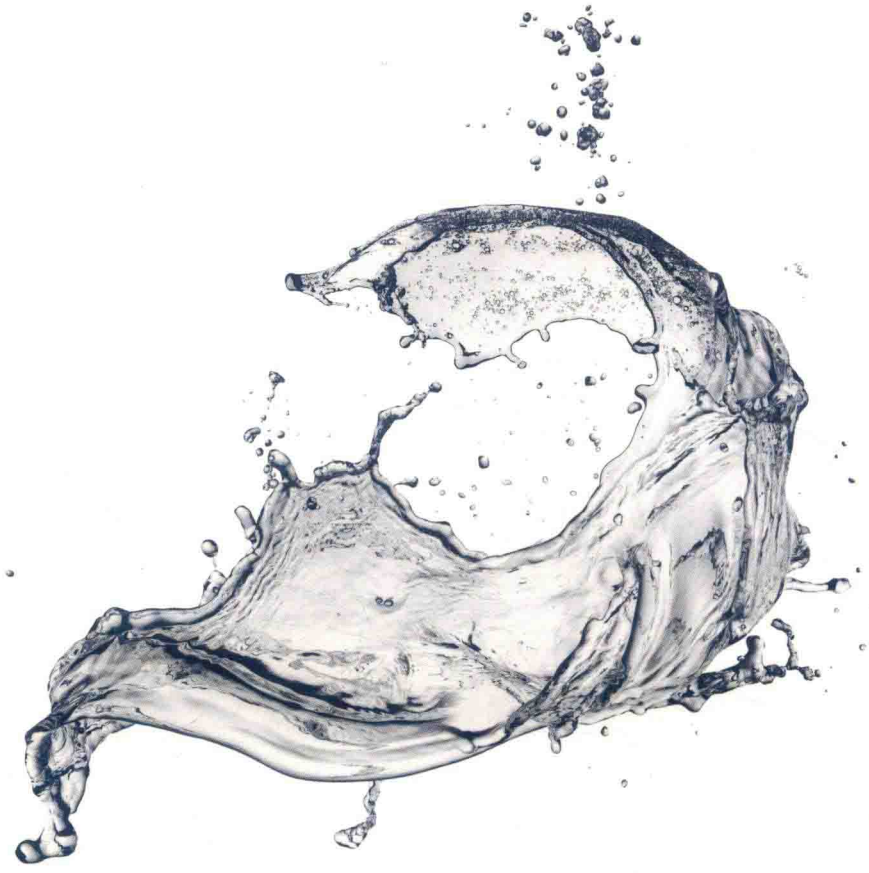
大数据人才培养规划教材

以解决实际问题为**学习目标**
以实战案例贯穿为**学习手段**



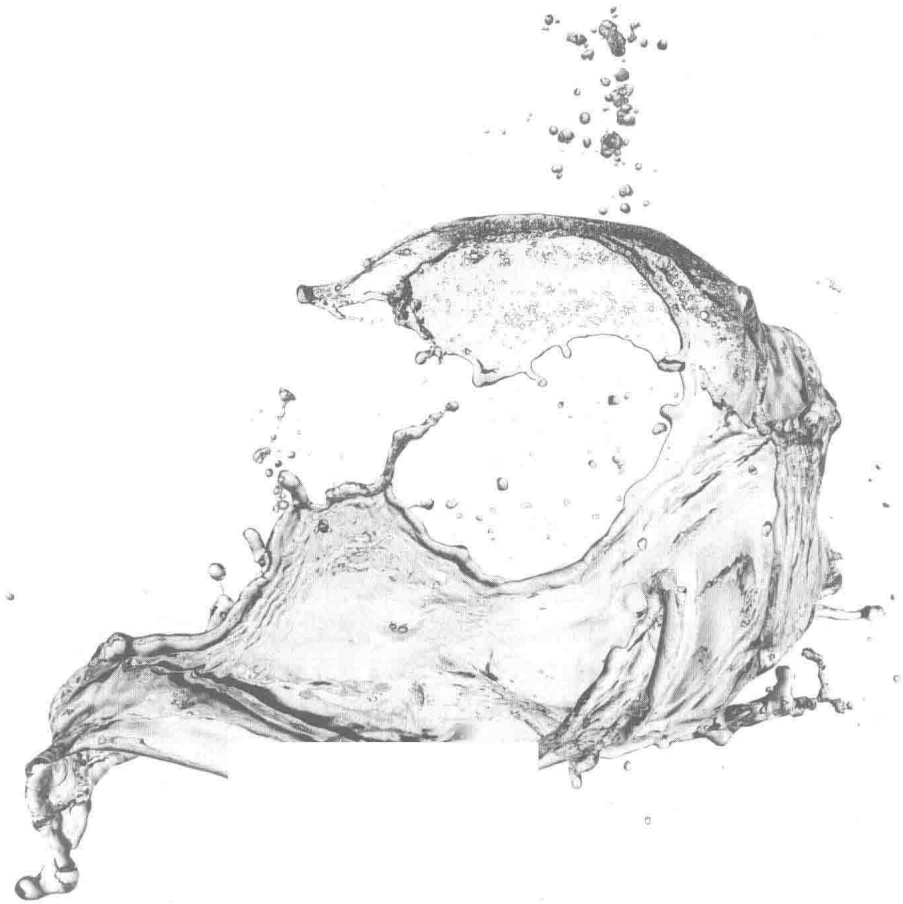
Python
网络爬虫技术

Web Scraping with Python

江吉彬 张良均 ◉ 主编
詹增荣 戴华炜 郭信佑 ◉ 副主编

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

大数据人才培养规划教材



Python 网络爬虫技术

Web Scraping with Python

江吉彬 张良均◉主编
詹增荣 戴华炜 郭信佑◉副主编

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

Python网络爬虫技术 / 江吉彬，张良均主编. -- 北京 : 人民邮电出版社，2019.4
大数据人才培养规划教材
ISBN 978-7-115-50506-4

Ⅰ. ①P… Ⅱ. ①江… ②张… Ⅲ. ①软件工具－程序设计－教材 Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字(2019)第001483号

内 容 提 要

