**《Python网络爬虫实务》教学大纲**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **课程名称：Python网络爬虫实务** | | | | | | | **课程类别（必修/选修）：选修** | | | | |
| **课程英文名称：** | | | | | | | | | | | |
| **总学时/周学时/学分：**48/3/3 | | | | | | | **其中实验/实践学时：**24/0 | | | | |
| **先修课程：**程序设计、数据结构、线性代数 | | | | | | | | | | | |
| **授课时间：**1-16周 周五 2-4节 | | | | | | | **授课地点：**实验楼 605 机房 | | | | |
| **授课对象：**2018计算机科学与技术(跨境电商)1班 | | | | | | | | | | | |
| **开课学院：**粤台产业科技学院 | | | | | | | | | | | |
| **任课教师姓名/职称：**郭燚 | | | | | | | | | | | |
| **答疑时间、地点与方式：**  1.每次课的课前、课间和课后，采用一对一的问答方式；  2.每次习题课，采用集中讲解方式 | | | | | | | | | | | |
| **课程考核方式：**开卷**（）**闭卷**（）**课程论文**（✔）**其它**（✔）** | | | | | | | | | | | |
| **使用教材：**  Python 3网络爬虫开发实战，崔庆才 著，出版社： 人民邮电出版社ISBN：9787115480347  **教学参考资料：**  自编材料，python爬虫视频教程基础到项目实战开发（网页版）  **优学院授课网站:** | | | | | | | | | | | |
| **课程概述**  **人工智能时代的来临，随着互联网数据越来越开放，越来越丰富。基于大数据来做的事也越来越多。数据分析服务、互联网金融、数据建模、医疗病例分析、自然语言处理、信息聚类，这些都是大数据的应用场景，而大数据的来源都是利用网络爬虫来实现。**  **课程目标**  **（一）总体目标**  **学完本课程，学生能够搭建python开发环境，安装python开发工具，配置python环境，熟练操作pycharm编写python程序，使用python编写网络爬虫程序爬取网络数据。**  **（二）具体目标**  **1、能力目标**  **（1）能熟练使用pycharm开发软件编写python程序； （2）能熟练应用requests库实现HTTP网络请求； （3）能熟练运用正则表达式、xpath、beautifulsoup解析请求响应内容； （4）会创建scrapy爬虫项目；**  **2、知识目标 （1）了解网页结构； （2）理解HTTP请求； （3）掌握正则表达式、xpath、beautifulsoup解析数据。**    **（三）素质目标**  **（1）能够按照指定的任务描述，利用所学知识和技术，分析网站布局方式，确定合适的解决方案，完成指定任务；**  **（2）能够合理利用所需技术和工具，采用科学的方法确定解决方案，设计相应的实验方案，搭建实验环境，开展实验研究，完成实验报告；**  **（3）能够利用图书馆和互联网进行文献检索和资料查询，解决学习和实践中遇到的问题。**  **课程内容和要求**  **1. Python基础语法学习（基础知识）**  **2. 对HTML页面的内容抓取（数据抓取）**  **3. 对HTML页面的数据提取（数据提取）**  **4. Scrapy框架以及scrapy-redis分布式策略（第三方框架）**  **5. 爬虫(Spider)、反爬虫(Anti-Spider)、反反爬虫(Anti-Anti-Spider)**  **课程基本要求是**  **本课程共分为5个部分：基础入门/进阶/运维/实战开发/项目开发，教程内容涉及编程/爬虫/运维开发/项目案例，帮助学生学习和更快理解并能解决实际应用中的问题。**  **本课程适合于任何想学习Python编程的人员，无论是否从事软件类行业，无论是否接触过Python，只要之前有接触过计算机语言，通过学习均可快速掌握Python的基础操作和技巧。** | | | | | | | | | | | |
