《数据采集与爬虫技术》教学大纲

**课程名称：数据采集与爬虫技术**

**课程类别：**必修

**适用专业：**大数据技术类相关专业

**总学时：**64学时（其中理论24学时，实验40学时）

**总学分：**4.0学分

# 课程的性质

数字经济时代，数字资源已经成为互联网竞争和经营的生产要素和核心竞争力，而如何获取数据资源并基于此产出有价值的数据，已成为重要的资源配置。数据企业能够收集、获取的数据越多，越可能在行业竞争中具有优势地位。行业的发展带动岗位的需求，越来越多的爬虫工程师岗位涌现，工作中对爬虫技术的需求也越来越多。网络爬虫技术是数据分析、数据挖掘、人工智能等技术的数据基础，是从互联网上批量获取数据的重要技术之一，特开设Python网络爬虫技术课程。

# 课程的任务

通过本课程的学习，掌握使用Python基本语法完成爬虫任务编写，使用Requests库向指定网址发送请求，Xpath或Beautiful Soup库对静态网页进行解析，Selenium库爬取动态页面；使用JSON文件、MySQL数据库、MongoDB数据库对爬取下来的数据进行存储；使用表单登录方法、Cookie登录方法实现模拟登录；使用HTTP Analyzer和Fiddler工具抓包，并分析终端协议；使用Scrapy框架进行网页内容爬取，理论结合实践，每个章节中都配有多个案例，为学生将来从事数据采集、数据爬取的工作、研究奠定基础。

# 课程学时分配

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 教学内容 | 理论学时 | 实验学时 | 其它 |
| 1 | 第1章 了解爬虫与Python爬虫环境 | 2 | 2 |  |
| 2 | 第2章 爬虫基础知识准备 | 2 | 4 |  |
| 3 | 第3章 简单静态网页爬取——获取某企业官网基本信息 | 6 | 10 |  |
| 4 | 第4章 爬取动态网页——获取图书基本信息 | 6 | 8 |  |
| 5 | 第5章 模拟登录——登录某企业官网 | 2 | 4 |  |
| 6 | 第6章 终端协议分析——获取某音乐PC客户端和APP数据 | 2 | 6 |  |
| 7 | 第7章 使用Scrapy爬虫——爬取某企业官网新闻动态 | 4 | 6 |  |
|  | 总计 | 24 | 40 |  |

# 教学内容及学时安排

### 理论教学

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **章节名称** | **主要内容** | **教学目标** | **学时** |
| 1 | 了解爬虫与Python爬虫环境 | 1. 爬虫的概念 2. 爬虫的分类 3. 爬虫的合法性与robots协议 4. 网站反爬虫的目的与手段 5. 爬取策略制定 6. 爬虫需要用到的编程环境、工具介绍 | 1.掌握爬虫的概念和分类  2.了解爬虫的法律问题  3.掌握爬虫robots协议以及该协议的常见字段  4.了解反爬虫的目的和常用手段  5.掌握制定爬虫策略的方法  6.了解爬虫相关库及环境配置的方法 | 2 |
| 2 | 爬虫基础知识准备 | 1. 了解网页开发技术 2. 了解网页的结构 3. 了解网页的分类 4. 了解网页的数据结构 5. 熟悉HTTP请求方法与过程 6. 熟悉常见HTTP状态码 7. 熟悉HTTP头部信息 8. 熟悉Cookie | 1. 了解常用的网页开发技术 2. 掌握网页的结构和常用的标签，并能完成简单的网页创建 3. 了解不同类型网页的区别，并能加以区分 4. 掌握网页中常用的数据结构 5. 掌握HTTP请求方法和过程，并能分析HTTP协议中的头部信息 | 2 |
| 3 | 简单静态网页爬取——获取某企业官网基本信息 | 1. 使用Chrome浏览器查看网页 2. 使用urllib3库实现HTTP请求 3. 使用Requests库实现HTTP请求 4. 使用Xpath解析网页 5. 使用Beautiful Soup库解析网页 6. 使用正则表达式解析网页 7. 将数据存储为JSON文件 8. 将数据存储到MySQL数据库 | 1. 掌握Chrome浏览器开发者工具的使用方法 2. 掌握使用urllib 3和Requests库实现HTTP请求的流程 3. 掌握Xpath的基本语法和常用函数 4. 掌握创建Beautiful Soup对象的方法 5. 掌握正则表达式解析网页的常用方法 6. 掌握使用JSON模块和PyMySQL库存储数据的方法 | 6 |
| 4 | 爬取动态网页——获取图书基本信息 | 1. 逆向分析判断网页类型 2. 逆向分析获取动态网页信息 3. 使用Selenium库打开浏览器对象并访问页面 4. Selenium库进行元素选取、交互、页面操作、等待 5. MongoDB数据库和MySQL数据库的区别 6. 将数据存储到MongoDB数据库 | 1. 掌握逆向分析定位动态数据所在网址的方法 2. 掌握Selenium库的安装和配置方法 3. 掌握Selenium库的常用函数 4. 了解MongoDB数据库和MySQL数据库的区别 5. 掌握MongoDB存储数据的基本方法 | 6 |
| 5 | 模拟登录——登录某企业官网 | 1. 查找提交入口 2. 查找并获取需要提交的表单数据 3. 使用POST请求方法登录 4. 使用浏览器Cookie登录 5. 基于表单登录的Cookie登录 | 1. 掌握表单登录和Cookie登录的流程 2. 掌握使用Chrome开发者工具查找登录入口的方法 3. 掌握发送POST请求的方法 4. 掌握保存和加载Cookie实现登录的方法 | 2 |
| 6 | 终端协议分析——获取某音乐PC客户端和APP数据 | 1. 了解HTTP Analyzer工具 2. 爬取酷我音乐PC客户端数据 3. 了解Fiddler工具 4. 分析酷我音乐APP数据 | 1. 了解HTTP Analyzer工具的基础知识 2. 掌握酷我音乐PC客户端数据的爬取流程 3. 了解Fiddler工具的基础知识 4. 掌握HTTP Analyzer和Fiddler工具的抓包和数据分析方法 5. 掌握酷我音乐App客户端的数据分析方法 | 2 |
| 7 | 使用Scrapy爬虫——爬取某企业官网新闻动态 | 1. 了解Scrapy爬虫框架 2. 熟悉Scrapy常用命令 3. 创建Scrapy爬虫项目 4. 修改items/pipelines脚本 5. 编写spider脚本 6. 修改settings脚本 7. 定制下载中间件 8. 定制Spider中间件 | 1. 了解Scrapy框架各组件的作用以及数据流向 2. 熟悉Scrapy的常用命令及其作用 3. 掌握Scrapy爬取文本信息的方法 4. 掌握下载器中间件、Spider中间件的定制方法 | 4 |
| 学 时 合 计 | | | | 24 |

