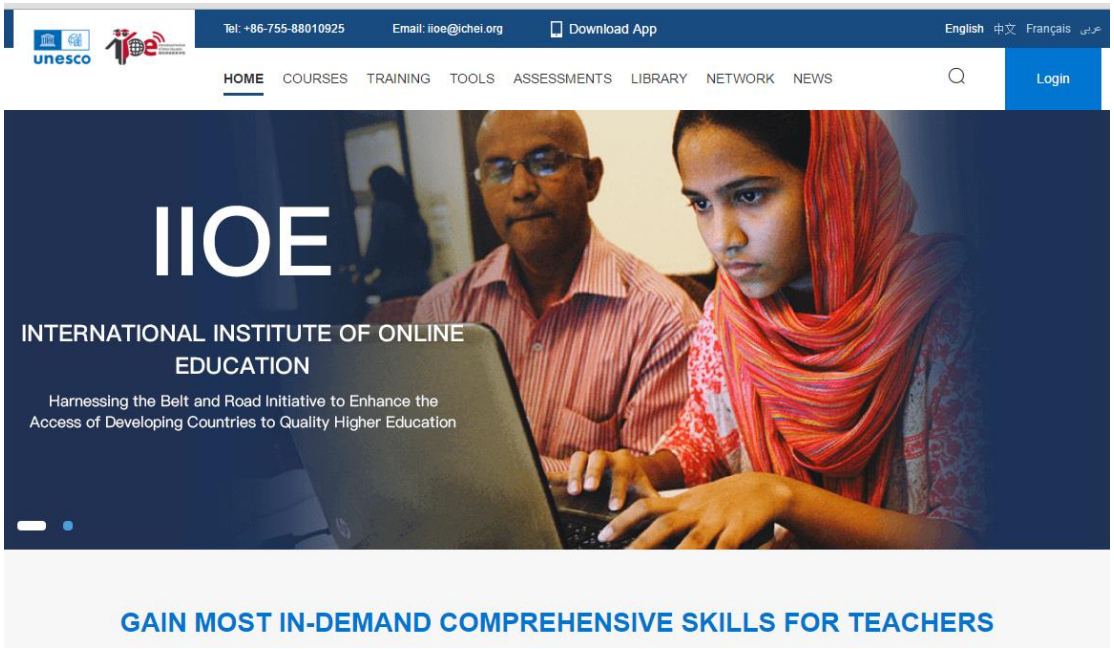
联合国教科文组织 IIOE 国际课程

“一带一路”（Belt and Road）国际课程



Python 网络爬虫程序技术

Python Web Crawler Program Technology

附件 1：国际网络教育学院课程信息表（中英文）

课程名称	**Python 网络爬虫程序技术**	
开课学校	深圳信息职业技术学院	
院系（或部门）	软件学院	
授课语言	英语	
课程时长	10 小时	
建议学分	2	
课程负责人	姓名：黄锐军 职位/职称：副教授	E-mail：r_j_huang@163.com 手机：13543347131
开课学校协调人	姓名： 职位/职称：	E-mail： 手机：
课程所属	☐专业基础课• √专业核心课 ☐其他	

课程所属专业	计算机软件技术
预备知识	Python 语言基础、HTML 与 CSS 基础、JavaScript、数据库基础

课程简介（字数控制在 100~300 以内）

网络爬虫就是一组能自动从网站的相关网页中自动搜索与提取数据的程序，本课程的目标就是学习如何使用 Python 编写高效稳定的爬虫。课程采用项目驱动的形式，分四个项目讲解了 Web 程序架构、正则表达式、BeautifulSoup、CSS 查找与爬取数据、多线程爬取、深度优先与广度优先顺序爬取路径、Selenium 动态网页数据爬取、数据的存储等内容。课程采用线上网络教学与线下现场教学相结合的形式，以项目为依托、循序渐进、由浅入深地进行知识的传授与技能的教学。