| **课程教学目标**  **一、知识目标：**  **（1）了解网页结构；**  **（2）理解HTTP请求；**  **（3）掌握正则表达式、xpath、beautifulsoup解析数据。**  **二、能力目标：**  **（1）能熟练使用pycharm开发软件编写python程序；**  **（2）能熟练应用requests库实现HTTP网络请求；**  **（3）能熟练运用正则表达式、xpath、beautifulsoup解析请求响应内容；**  **（4）会创建scrapy爬虫项目；**  **三、素质目标：**  **（1）能够按照指定的任务描述，利用所学知识和技术，分析网站布局方式，确定合适的解决方案，完成指定任务；**  **（2）能够合理利用所需技术和工具，采用科学的方法确定解决方案，设计相应的实验方案，搭建实验环境，开展实验研究，完成实验报告；**  **（3）能够利用图书馆和互联网进行文献检索和资料查询，解决学习和实践中遇到的问题。** | | | | | | | | **本课程与学生核心能力培养之间的关联(授课对象为理工科专业学生的课程填写此栏）：**  通过本专业的学习，学生应具备如下核心能力：  ■核心能力1（交叉知识的运用能力）：具有运用数学、基础科学及计算机科学与技术相关知识的能力；  ■核心能力2（实验与数据解读能力）：具有计算机软件开发与数据搜寻分析解释的能力；  ■核心能力3（技术工具的应用能力）具有计算器软件工程技术应用、数据搜集分析应用跨境电商运营知识与技能、及大数据技术的专业所需的技术、技能和使用软硬件辅助工具的能力；  ■核心能力4（计科与大数据分析专业能力）：具有编程设计能力并能应用计算器与数据分析科技来辅助、及大数据技术分析，促进跨境电商运营的能力；  ■核心能力5（项目管理与团队合作能力）具有项目管理、有效沟通、领域整合与团队合作的能力；  ■核心能力6（解决复杂问题的能力）：具有运用计算机科学与技术理论及应用知识，整合计算机应用技术、数据分析应用及跨境电商运营专业，解决相关问题和进行研发或创新的能力；  ■核心能力7（持续学习与创新超越能力）：具有应对计算器科学与技术快速变迁的能力，培养自我持续学习的习惯与能力，了解所学专业技术对环境、社会及全球的影响，并在学习中敢于创新超越。  □核心能力8（专业伦理、社会责任与国际视野）：具有理解职业道德、工程专业伦理、认知社会责任、国际观以及开拓全球视野及尊重多元观点的能力。 | | | |
| **理论教学进程表** | | | | | | | | | | | |
| **周次** | **教学主题** | | **主讲教师** | **学时数** | **教学的重点、难点、课程思政融入点** | | | **教学模式**  **（线上/混合式/线下** | **教学方法** | | **作业安排** |
| 1 | 网络爬虫概述 | | 郭燚 | 2 | 重点：  网络爬虫的应用背景、市场前景、相关就业方向、应用案例介绍  课程思政融入点：介绍网络爬虫的基本概念，培养实事求是的科学态度和职业道德。 | | | 线上与 优学院根据疫情发展灵活调整 | 讲授与上机练习 | | 教材课后习题  课程思政作业：要求学生每人阅读Python网络爬虫实务有关的文章或书籍 |
| 2 | 爬虫原理与数据抓取 | | 郭燚 | 2 | 重点：  了解爬虫技术原理，HTTP抓包工具  课程思政融入点：  介绍网络爬虫的基本技术，培养实事求是的科学态度和职业道德。 | | | 线上与 优学院根据疫情发展灵活调整 | 讲授与上机练习 | | 教材课后习题  课程思政作业：要求学生每人阅读Python网络爬虫实务有关的文章或书籍 |
| 3 | 非机构化数据与结构化数据提取 | | 郭燚 | 2 | 重点：  正则表达式、XPath、多线程爬虫  难点：  课程思政融入点：介绍正则表达式的编写规则、RE模块使用、了解beautifulSoup4使用方法，培养实事求是的科学态度和职业道德。 | | | 线上与 优学院根据疫情发展灵活调整 | 讲授与上机练习 | | 教材课后习题  课程思政作业：要求学生每人阅读Python网络爬虫实务有关的文章或书籍 |
| 4 | 动态HTML处理与验证码处理 | | 郭燚 | 2 | 重点：了解HTML、PhantomJS、模拟登陆、验证码识别。  课程思政融入点：介绍动态HTML处理，培养实事求是的科学态度和职业道德。 | | | 混合式:线下 与优学院根据疫情发展灵活调整 | 讲授与上机练习 | | 教材课后习题  课程思政作业：要求学生每人阅读Python网络爬虫实务有关的文章或书籍 |
| 5 | Scrapy框架 | | 郭燚 | 2 | 重点：  Scrapy安装、基础操作、运作流程  课程思政融入点：介绍Scrapy基础框架，培养实事求是的科学态度和职业道德。 | | | 混合式:线下 与优学院根据疫情发展灵活调整 | 讲授与上机练习 | | 教材课后习题  课程思政作业：要求学生每人阅读Python网络爬虫实务有关的文章或书籍 |
| 6 | Scrapy实战 | | 郭燚 | 2 | 重点：  Scrapy爬虫实战项目  课程思政融入点：介绍Scrapy实战项目，培养实事求是的科学态度和职业道德。 | | | 混合式:线下 与优学院根据疫情发展灵活调整 | 讲授与上机练习 | | 教材课后习题  课程思政作业：要求学生每人阅读Python网络爬虫实务有关的文章或书籍 |
