单概括网络爬虫的概念、作用、相关基础知识等，根据爬虫软件和爬虫程序的分类，介绍了常见的网络爬虫工具。为了便于读者进一步熟悉和掌握大数据获取的基本技术，第1.1节分别从两类爬虫工具中选择GooSeeker与Python进行简要介绍，第1.2节、1.3节对此进行详细说明。第1.2节介绍了GooSeeker的作用及主要应用领域，详细讲解了下载安装与操作步骤，展示了通过GooSeeker获取豆瓣电影数据的步骤。第1.3节介绍了Python和PyCharm的下载安装，展示了简单的网页数据爬虫操作。此外，本章还通过获取新浪微博“元宇宙”数据的案例，向读者展示网络爬虫技术的实际运用步骤，旨在抛砖引玉，启发读者，在日后的实际应用中不断精进操作能力。

获取大数据的网络爬虫工具多种多样，每个网站抓取的代码也各不相同，但背后的原理是相通的，为此需要辨析与掌握URL、HTML、正则表达式等相关的基础知识。学习并掌握网络爬虫技术，研究者和从业者可以更好地获取和利用网络数据，为新闻传播研究及实践提供有力支持。

核心概念

(1) 网络爬虫。网络爬虫又称为网络蜘蛛或Web信息采集器，是一种自动下载网页的计算机程序或自动化脚本，是搜索引擎的重要组成部分。

(2) URL。URL是URI的一种具体形式。URI是用于标识资源的标准标识符。URL不仅能标识一个资源，还能指明如何定位这个资源。URL由3部分组成，包括资源类型、存放资源的主机域名以及资源文件名。

(3) HTML。HTML是一种用于创建网页的标准标记语言。标准的HTML文件都具有一个基本的整体结构。在HTML中，一个网页从<html>开始，然后以</html>结束。HTML中的标签由元素组成，主要用于标记内容模块，同时也可以用来表明元素内容的意义。

(4) HTML DOM。HTML DOM(文档对象模型)定义了访问和操作HTML文档的标准方法。在HTML DOM中，一切事物都被视为节点，DOM将HTML文档表示为节点树形式，文档中的所有内容都是节点，每个节点(除文档节点外)都有父节点，大部分元素都有子节点，具有相同父节点的节点是同辈，也就是兄弟节点。

(5) 正则表达式。正则表达式是一种文本模式，又称为规则表达式，它是由普通字符和特殊字符组成的。正则表达式用于描述、匹配符合某种语法规则的文本字符串，通常用于检索、替换符合某个模式(规则)的文本。

(6) Python。Python是一种面向对象、解释型、通用、开源的脚本编程语言。它是目前最受欢迎的编程语言之一，广泛应用于Web开发、数据分析、人工智能、科学计算、桌面应用、游戏开发等多个领域。

思考题

(1) 什么是正则表达式？它与网络爬虫有什么关系？它在网络爬虫中起什么作用？

(2) 根据你所掌握的知识列举抓取新浪微博数据的方法。

(3) 运用GooSeeker采集京东网页中关于笔记本电脑的详细信息，包括标题、价格、评价数量、商品名称、显卡类别、裸机重量、屏幕尺寸、处理器等。翻页数量为5。

(4) 在案例“使用Python获取新浪微博数据”中，为什么需要找到网页的cookies内容？

(5) 请自行编写一段简单的代码，获取人民网首页的数据。

(6) 请使用Python的微博关键词爬虫工具，爬取关键词为“数字分身”的微博内容。