本书以任务为导向，较为全面地介绍了不同场景下 Python 爬取网络数据的方法，包括静态网页、动态网页、登录后才能访问的网页、PC 客户端、App 等场景。全书共 7 章，第 1 章介绍了爬虫与反爬虫的基本概念，以及 Python 爬虫环境的配置，第 2 章介绍了爬取过程中涉及的网页前端基础，第 3 章介绍了在静态网页中爬取数据的过程，第 4 章介绍了在动态网页中爬取数据的过程，第 5 章介绍了对登录后才能访问的网页进行模拟登录的方法，第 6 章介绍了爬取 PC 客户端、App 的数据的方法，第 7 章介绍了使用 Scrapy 爬虫框架爬取数据的过程。本书所有章节都包含了实训与课后习题，通过练习和操作实战，可帮助读者巩固所学的内容。

本书可以作为高校大数据技术类专业的教材，也可作为大数据技术爱好者的自学用书。

◆ 主　　编　江吉彬　张良均
副 主 编　詹增荣　戴华炜　郭信佑
责任编辑　左仲海
责任印制　马振武

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：11　　2019 年 4 月第 1 版
字数：252 千字　　2019 年 4 月北京第 1 次印刷

定价：39.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

大数据专业系列图书
编写委员会

序 PREFACE

随着大数据时代的到来，移动互联网和智能手机迅速普及，多种形态的移动互联应用蓬勃发展，电子商务、云计算、互联网金融、物联网等不断渗透并重塑传统产业，大数据当之无愧地成了新的产业革命核心。

未来 5～10 年，我国大数据产业将会进入一个飞速发展时期，社会对大数据相关专业人才有着巨大的需求。目前，国内各大高校都在争相设立或准备设立大数据相关专业，以适应地方产业发展对战略性新兴产业的人才需求。

