《解析 Python 网络爬虫：核心技术、Scrapy 框架、分布式爬虫》课程教学大纲

（课程英文名称）

课程编号：201800522062

学　　分：5 学分

学　　时：53 学时　（其中：讲课学时 41　上机学时：12）

先修课程：《Python 快速编程入门》

后续课程：Python 数据分析与数据挖掘

适用专业：信息技术及其计算机相关专业

开课部门：计算机系

一、课程的性质与目标

《解析 Python 网络爬虫：核心技术、Scrapy 框架、分布式爬虫》是面向计算机相关专业的一门专业进阶课，涉及抓取网页数据、数据解析、并发下载、抓取动态网页、图像识别与文字处理、存储爬虫数据、爬虫框架、分布式爬虫。通过本课程的学习，学生能够掌握 Python 爬虫的基础知识，可以独立使用框架开发 Python 爬虫的项目程序。

二、教学条件要求

操作系统：Windows 7 版本及以上

Python 环境：Python 3.6.2

开发工具：PyCharm 2016.3.2

三、课程的主要内容及基本要求

第 1 章　初识爬虫

章名	初识爬虫	学时	1

学习目标	1. 了解爬虫产生的背景 2. 知道什么是爬虫 3. 了解爬虫的用途 4. 熟悉不同维度下网络爬虫的几种类型			
知识点	了解	掌握	重点	难点
爬虫产生背景	√			
什么是爬虫		√	√	
爬虫的用途	√			
爬虫的分类		√	√	√

第 2 章 爬虫的实现原理和技术

章名	爬虫的实现原理和技术	学时	4	
学习目标	1. 掌握通用爬虫和聚焦爬虫的工作原理 2. 熟悉爬虫抓取网页的流程 3. 了解通用爬虫的网页分类， 4. 了解爬虫要遵守的协议，及智能抓取更新网页的文件 5. 熟悉防爬虫的一些应对策略 6. 了解使用 Python 语言做爬虫的优势			
知识点	了解	掌握	重点	难点
通用爬虫工作原理		√		√
聚焦爬虫工作原理			√	√
爬虫抓取网页的详细流程			√	
通用爬虫中网页的分类	√			
robots.txt 文件		√		
Sitemap.xml 文件	√			
防爬虫应对策略		√		√
为什么选择 Python 做爬虫	√			

第 3 章 网页请求原理

章名	网页请求原理	学时	3	
学习目标	1. 熟悉浏览器加载网页的过程 2. 掌握基于 HTTP 协议的请求原理 3. 掌握客户端 HTTP 请求格式 4. 掌握服务端 HTTP 响应格式 5. 熟悉 HTTP 抓包工具 Fiddler 的使用			
知识点	了解	掌握	重点	难点
统一资源定位符 URL	√			

计算机域名系统 DNS	√			
分析浏览器显示完整网页的过程			√	√
客户端 HTTP 请求格式		√	√	√
服务端 HTTP 响应格式		√	√	√
Fiddler 工作原理	√			
Fiddler 下载安装	√			
Fiddler 界面详解		√		
Fiddler 抓取 HTTPS 设置		√		
使用 Fiddler 捕获 Chrome 的会话	√			

第 4 章 抓取网页数据

章名	抓取网页数据		学时	6
学习目标	1. 了解什么是 urllib 库 2. 会使用 urllib 库爬取网页 3. 掌握如何转换 URL 编码，可以使用 GET 和 POST 两种方式实现数据传输 4. 知道伪装浏览器的用途 5. 掌握如何自定义 opener 6. 了解服务器的超时 7. 熟悉一些常见的网络异常 8. 掌握 requests 库的使用			
知识点	了解	掌握	重点	难点
什么是 urllib 库	√			
快速爬取一个网页	√			
分析 urlopen 方法		√		
使用 HTTPResponse 对象		√		
构造 Request 对象		√		
URL 编码转换		√		
处理 GET 请求		√	√	
处理 POST 请求		√	√	√
添加特定 Headers		√	√	
简单的自定义 opener		√		√
设置代理服务器		√		√
超时设置		√		
URLError 异常和捕获	√			
HttpError 异常和捕获	√			
什么是 requests 库	√			
使用 requests 发送请求		√	√	
返回 Response 响应		√	√	

第 5 章 数据解析

章名	数据解析	学时	8
学习目标	1. 了解服务器返回的数据格式 2. 会通过浏览器查看网页的结构 3. 熟悉解析数据的几种技术 4. 掌握正则表达式的使用，会使用 re 模块解析网页数据 5. 掌握 XPath 语法的使用，会使用 lxml 库解析网页数据 6. 掌握 BeautifulSoup 的使用，会使用 bs4 库解析网页数据 7. 掌握 JSONPath 语法的使用，会使用 json 模块解析网页数据		

