**《Python网络爬虫与信息提取》学习大纲**

**课程概述：**

该课程旨在向学生介绍使用Python进行网络数据抓取和信息提取的技术和方法，是一门旨在教授学生如何使用Python语言进行网络数据爬取和信息提取的课程。本课程涵盖了网络爬虫的工作原理、相关技术和常用的数据提取方法。学生将学习基本的网络爬虫原理和HTTP协议，了解如何通过发送HTTP请求和接收响应来访问网页。接着，学生将学习如何使用Python库（如requests和beautifulsoup）来编写简单的爬虫程序，并从网页中提取所需的信息。之后，将进一步学习如何处理动态网页，涵盖使用Selenium库和PhantomJS等工具模拟浏览器行为进行数据提取。学生将学习如何处理JavaScript生成的网页内容，并使用自动化工具完成数据提取。最后，通过练习项目和实践案例，学生将应用所学知识，自己动手编写并执行爬虫程序，从实际的网站上采集和提取所需的数据，并进行保存、处理和分析。课程还将涵盖伦理和法律问题，引导学生遵守网络使用规范和隐私保护原则。

**授课目标：**

网络爬虫在当今信息时代具有重要的作用，可以帮助我们从互联网上获取大量数据并应用于各种领域，如数据分析、机器学习、商业竞争情报等。通过学习《Python网络爬虫与信息提取》课程，学生将获得有关网络爬虫和信息提取的实用技能，具备处理实际数据采集与信息提取任务的能力，能够应对实际问题中的数据收集和信息提取任务。这门课程不仅适用于对网络数据感兴趣的学生，还适用于在撰写学术性论文时，需要进行大规模数据采集和处理的情境。

**学习内容及要求:**

1. 网络爬虫基础
2. 爬虫的定义与应用领域
3. HTTP协议和请求/response过程
4. URL解析和构建
5. requests库的基本使用
6. 静态网页爬取与数据提取
7. 静态网页结构解析
8. HTML解析库的使用（如beautifulsoup）
9. 定位和提取所需数据
10. 数据存储和导出
11. 动态网页爬取与数据提取
12. 动态网页的特点与处理方法
13. Selenium库的使用
14. 使用PhantomJS等自动化工具模拟浏览器行为
15. 提取JavaScript生成的数据
16. 反爬虫机制与策略
17. 常见反爬虫技术及对策
18. User-Agent与Cookies的处理
19. IP代理和访问频率控制
20. 验证码处理方法
21. 高级爬虫技术
22. 使用多线程和多进程提高爬取效率
23. 分布式爬虫架构与任务调度
24. 自动化登录与数据交互
25. 大规模数据存储与处理
26. 实践项目
27. 学生将分组完成一个实际的网络爬虫项目
28. 从不同类型的网站上爬取数据
29. 分析和提取有用的信息
30. 数据可视化和展示
31. 伦理与法律问题
32. 网络爬虫的道德和法律问题
33. 隐私保护与数据使用合规性
34. 操作规范和良好的职业道德

# **参考书目：**

[1] 《Python语言程序设计基础(第2版)》，嵩天、礼欣、黄天羽著，高等教育出版社，2017.2（讲授Python 3版本）

[2] 《数据采集技术—Python网络爬虫项目化教》，ISBN：9787040497816，黄锐军，高等教育出版社，2018；

[3] 《Python编程 从入门到实践》第3版，[美] 埃里克·马瑟斯 著，袁国忠 译

# **网络资源：**

# Python网络爬虫程序技术，黄锐军等，国家精品课程（中国大学MOOC）

# 【Python+爬虫】只用260分钟学会Python爬虫（B站）