人才培养离不开教材，大数据专业是 2016 年才获批的新专业，目前还没有成套的系列教材，已有教材也存在企业案例缺失等亟须解决的问题。由广州泰迪智能科技有限公司和人民邮电出版社策划、校企联合编写的这套图书，犹如大旱中的甘露，可以有效解决高校大数据相关专业教材紧缺的困难。

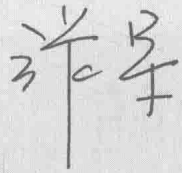
实践教学是在一定的理论指导下，通过引导学习者的实践活动，传承实践知识、形成技能、发展实践能力、提高综合素质的教学活动。目前，高校教学体系的设置有诸多限制因素，过多地偏向理论教学，课程设置与企业实际应用契合度不高，学生无法把理论转化为实践应用技能。课程内容设置方面看似繁多又各自为“政”，课程冗余、缺漏，体系不健全。本套图书的第一大特点就是注重学生实践能力的培养，根据高校实践教学中的痛点，首次提出“鱼骨教学法”的概念。以企业真实需求为导向，学生所学技能紧紧围绕企业实际应用需求，将学生需掌握的理论知识通过企业案例的形式进行衔接，达到知行合一、以用促学的目的。

大数据专业应该以大数据技术应用为核心，紧紧围绕大数据应用闭环的流程进行教学，才能够使学生从宏观上理解大数据技术在行业中的具体应用场景及应用方法。高校现有的大数据课程集中在如何进行数据处理、建模分析、参数调整，使得模型的结果更加准确。但是，完整的大数据应用却是一个容易被忽视的部分。本套图书的第二大特点就是围绕大数据应用的整个流程，从数据采集、数据迁移、数据存储、数据

分析与挖掘，最终到数据可视化，覆盖完整的大数据应用流程，涵盖企业大数据应用中的各个环节，符合企业大数据应用真实场景。

我很高兴看到这套书的出版，也希望这套书能给更多的高校师生带来教学上的便利，帮助读者尽快掌握本领，成为有用之才！

教育部长江学者特聘教授
国家杰出青年基金获得者
电气电子工程师学会会士（IEEE Fellow）
华南理工大学计算机科学与工程学院院长



2017 年 12 月

前 言 FOREWORD

随着云时代的来临，数据分析技术将帮助企业用户在合理时间内获取、管理、处理及整理海量数据，为企业经营决策提供积极的帮助。大数据分析作为一门前沿技术，广泛应用于物联网、云计算、移动互联网等战略性新兴产业。在大数据的研究和应用中，爬虫作为数据获取来源之一，扮演着至关重要的角色。

本书特色

本书作者以任务为导向，以将 Python 爬虫常用技术和真实案例相结合的方式，介绍使用 Python 进行数据爬取的主要方法。每一章都由任务、小结、实训和课后习题组成。设计思路以应用为导向，从而让读者明确所学知识是如何解决问题的。本书通过实训和课后习题巩固所学知识，使读者真正理解并能够应用所学知识。本书的内容由浅入深：第 1 章介绍了爬虫与反爬虫的基本概念，让读者在宏观上理解爬虫能够解决什么问题；第 2 章介绍了爬取过程中涉及的网页前端基础；第 3 ~ 6 章结合具体的的任务，介绍了 Python 在静态网页、动态网页、需要登录后才能访问的网页、PC 客户端、App 中爬取数据的方法；第 7 章介绍了使用 Scrapy 爬虫框架爬取数据的过程。

本书适用对象

- 开设有数据分析、Python 爬虫课程的高校的教师和学生

目前，国内不少高校将数据分析引入了教学，并在数学、计算机、自动化、电子信息、金融等专业开设了与大数据分析技术相关的课程，但目前这一课程的教学仍然主要限于理论介绍。单纯的理论教学过于抽象，学生理解起来往往比较困难，教学效果也不甚理想。本书提供的基于实践的教学模式，能够使师生充分发挥互动性和创造性，实现最佳的教学效果。