知识点	了解	掌握	重点	难点
网页数据格式	√			
查看网页结构	√			
数据解析技术		√		
正则表达式		√	√	
什么是 XPath	√			
XPath 语法		√		√
XPath 开发工具	√			
什么是 lxml 库		√	√	
lxml 库的基本使用		√	√	√
什么是 Beautiful Soup	√			
构建 BeautifulSoup 对象		√		
通过操作方法进行解读搜索		√	√	√
通过 CSS 选择器进行搜索		√		√
什么是 JSON	√			
JSON 与 XML 语言比较	√			
json 模块介绍	√			
json 模块基本使用		√	√	√
JSONPath 介绍	√			
JSONPath 语法对比		√		

第 6 章 并发下载

章名	并发下载	学时	3
学习目标	1. 了解多线程爬虫的流程 2. 掌握 queue 模块的使用，可以利用它实现多线程爬虫 3. 熟悉协程的使用，能够用协程技术实现并发爬虫		

知识点	了解	掌握	重点	难点
多线程爬虫流程分析	√		√	
queue（队列）模块简介		√		
Queue 类简介		√	√	
协程爬虫的流程分析	√		√	
第三方库 gevent		√		√

第 7 章 抓取动态内容

章名	抓取动态内容	学时	4
学习目标	1. 知道什么是动态网页 2. 掌握抓取动态网页的 selenium 和 PhantomJS 技术，学会安装和配置它们 3. 掌握 selenium 和 PhantomJS 的基本使用		

知识点	了解	掌握	重点	难点
动态网页介绍	√			
selenium 和 PhantomJS 概述		√		
selenium 和 PhantomJS 安装配置		√	√	
入门操作		√	√	√
定位 UI 元素		√	√	
鼠标动作链		√		
填充表单		√		√
弹窗处理		√		
页面切换		√		
页面前进和后退		√		
获取页面 Cookies		√		
页面等待		√		√

第 8 章 图像识别与文字处理

章名	图像识别与文字处理	学时	4
学习目标	1. 了解什么是 OCR 技术 2. 会安装 Tesseract 工具 3. 熟悉 PIL 和 pytesseract 库 4. 知道什么是文字规范的图像，能够利用 pytesseract 识别和处理字符 5. 了解验证码的分类，能够利用 pytesseract 识别简单的图形验证码		

知识点	了解	掌握	重点	难点
OCR 技术简介	√			
Tesseract 引擎的下载和安装		√		

pytesseract 库简介		√		
PIL 库简介		√		
处理图像中格式规范的文字		√	√	
对图片进行阈值过滤和降噪处理		√	√	√
识别图像的中文字符		√	√	√
验证码分类	√			
简单识别图形验证码		√		

第 9 章 存储爬虫数据

章名	存储爬虫数据	学时	4	
学习目标	1. 熟悉数据存储的几种方式 2. 了解什么是 MongoDB 数据库 3. 会在 Windows 平台安装和配置 MongoDB 数据库 4. 掌握 PyMongo 库的使用			
知识点	了解	掌握	重点	难点
数据存储简介	√			
什么是 MongoDB	√			
Windows 平台安装 MongoDB 数据库		√		
比较 MongoDB 和 MySQL 的术语		√	√	
什么是 PyMongo	√			
PyMongo 的基本操作		√	√	√

第 10 章 初识爬虫框架 Scrapy

章名	初识爬虫框架 Scrapy	学时	3	
学习目标	1. 了解常见的爬虫框架 2. 掌握 Scrapy 框架的架构 3. 熟悉 Scrapy 框架的运作流程 4. 学会在不同的平台上安装 Scrapy 框架 5. 掌握 Scrapy 框架的基本操作			
知识点	了解	掌握	重点	难点
常见爬虫框架介绍	√			
Scrapy 框架的架构		√		√
Scrapy 框架的运作流程		√		√
安装 Scrapy 框架		√	√	
新建一个 Scrapy 项目		√	√	
添加 Item 实体数据		√	√	
制作 Spiders 爬取网页		√	√	√
永久性存储数据	√			

