《Python 网络爬虫程序技术》教学大纲

课程类别： 专业课

适用对象：软件技术专业及相关专业

总 学 时：56　讲授学时：26　课内实践学时：30

独立实践学时：（0）

一、课程的性质、任务与课程的教学目标

（一）课程的性质、任务

1.课程性质

本课程注重对学生职业能力和创新精神、实践能力的培养。本课程旨在对学生的程序设计思想和技能进行，培养学生利用主流 scrapy 框架进行爬虫项目的设计与开发。《Python 网页爬虫程序技术》课程是软件技术专业 Python 方向的专业核心课程，是融理论与实践一体化，教、学、做一体化的专业课程，是基于设计的工作过程系统化学习领域课程，是工学结合课程。

2．课程的任务

本门课程旨在通过学习与实践培养学生的爬虫程序开发能力，为社会输送急需人才；课程对应的网页爬虫开发工程师岗位有着相对较高的薪酬水平和较为广阔的发展前景，可以为参加学习的学生提供良好职业预期发展。

（二）课程的教学目标

1. 基本理论要求：

本课程主要面向岗位为网页爬虫开发工程师，能力辐射岗位有：Web 开发工程师、数据分析师、测试工程师、文档工程师、售前/售后工程师等。

2.基本技能要求：

掌握使用 Python 语言开发应用程序的一般方法，课程应多在计算机教室进行一体化教学。该课程实践性较强，需要学员多作练习，理论与实验比例应为 1：1。 要紧密围绕培养目标，突出重点。

3.职业素质要求：

要求学生在学习过程中认真学习，能严格遵守课题纪律和实验实训规章制度，以培养学生的职业素质.

一、课程教学要求与内容

（一）课程教学要求

本课程内容涵盖了对学生在“基本理论”、“基本技能”和“职业素质”三个层次的培养。以网页爬虫开发岗位必备的开发技能为重点并具备相应的理论基础的同时，注重综合职业素质的养成，课程采用启发诱导式教学，鼓励学生“勤于思考，勤于动手”。

1)基本理论要求：

- 掌握爬虫程序设计理念；
- 掌握数据提取与存储思想
- 掌握 scrapy 爬虫框架设计思想。

2)基本技能要求：

- 熟练掌握 ullib 网页下载方法；
- 熟练掌握正则表达式选取数据的规则；
- 熟练掌握 BeautifulSoup 工具选择数据的方法；
- 熟练掌握 xpath、css 选择数据的方法；
- 熟练掌握 scrapy 网页爬取的工作流程；
- 熟练掌握 scrapy 中 Item、Pipeline 数据的序列化输出方法；
- 熟练掌握 scrapy 中 Spider 的网页递归爬取技术；
- 熟练掌握 scrapy 中中间件的使用方法；

3)职业素质要求：

- 能够完成真实业务逻辑向代码的转化；
- 能够独立分析解决技术问题；
- 自学能力强，能够快速准确地查找参考资料；
- 能够按照规范编写技术文档；
- 沟通能力强，能够与小组其他成员通力合作。

本门着重培养学生独立完成交互式爬虫程序项目的设计、开发以及测试等能力。

（二）课程教学内容

项目 1 爬取外汇网站数据

1.1 外汇网站项目任务
1.2 搭建爬虫程序开发环境
1.2.1 理解爬虫程序
1.2.2 搭建开发环境
1.3 使用 Flask 创建 Web 网站
1.3.1 安装 Flask 框架
1.3.2 创建模拟外汇网站
1.3.3 获取网站 HTML 代码
1.4 使用 GET 方法访问网页
1.4.1 客户端 GET 方式发送数据
1.4.2 服务器获取 GET 发送的数据
1.5 使用 POST 方法访问网页
1.5.1 客户端 POST 发送数据
1.5.2 服务器获取 POST 的数据
1.5.3 混合使用 GET 与 POST
1.6 使用正则表达式匹配数据
1.6.1 使用正则表达式匹配字符串
1.6.2 使用正则表达式爬取数据
1.7 综合项目 爬取模拟外汇网站数据
1.7.1 创建模拟外汇网站
1.7.2 解析网站 HTML 代码
1.7.3 设计存储数据库
1.7.4 编写爬虫程序
1.7.5 执行爬虫程序
1.8 实战项目 爬取实际外汇网站数据
1.8.1 解析网页 HTML 代码
1.8.2 爬取网页外汇数据
1.8.3 设计存储数据库
1.8.4 编写爬虫程序
1.8.5 执行爬虫程序

项目 2 爬取名言网站数据

2.1 名言网站项目任务
2.2 BeautifulSoup 装载 HTML 文档
2.2.1 创建模拟名言网站
2.2.2 安装 BeautifulSoup 库
2.2.3 BeautifulSoup 装载 HTML 文档
2.3 BeautifulSoup 查找 HTML 元素
2.3.1 使用 find 函数查找
2.3.2 查找元素属性与文本

6.4.1 存储数据
6.4.2 读取数据
6.4.3 修改数据
6.4.4 删除数据
6.4.5 扫描数据
6.4.6 删除数据库表
6.5 综合项目 爬取模拟景区网站数据
6.5.1 创建模拟景区网站
6.5.2 编写爬虫程序
6.5.3 执行爬虫程序
6.6 实战项目 爬取旅游景区网站数据
6.6.1 解析网站 HTML 代码
6.6.2 爬取网站景区数据
6.6.3 爬取全部页面的数据
6.6.4 设计存储数据库
6.6.5 编写爬虫程序
6.6.6 执行爬虫程序

三、课程学时分配

学时分配表（以课题或知识单元编排）

序号	模 块 名 称	学 时	其中	
			讲授	实践
1	Web 网站与访问	8	4	4
2	网页数据爬取方法	10	4	6
3	网站数据爬取路径	8	4	4
4	scrapy 框架爬虫程序	12	6	6
5	selenium 爬取动态数据	10	4	6
6	数据爬取与 NoSql 数据库	8	4	4
	总计	56	26	30

注：12 学时作为考核、系统测试、验收、发布、讲评等

五、考核方式及成绩评定

1．基本思路

该课程是面向工作过程的工学结合特点的课程，因此重视学生的知识与能力的培养与训练，采用完成程序结果与过程考核相结合的方法。

2．课程设计的评分标准（共100分）

考核内容	考核细节名称	考核权重	权重
学习情境考核	Web网站与访问	15%	60%
	网页数据爬取方法	20%	
	网站数据爬取路径	20%	
	scrapy框架爬虫程序	20%	
	selenium爬取动态数据	15%	
	数据爬取与NoSql数据库	10%	
学习过程考核	上课出勤	25%	40%
	学习态度	25%	
	合作精神	25%	
	组织协调	25%	
合计			100%

六、必要说明

（一）课程开设的基本条件

本课程需要Python3.6 + vsCode IDE。

（二）建议使用的教材及教学参考书

[教材]
《Python爬虫项目教程》（微课版），黄锐军编，人民邮电出版社，2021年
[参考书]
《数据采集技术-Python网络爬虫项目化教程》，黄锐军编，高等教育出版社，2018年

大纲编订部门： 软件学院　　执笔人： 黄锐军

大纲审订部门：　　审订部门负责人：

编订日期： 2017-09-01　　修订日期： 2022-02-10