- 数据分析、Python 开发等相关人员

这类人员可以通过本书理解大数据分析技术中的数据获取的爬虫方法，并掌握相关实现方法，从而对爬虫技术有一个全面而深入的了解。

代码下载及问题反馈

为了帮助读者更好地使用本书，本书提供了相关计算过程的数据文件、Python 程序代码。读者可以从泰迪云课堂（https://edu.tipdm.org/course/96）免费下载，也可登录人民邮电出版社的人邮教育社区（http://www.ryjiaoyu.com）下载。此外，为了帮助读者

更好地学习，泰迪云课堂（https://edu.tipdm.org）还提供了配套的教学视频。

为方便教师授课，本书还提供了 PPT 课件，读者可以从泰迪云课堂（https://edu.tipdm.org/course/96）下载申请表，填写后发送至指定邮箱。其他图书资源，可通过热线电话（40068-40020）或以下微信公众号咨询获取。

我们已经尽最大努力避免在文本和代码中出现问题，但是由于水平有限，编写时间仓促，书中难免出现一些疏漏和不足的地方。如果您有相关的意见和建议，欢迎发送邮件至邮箱 13560356095@qq.com，期待能够得到您真挚的反馈。同时，本书内容更新将及时在泰迪云课堂（https://edu.tipdm.org/course/96）上发布，读者可以登录网站或关注泰迪大数据挖掘微信公众号（TipDataMining）查阅相关信息。更多本系列图书的信息可以在“泰迪杯”数据挖掘挑战赛网站（http://www.tipdm.org/tj/index.jhtml）上查阅。

编　者

2018 年 12 月

目录 CONTENTS

第 1 章 Python 爬虫环境与爬虫简介

随着互联网的快速发展，越来越多的信息被发布到互联网上。这些信息都被嵌入到各式各样的网站结构及样式当中，虽然搜索引擎可以辅助人们寻找到这些信息，但也拥有其局限性。通用的搜索引擎的目标是尽可能覆盖全网络，其无法针对特定的目的和需求进行索引。面对如今结构越来越复杂，且信息含量越来越密集的数据，通用的搜索引擎无法对其进行有效的发现和获取。在这样的环境和需求的影响下，网络爬虫应运而生，它为互联网数据的应用提供了新的方法。

学习目标

（1）认识爬虫的概念及原理。
（2）认识反爬虫的概念及对应爬取策略。
（3）掌握 Python 爬虫的环境配置方法。

任务 1.1 认识爬虫

任务描述

网络爬虫作为收集互联网数据的一种常用工具，近年来随着互联网的发展而快速发展。使用网络爬虫爬取网络数据首先需要了解网络爬虫的概念和主要分类，各类爬虫的系统结构、运作方式，常用的爬取策略，以及主要的应用场景，同时，出于版权和数据安全的考虑，还需了解目前有关爬虫应用的合法性及爬取网站时需要遵守的协议。

任务分析

（1）认识爬虫的概念。
（2）认识爬虫的原理。
（3）了解爬虫运作时应遵守的规则。

1.1.1 爬虫的概念

网络爬虫也被称为网络蜘蛛、网络机器人，是一个自动下载网页的计算机程序或自动化脚本。网络爬虫就像一只蜘蛛一样在互联网上爬行，它以一个被称为种子集的 URL 集合为起点，沿着 URL 的丝线爬行，下载每一个 URL 所指向的网页，分析页面内容，提取新

的 URL 并记录下每个已经爬行过的 URL，如此往复，直到 URL 队列为空或满足设定的终止条件为止，最终达到遍历 Web 的目的。

1.1.2 爬虫的原理

网络爬虫按照其系统结构和运作原理，大致可以分为 4 种：通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层网络爬虫。