| 7 | 期中总结、复习与期中考 | | 郭燚 | 2 | 期中总结、复习与期中考 | | | 线上与 优学院根据疫情发展灵活调整 |  | |  |
| 10 | Scrapy分布式组件 | | 郭燚 | 2 | 重点：  Scrapy的分布式组件例如Connection、Dupefilter、Pickleconmpat、管道、队列、调度等  难点：  课程思政融入点：介绍Scrapy的分布式组件，培养实事求是的科学态度和职业道德。 | | | 线上与 优学院根据疫情发展灵活调整 | 讲授与上机练习 | | 教材课后习题  课程思政作业：要求学生每人阅读Python网络爬虫实务有关的文章或书籍 |
| 11 | Scrapy-redis实战 | | 郭燚 | 2 | 重点：  Scrapy -redis实战项目演练  课程思政融入点：介绍分布式爬虫、新闻资讯爬虫，在线数据表爬虫的项目实战案例，培养实事求是的科学态度和职业道德。 | | | 线上与 优学院根据疫情发展灵活调整 | 讲授与上机练习 | | 教材课后习题  课程思政作业：要求学生每人阅读Python网络爬虫实务有关的文章或书籍 |
| 12 | 期末总结与期末考 | | 郭燚 | 2 | 期末总结与期末考 | | |  |  | |  |
| **合计：** | | | | 24 |  | | |  |  | |  |
| **实践教学进程表** | | | | | | | | | | | |
| **周次** | **实验项目名称** | **主讲教授** | | **学时** | **重点、难点、课程思政融入点** | | | **项目类型（验证/综合/设计）** | **教学**  **手段** | | |
| 2-3 | 百科采集案例 | 郭燚 | | 3 |  | | | **综合** | 线上与 优学院根据疫情发展灵活调整 | | |
| 2-3 | 多线程爬虫案例 | 郭燚 | | 3 |  | | | **综合** | 线上与 优学院根据疫情发展灵活调整 | | |
| 3 | Beautifulsoup4爬虫案例 | 郭燚 | | 3 |  | | | **综合** | 线上与 优学院根据疫情发展灵活调整 | | |
| 4 | Tesseract验证码识别案例 | 郭燚 | | 3 |  | | | **综合** | 线上与 优学院根据疫情发展灵活调整 | | |
| 5 | 手机App抓包、阳光热线平台、新浪爬虫、图片下载 | 郭燚 | | 3 |  | | | **综合** | 线上与 优学院根据疫情发展灵活调整 | | |
| 6 | MongoDB、模拟登陆、Fiddler手机抓包 | 郭燚 | | 3 |  | | | **综合** | 线上与 优学院根据疫情发展灵活调整 | | |
| 7 | 新浪新闻分类爬虫 | 郭燚 | | 3 |  | | | **综合** | 线上与 优学院根据疫情发展灵活调整 | | |
| 8 | IT桔子数据爬取 | 郭燚 | | 3 |  | | | **综合** | 线上与 优学院根据疫情发展灵活调整 | | |
| 合计： | | | | 24 |  | | |  |  | | |
| **考核方法及标准** | | | | | | | | | | | |
| **考核形式** | | | | | | **评价标准** | | | | **权重** | |
| 平时成绩 | | | | | | 出勤状况  1. 评价标准：不迟到，请假须有辅导员签字的请假条。  **线上教学**课堂表现  1. 评价标准：参与优学院学习参与成绩,课堂程度及随堂测验。  2. 要求：参与课堂程度高。  平时作业  1. 评价标准：按照作业完成情况评分。  2. 要求：按时作业，作业工整规范。 | | | | 40% | |
| 期中考试 | | | | | | 1. 评价标准：按照试卷参考解答及评分标准给分。  2. 要求：能灵活运用所学算法基础知识和方法进行求解，独立、按时完成考试。 | | | | 30% | |
| 期末考试 | | | | | | 1. 评价标准：按照试卷参考解答及评分标准给分。  2. 要求：能灵活运用所学算法基础知识和方法进行求解，独立、按时完成考试。 | | | | 30% | |
|  | | | | | |  | | | |  | |
|  | | | | | |  | | | |  | |
| **大纲编写时间：**2020**年**8**月**14**日** | | | | | | | | | | | |
| **系（部）审查意见：**  系（部）主任签名： 时维宁  日期：2020年 9 月 7日 | | | | | | | | | | | |