教学大纲（章节到 2 级标题）

项目 1 开发和请求网站

1.1 开发 Python Web Spider 应用程序

1.1.1 Python 平台

1.1.2 爬虫程序

1.2 开发和请求网站

1.2.1 开发网站

1.2.2 请求网站

1.3 通过获取和发布方法请求网站

1.3.1 通过 Get 方法请求

1.3.2 通过邮寄方式请求

1.3.3 服务器接收数据

1.4 正则表达式

1.4.1 正则表达式简介

1.4.2 正则表达式规则

1.5 综合训练：下载图像

1.5.1 开发网站

1.5.2 开发 Spider 程序

项目二 爬取网站数据

2.1 BeautifulSoup 和 HTML

2.1.1 BeautifulSoup 安装

2.1.3 BeautifulSoup 加载 HTML

2.2 BeautifulSoup 搜索 HTML 元素

2.2.1 BeautifulSoup 搜索元素

2.2.2 BeautifulSoup 获取属性

2.2.3 BeautifulSoup 获取文本

2.3 BeautifulSoup 高级搜索

2.3.1 属性规则

2.3.2 搜索后代

2.3.3 搜索兄弟姐妹

2.4 综合训练：获取名言名句

2.4.1 解析网页

2.4.2 开发爬虫程序

2.5 培训项目：获取与存储名言名句

2.5.1 解析下一个网页

2.5.2 获取和存储数据

项目 3 递归快速爬取数据

3.1 递归访问 Web 服务器

3.1.1 具有多个页面的 Web 服务器

3.1.2 递归请求页面

3.2 多线程程序

3.2.1 多线程

3.2.2 线程等待

3.3 在多线程中下载图像

3.3.1 单线程下载

3.3.2 多线程下载

3.4 综合训练：获取旅行信息

3.4.1 解析网页

3.4.2 开发爬虫程序

3.5 综合训练：获取和存储旅行信息

3.5.1 解析网页

3.5.2 开发爬虫程序

项目 4 Selenium 与动态网站

4.1 BeautifulSoup 困扰问题

4.1.1 网页使用 JavaScript

4.1.2 BeautifulSoup 爬虫程序

4.2 Selenium 解决方案

4.2.1 Selenium 安装

4.2.3 Selenium 爬虫程序

4.3 综合训练：获取和存储头条新闻

4.3.1 查找元素

4.3.2 解析网页

4.3.3 开发爬虫程序

主讲教师简介（可为多人，每人字数控制在 100-300 以内）

黄锐军，副教授，从事计算机软件技术的研究与教学，在近 30 年的教学中。讲授了大量的计算机程序设计课程，有丰富的教学经验。先后出版了 8 本计算机教材，主持开发与教授 3 门 MOOC 课程，受到广大师生的好评。参加与主持多个深圳市软件项目设计与开发。由于工作表现突出，获得深圳市优秀教师称号。

主要教材(讲义、参考书)名称、主编、出版社、出版日期

1.《Python 程序设计》，黄锐军，高等教育出版社，2018，ISBN:9787040493726
2.《数据采集技术—Python 网络爬虫项目化教程》，黄锐军，高等教育出版社，2018，ISBN:9787040497816

课程预告素材：

1. 课程封面图片（540*300）

2. 教师头像图片（最小 120*120）

课程建设团队负责人签字：

年　月　日

开课单位意见：

负责人签字：

公章：

年　月　日

Course Information

Name	Python Web Crawler Program Technology	
Institute	Shenzhen Institute of Information Technology	
Department	Software College	
Language	English	
Course Duration	10 hours	
Credits Suggested	2	
Lecturer	Name: Huang RuiJun Title: Associate Professor	E-mail：r_j_huang@163.com Mobile: 13543347131
Institute Coordinator	Name: Title:	F-mail: Mobile:
Category	☐Foundation ✓Core ☐Other	
Major	Computer Software Technology	
Prerequisite knowledge	Python Program Foundation、HTML and CSS、JavaScript、Database	

Introduction (100~300 words)

Web crawlers or sipders are a set of programs that can automatically search and extract data from the relevant pages of a website. The goal of this course is to learn how to develop efficient and stable crawlers using Python. The course is divided into four projects to explain web program architecture, regular expressions, BeautifulSoup, CSS searching, multi-threaded crawling, depth-search-first and breadth-search-first crawl paths, and Selenium dynamic web data crawling, and data saving, etc. The course adopts the combination of online and offline teaching processes. It is based on projects, and it gradually teaches you knowledge and skills step by step.

Syllabus (Chapter includes first and second headings)

2.1.3 BeautifulSoup Loading HTML

2.2 BeautifulSoup Searching HTML Element

2.2.1 BeautifulSoup Searching Element

2.2.2 BeautifulSoup Getting Attribute

2.2.3 BeautifulSoup Getting Text

2.3 BeautifulSoup Adavnce Searching

2.3.1 Attribute Rules

2.3.2 Searching Descendants

2.3.3 Searching Siblings

2.4 Training Project: Getting Famous Quotes

2.4.1 Parsing Web Page

2.4.2 Developing Spider Program

2.5 Training Project: Getting and Storing Famous Quotes

2.5.1 Parsing Next Web Page

2.5.2 Getting and Storing Data

Project 3 Getting Data Recursively and Quickly

3.1 Requesting Web Server Recursively

3.1.1 Web Server With Many Pages

3.1.2 Requesting Pages Recursively

3.2 Program of Multiple Threads

3.2.1 Multiple Threads

3.2.2 Thread Waiting

3.3 Downloading Images in Multiple Threads

3.3.1 Downloading in Single Thread

3.3.2 Downloading in Multiple Threads

3.4 Training Project: Getting Travel Information

3.4.1 Parsing Web Page

3.4.2 Developing Spider Program

3.5 Training Project: Getting and Storing Travel Information

3.5.1 Parsing Web Page

3.5.2 Developing Spider Program

Project 4 Selenium and Dynamic Website

4.1 BeautifulSoup Problem

4.1.1 Web Page with JavaScript

4.1.2 BeautifulSoup Spider Program

4.2 Selenium Solution

4.2.1 Selenium Installation

4.2.2 Selenium Spider Program

4.3 Training Project: Downloading Top News

4.3.1 Searching Elements

4.3.2 Parsing Web Page

4.3.3 Developing Spider Program

Lectures (Each lecture includes 100~300 words' introduction)

Huang Ruijun, associate professor, has been engaged in the research and teaching of computer software technology for nearly 30 years. He has taught a lot of computer programming courses and has rich teaching experience. He has published more than 8 computer textbooks, presided over the development and teaching of 3 MOOC courses, and is widely and warmly welcomed by teachers and students. He has participated and presided over the design and development of multiple Shenzhen software projects. Due to his outstanding work and performance, he was awarded the title of Shenzhen Excellent Teacher.

Major textbook or references with editor, press and publication date.

1. "Python Programming", Huang Ruijun, Higher Education Press, 2018, ISBN: 9787040493726
2. "Data Acquisition Technology-Python Web Crawler Project Tutorial", Huang Ruijun, Higher Education Press, 2018, ISBN: 9787040497816

Course preview materials includes:
a) Cover photo (540*300 PPI);

b) Lecturer portrait (120*120 PPI)

Course leader:
Signature: Date:

Institute:
Signature: Date:

Official seal