1. 通用网络爬虫

通用网络爬虫又称全网爬虫，其爬取对象由一批种子 URL 扩充至整个 Web，主要由搜索引擎或大型 Web 服务提供商使用。这类爬虫的爬取范围和数量都非常大，对于爬取的速度及存储空间的要求都比较高，而对于爬取页面的顺序要求比较低，通常采用并行工作的方式来应对大量的待刷新页面。

该类爬虫比较适合为搜索引擎搜索广泛的主题，常用的爬取策略可分为深度优先策略和广度优先策略。

（1）深度优先策略

该策略的基本方法是按照深度由低到高的顺序，依次访问下一级网页链接，直到无法再深入为止。在完成一个爬取分支后，返回上一节点搜索其他链接，当遍历完全部链接后，爬取过程结束。这种策略比较适合垂直搜索或站内搜索，缺点是当爬取层次较深的站点时会造成巨大的资源浪费。

（2）广度优先策略

该策略按照网页内容目录层次的深浅进行爬取，优先爬取较浅层次的页面。当同一层中的页面全部爬取完毕后，爬虫再深入下一层。比起深度优先策略，广度优先策略能更有效地控制页面爬取的深度，避免当遇到一个无穷深层分支时无法结束爬取的问题。该策略不需要存储大量的中间节点，但是缺点是需要较长时间才能爬取到目录层次较深的页面。

2. 聚焦网络爬虫

聚焦网络爬虫又被称作主题网络爬虫，其最大的特点是只选择性地爬取与预设的主题相关的页面。与通用网络爬虫相比，聚焦爬虫仅需爬取与主题相关的页面，极大地节省硬件及网络资源，能更快地更新保存的页面，更好地满足特定人群对特定领域信息的需求。

按照页面内容和链接的重要性评价，聚焦网络爬虫策略可分为以下 4 种。

（1）基于内容评价的爬取策略

该策略将用户输入的查询词作为主题，包含查询词的页面被视为与主题相关的页面。其缺点为，仅包含查询词，无法评价页面与主题的相关性。

（2）基于链接结构评价的爬取策略

该策略将包含很多结构信息的半结构化文档 Web 页面用来评价链接的重要性，其中，一种广泛使用的算法为 PageRank 算法，该算法可用于排序搜索引擎信息检索中的查询结构，也可用于评价链接重要性，其每次选择 PageRank 值较大页面中的链接进行访问。

（3）基于增强学习的爬取策略

该策略将增强学习引入聚焦爬虫，利用贝叶斯分类器基于整个网页文本和链接文本来对超链接进行分类，计算每个链接的重要性，按照重要性决定链接的访问顺序。

（4）基于语境图的爬取策略

该策略通过建立语境图来学习网页之间的相关度，具体方法是，计算当前页面到相关页面的距离，距离越近的页面中的链接越优先访问。

3. 增量式网络爬虫

增量式网络爬虫只对已下载网页采取增量式更新，或只爬取新产生的及已经发生变化的网页，这种机制能够在某种程度上保证所爬取的页面尽可能的新。与其他周期性爬取和刷新页面的网络爬虫相比，增量式网络爬虫仅在需要的时候爬取新产生或者有更新的页面，而没有变化的页面则不进行爬取，能有效地减少数据下载量并及时更新已爬取过的网页，减少时间和存储空间上的浪费，但该算法的复杂度和实现难度更高。

增量式网络爬虫需要通过重新访问网页来对本地页面进行更新，从而保持本地集中存储的页面为最新页面，常用的方法有以下 3 种。

（1）统一更新法

爬虫以相同的频率访问所有网页，不受网页本身的改变频率的影响。

（2）个体更新法

爬虫根据个体网页的改变频率来决定重新访问各页面的频率。

（3）基于分类的更新法

爬虫按照网页变化频率将网页分为更新较快的网页和更新较慢的网页，并分别设定不同的频率来访问这两类网页。

为保证本地集中页面的质量，增量式网络爬虫需要对网页的重要性进行排序，常用的策略有广度优先策略和 PageRank 优先策略，其中，广度优先策略按照页面的深度层次进行排序，PageRank 优先策略按照页面的 PageRank 值进行排序。