第 11 章 Scrapy 终端与核心组件

章名	Scrapy 终端与核心组件		学时	4
学习目标	1. 会启动和使用 Scrapy 框架自带的 shell 2. 掌握 Spiders 组件，能够更深一步认识并使用这个组件 3. 掌握 Item Pipeline 组件，会自定义管道来处理数据 4. 掌握 Downloader Middlewares 组件，可以通过随机 IP 和随机用户代理应对反爬虫行为 5. 掌握 Settings 组件，能够明确和定制各个 Scrapy 组件的行为			
知识点	了解	掌握	重点	难点
启用 Scrapy shell	√			
使用 Scrapy shell		√	√	
Spiders—抓取和提取结构化数据		√	√	√
自定义 Item Pipeline		√	√	√
Downloader Middlewares—防止反爬虫		√		√
Settings—定制 Scrapy 组件		√	√	

第 12 章 自动抓取网页的爬虫 CrawlSpider

章名	自动抓取网页的爬虫 CrawlSpider		学时	3
学习目标	1. 明确 CrawlSpider 爬虫类的用途，可以创建使用 CrawlSpider 模板的爬虫 2. 掌握 CrawlSpider 类的原理 3. 掌握 Rule 类的使用，能够运用该类制定爬虫的爬取规则 4. 掌握 LinkExtractor 类的使用，能够提取需要跟踪爬取的链接			
知识点	了解	掌握	重点	难点
初识爬虫类 CrawlSpider	√			
CrawlSpider 类的工作原理		√	√	√
通过 Rule 类决定爬取规则		√	√	√
通过 LinkExtractor 类提取链接		√	√	√

第 13 章 Scrapy-Redis 分布式爬虫

章名	Scrapy-Redis 分布式爬虫		学时	6
学习目标	1. 了解什么是 Scrapy-Redis，明确 Scrapy 和 Scrapy-Redis 的关系 2. 熟悉 Scrapy-Redis 的架构和运作流程 3. 掌握 Scrapy-Redis 的主要组件 4. 会在不同的平台上独立搭建 Scrapy-Redis 开发环境 5. 理解分布式采用的策略，可以测试服务器端和爬虫端是否能远程连接 **6.** 掌握 Scrapy-Redis 的基本使用，可以在 Scrapy 项目的基础上实现分布式爬取			
知识点	了解	掌握	重点	难点
Scrapy-Redis 简介	√			
Scrapy-Redis 的完整架构		√		√
Scrapy-Redis 的运作流程		√		√
Scrapy-Redis 的主要组件		√		√
安装 Scrapy-Redis		√		
安装和启动 Redis 数据库		√		
修改配置文件 redis.conf	√			
分布式策略		√	√	
测试 Slave 端远程连接 Master 端		√		
设置 Scrapy-Redis 组件		√	√	
制作 Spider 爬取网页		√	√	√
执行分布式爬虫		√	√	√
使用多个管道存储		√	√	√
处理 Redis 数据库里的数据	√			√

四、学时分配

章目	讲课	上机	合计
第 1 章　初识爬虫	1 学时	0 学时	1 学时
第 2 章　爬虫的实现原理和技术	3 学时	1 学时	4 学时
第 3 章　网页请求原理	2 学时	1 学时	3 学时
第 4 章　抓取网页数据	5 学时	1 学时	6 学时
第 5 章　数据解析	7 学时	1 学时	8 学时
第 6 章　并发下载	2 学时	1 学时	3 学时
第 7 章　抓取动态内容	3 学时	1 学时	4 学时
第 8 章　图像识别与文字处理	3 学时	1 学时	4 学时

第 9 章　存储爬虫数据	3 学时	1 学时	4 学时
第 10 章　初识爬虫框架 Scrapy	2 学时	1 学时	3 学时
第 11 章　Scrapy 框架的 shell 和组件	3 学时	1 学时	4 学时
第 12 章　自动抓取网页的爬虫 CrawlSpider	2 学时	1 学时	3 学时
第 13 章　Scrapy-Redis 分布式爬虫	5 学时	1 学时	6 学时
合计	41 学时	12 学时	53 学时

五、考核模式与成绩评定办法

本课程为考试课程，期末考试采用百分制的闭卷考试模式。学生的考试成绩由平时成绩（30%）和期末考试（70%）组成，其中，平时成绩包括出勤（5%）、作业（5%）、上机成绩（20%）。

六、选用教材和主要参考书

本大纲是根据教材《Python 快速编程入门》所设计的。

参考书籍：

传智播客.《Python 快速编程入门》人民邮电出版社.201709

七、大纲说明

本课程的授课模式为：课堂授课+上机，其中，课堂主要采用多媒体的方式进行授课，并且会通过测试题阶段测试学生的掌握程度；上机主要是编写程序，要求学生动手完成指定的程序设计或验证。

撰写人：　　审定人：

批准人：　　执行时间：