### 实验教学

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **实验项目名称** | **实验要求** | **学时** |
| 1 | Python网络爬虫环境搭建 | 1. 在Windows/Linux系统上安装Python、PyCharm、MySQL、MongoDB数据库 2. 使用PyCharm进行代码编写、运行、安装卸载第三方库操作 3. 学会MySQL、MongoDB数据库配置操作 | 2 |
| 2 | 爬虫基础知识准备 | 1. 使用html标签和CSS样式创建网页 2. 分析HTTP请求的过程 3. 使用浏览器查看HTTP报文并分析报文的头部信息 4. 从浏览器中查看报文中Cookie信息的方法 | 4 |
| 3 | 简单静态网页爬取——获取某企业官网基本信息 | 1. 使用Chrome浏览器查看网页源码，并定位节点 2. 使用urllib3和Requests实现HTTP请求发送和接收响应结果 3. 使用Xpath、Beautiful Soup或者正则表达式对响应的内容进行解析 4. 将数据存储到JSON文件中 5. 将数据存储到MySQL数据库中 | 10 |
| 4 | 爬取动态网页——获取图书基本信息 | 1. 判断网页类型 2. 使用逆向分析的方法获取网页信息 3. Selenium库及浏览器驱动安装 4. 使用Selenium库爬取动态网页 5. 网页元素选取 6. 网页元素交互 7. 网页页面操作 8. 网页页面等待 9. 将数据存储到MongoDB数据库中 | 8 |
| 5 | 模拟登录——登录某企业官网 | 1. 查找提交的入口 2. 查找并获取需要提交的表单数据 3. 使用POST方法发送登录请求实现模拟登录 4. 从浏览器中查找Cookie 5. 使用Cookie模拟登录 | 4 |
| 6 | 终端协议分析——获取某音乐PC客户端和APP数据 | 1. 安装HTTP Analyzer工具，并掌握基本用法 2. 使用HTTP Analyzer爬取酷我音乐PC端数据 3. 安装Fiddler工具，并掌握基本用法 4. 使用Fiddler分析酷我音乐APP，并获取数据 | 6 |
| 7 | 使用Scrapy爬虫——爬取某企业官网新闻动态 | 1. 使用Scrapy常用命令 2. 通过Scrapy爬取网页信息 3. 修改items/pipelines脚本 4. 编写spider脚本 5. 修改settings脚本 6. 根据需求定制下载器中间件 7. 根据需求定制Spider中间件 | 6 |
| 学 时 合 计 | | | 40 |

# 考核方式

突出学生解决实际问题的能力，加强过程性考核。课程考核的成绩构成 = 平时作业（10%）+ 课堂参与（20%）+ 期末考核（70%），期末考试建议采用开卷形式，试题应包括发送HTTP请求、解析静态网页内容、解析动态网页内容、数据存储为JSON文件、数据存储到MySQL数据库、数据存储到MongoDB数据库、使用表单和Cookie模拟登录、使用HTTP Analyzer获取PC端数据、使用Fiddler获取APP端数据、Scrapy框架使用等部分，题型可采用判断题、选择、简答、编程题等方式。

# 教材与参考资料

### 教材

池瑞楠．Python网络爬虫技术（第2版）（微课版）[M]．人民邮电出版社．2023．

### 参考资料

1. 江吉彬．Python网络爬虫技术 [M]．人民邮电出版社．2019．
2. 黄瑞军．Python爬虫项目教程（微课版）[M]．人民邮电出版社．2021．
3. 黑马程序员．Python网络爬虫基础教程 [M] ．人民邮电出版社．2022．