4. 深层网络爬虫

Web 页面按照存在方式可以分为表层页面和深层页面两类。表层页面是指传统搜索引擎可以索引到的页面，以超链接可以到达的静态页面为主。深层页面是指大部分内容无法通过静态链接获取，隐藏在搜索表单后的，需要用户提交关键词后才能获得的 Web 页面，如一些登录后可见的网页。深层页面中可访问的信息量为表层页面中的几百倍，为目前互联网上发展最快和最大的新型信息资源。

深层网络爬虫爬取数据过程中，最重要的部分就是表单填写，包含以下两种类型。

（1）基于领域知识的表单填写

该方法一般会维持一个本体库，并通过语义分析来选取合适的关键词填写表单。该方法将数据表单按语义分配至各组中，对每组从多方面进行注解，并结合各组注解结果预测最终的注解标签。该方法也可以利用一个预定义的领域本体知识库来识别深层页面的内容，并利用来自 Web 站点的导航模式识别自动填写表单时所需进行的路径导航。

（2）基于网页结构分析的表单填写

该方法一般无领域知识或仅有有限的领域知识，其将 HTML 网页表示为 DOM 树形式，将表单区分为单属性表单和多属性表单，分别进行处理，从中提取表单各字段值。

1.1.3 爬虫的合法性与 robot.txt 协议

1. 爬虫的合法性

网络爬虫领域现在还处于早期的拓荒阶段，虽然已经由互联网行业自身的协议建立起一定的道德规范，但法律部分还在建立和完善中。

目前，多数网站允许将爬虫爬取的数据用于个人使用或者科学研究。但如果将爬取的数据用于其他用途，尤其是转载或者商业用途，则依据各网站的具体情况有不同的后果，严重的将会触犯法律或者引起民事纠纷。

同时，也需要注意，以下两种数据是不能爬取的，更不能用于商业用途。

（1）个人隐私数据，如姓名、手机号码、年龄、血型、婚姻情况等，爬取此类数据将会触犯个人信息保护法。

（2）明确禁止他人访问的数据，例如，用户设置过权限控制的账号、密码或加密过的内容等。

另外，还需注意版权相关问题，有作者署名的受版权保护的内容不允许爬取后随意转载或用于商业用途。

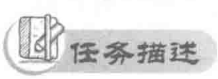
2. robot.txt 协议

当使用爬虫爬取网站的数据时，需要遵守网站所有者针对所有爬虫所制定的协议，这便是 robot.txt 协议。

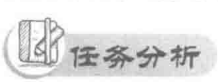
该协议通常存放在网站根目录下，里面规定了此网站中哪些内容可以被爬虫获取，以及哪些网页内容是不允许爬虫获取的。robot.txt 协议并不是一份规范，只是一个约定俗成的协议。爬虫应当遵守这份协议，否则很可能会被网站所有者封禁 IP，甚至网站所有者会采取进一步法律行动。在著名的百度与 360 的爬虫之争中，由于 360 没有遵守百度的 robot.txt 协议，爬取了百度网站的内容，而最终被判处 70 万元的罚款。

由于爬虫爬取网站时模拟的是用户的访问行为，所以必须约束自己的行为，接受网站所有者的规定，避免引起不必要的麻烦。

任务 1.2 认识反爬虫



网站所有者并不欢迎爬虫，往往会针对爬虫做出限制措施。爬虫制作者需要了解网站所有者反爬虫的原因和想要通过反爬虫达成的目的，并针对网站常用的爬虫检测方法和反爬虫手段，制定相应的爬取策略来规避网站的检测和限制。



（1）了解反爬虫的目的和常用手段。

（2）针对反爬虫的常用手段制定相应的爬取策略。

1.2.1 网站反爬虫的目的与手段

网站所有者从所有网站来访者中识别出爬虫并对其做出相应处理（通常为封禁 IP）的过程，被称为反爬虫。对于网站所有者而言，爬虫并不是一个受欢迎的客人。爬虫会消耗

大量的服务器资源，影响服务器的稳定性，增加运营的网络成本。可供免费查询的资源也有极大可能被竞争对手使用爬虫爬走，造成竞争力下降。以上种种因素导致网站所有者非常反感爬虫，想方设法阻止爬虫爬取自家网站的数据。

爬虫的行为与普通用户访问网站的行为极为类似，网站所有者在进行反爬虫时会尽可能地减少对普通用户的干扰。网站针对爬虫的检测方法通常分为以下几种。

1. 通过 User-Agent 校验反爬

浏览器在发送请求时，会附带一部分浏览器及当前系统环境的参数给服务器，这部分数据放在 HTTP 请求的 Headers 部分，Headers 的表现形式为 key-value 对，其中，User-Agent 标示一个浏览器的型号，图 1-1 所示为 Chrome 浏览器中某网页的 User-Agent。服务器会通过 User-Agent 的值来区分不同的浏览器。

图 1-1 Chrome 浏览器中某网页的 User-Agent

2. 通过访问频度反爬

普通用户通过浏览器访问网站的速度相对爬虫而言要慢得多，所以不少网站会利用这一点对访问频度设定一个阈值，如果一个 IP 单位时间内的访问频度超过预设的阈值，则网站将会对该 IP 做出访问限制。通常情况下，该 IP 需要经过验证码验证后才能继续正常访问，严重时，网站甚至会在一段时间内禁止该 IP 的访问。

3. 通过验证码校验反爬

与通过访问频度反爬不同，有部分网站不论访问频度如何，一定要来访者输入验证码才能继续操作。例如，在 12306 网站上，不管是登录还是购票，全部需要验证验证码，与访问频度无关。

4. 通过变换网页结构反爬

一些社交网站常常会更换网页结构，而爬虫大部分情况下都需要通过网页结构来解析需要的数据，所以这种做法也能起到反爬虫的作用。在网页结构变换后，爬虫往往无法在原本的网页位置找到原本需要的内容。

5. 通过账号权限反爬

还有部分网站需要登录才能继续操作，这部分网站虽然并不是为反爬虫才要求登录操作的，但确实起到了反爬虫的作用。

1.2.2 爬取策略制定

针对 1.2.1 节介绍的常见的反爬虫手段，可以制定相应的爬取策略。

1. 发送模拟 User-Agent

爬虫可通过发送模拟 User-Agent 来通过服务器的 User-Agent 检验，模拟 User-Agent 指的是，将要发送至网站服务器的请求的 User-Agent 值伪装成一般用户登录网站时使用的 User-Agent 值。通过这种方法能很好地规避服务器检验，有时有些服务器可能会禁止某种特定组合的 User-Agent 值，这时就需要通过手动指定来进行测试，直到试出服务器所禁止的组合，再进行规避即可。

2. 调整访问频度

目前，大部分网站都会通过 User-Agent 值做基础反爬检验，在此基础上，还有部分网站会再设置访问频度阈值，并通过访问频度反爬。爬虫爬取此类网站时，如果设置的访问频度不当，则有极大可能会遭到封禁或需要输入验证码，所以需要通过备用 IP 测试网站的访问频度阈值，然后设置比阈值略低的访问频度。这种方法既能保证爬取的稳定性，又能使效率不至于过于低下。如果仍然觉得访问频度设置得不足以满足需求，那么可以考虑使用异步爬虫和分布式爬虫。

3. 通过验证码校验

若因为访问频度问题导致需要通过验证码检验，则按照访问频度的方案实施即可，也可以通过使用 IP 代理或更换爬虫 IP 的方法来规避反爬虫。但对于一定要输入验证码才能进行操作的网站，则只能通过算法识别验证码或使用 Cookie 绕过验证码才能进行后续操作。需要注意的是，Cookie 有可能过期，过期的 Cookie 无法使用。

4. 应对网站结构变化

根据爬取需求，应对这类网站的方法可分为两种：如果只爬取一次，那么要尽量赶在其网站结构调整之前，将需要的数据全部爬取下来；如果需要持续性爬取，那么可以使用脚本对网站结构进行监测，若结构发生变化，则发出告警并及时停止爬虫，避免爬取过多无效数据。

5. 通过账号权限限制

对于需要登录的网站，可通过模拟登录的方法进行规避。模拟登录时除需要提交账号和密码外，往往也需要通过验证码检验。

6. 通过代理 IP 规避

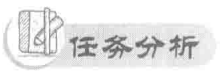
网站识别爬虫进行反爬虫的一个常用标识就是 IP，通过代理进行 IP 更换能够有效地规避网站的检测。需要注意的是，公用 IP 代理池往往已经被网站所有者识别为重点监测对象，使用这些公用 IP 代理时需要注意。

任务 1.3 配置 Python 爬虫环境

任务描述

Python 中整合了许多用于爬虫开发的库，使用 Python 开发爬虫需要了解 Python 中常用的爬虫库，各爬虫库的特性、功能和配置方法。爬虫爬取的数据需要存储在数据库中，本任务可使读者了解在 Windows 和 Linux 环境下的 MySQL 数据库和 MongoDB 数据库的

配置方法。



（1）了解 Python 中常用的爬虫库。
（2）掌握 MySQL 数据库的配置方法。
（3）掌握 MongoDB 数据库的配置方法。

1.3.1 Python 爬虫相关库介绍与配置

目前，Python 有着形形色色的与爬虫相关的库，按照库的功能，整理可得表 1-1。

表 1-1 与爬虫相关的库

类 型	库 名	简 介
通用	urllib	urllib 是 Python 内置的 HTTP 请求库，提供一系列用于操作 URL 的功能
	Requests	基于 urllib，采用 Apache2 Licensed 开源协议的 HTTP 库
	urllib 3	urllib 3 提供很多 Python 标准库里所没有的重要特性：线程安全，连接池，客户端 SSL/TLS 验证，文件分部编码上传，协助处理重复请求和 HTTP 重定位，支持压缩编码，支持 HTTP 和 SOCKS 代理，100% 测试覆盖率
框架	Scrapy	Scrapy 是一个为爬取网站数据、提取结构性数据而编写的应用框架。可应用在数据挖掘、信息处理或历史数据存储等一系列的程序中
HTML/XML 解析器	lxml	C 语言编写的高效 HTML/XML 处理库，支持 XPath
	Beautiful Soup 4	纯 Python 实现的 HTML/XML 处理库，效率相对较低

除 Python 自带的 urllib 库外，Requests、urllib 3、Scrapy、lxml 和 Beautiful Soup 4 等库都可以通过 pip 工具进行安装。pip 工具支持直接在命令行上运行，但需将 Python 安装路径下的 scripts 目录加入到环境变量 Path 中。另外，pip 工具支持指定版本库的安装，通过使用==、>=、<=、>、<符号来指定版本号。同时，如果有 requirements.txt 文件，也可使用 pip 工具来调用。使用 pip 工具安装 Requests 库的程序如代码 1-1 所示。

代码 1-1 使用 pip 工具安装 Requests 库

```
pip install requests  # 安装 Requests 库
pip install 'requests <2.19.0'  # 安装特定版本的 Requests 库
pip install 'requests >2.18.3,<2.19.0'
pip install -r requirements.txt  # 调用 requirements.txt 文件
```

1.3.2 配置 MySQL 数据库

MySQL 是目前广泛应用的关系型数据库管理系统之一，由瑞典 MySQL AB 公司开发，现属于 Oracle 公司。关系型数据库将数据保存在不同的表中，来增加运行速度和存储的灵活性。由于 MySQL 数据库具备体积小、速度快、成本低、开放源码等特点，中小型网站