湖南石油化工职业技术学院

学 期 授 课 计 划

（ 二〇二〇至二〇二一年第二学期 ）

课程名称 网络爬虫

授课班级 大数据 31901/31902/31903

授课教师 孙 检

审 批 签 字

教研室主任		
二级学院院长		

学期授课计划编制说明

课程标准名称、批准单位及时间	《网络爬虫》 湖南石油化工职业技术学院、2020年7月
教学内容(授课内容起止章节)	全面介绍了数据采集、数据存储、动态网站爬取、App爬取、验证码破解、模拟登录、代理使用、爬虫框架、分布式爬取等知识
教材名称、编者及出版单位	《Python3网络爬虫开发实战》、崔庆才编著、人民邮电出版社

课　时　分　配

本课程总时数	72	已讲授时数	0	尚需讲授时数	72

计划授课周、时数	本学期教学总周数	本学期实习周数	本学期理论教学周数	本学期理论教学周课时	本学期计划课时分配					
					新课讲授	实践(实验)	练习(复习)	考试(测验)	机动	其它
	18	0	18	72	34	32	4	2	0	
实际完成周、时数	18	0	18	72	34	32	4	2	0	

学生知识现状的调查与分析	本学期是大三的上学期本学期所授科目为网络爬虫，学生在大一接触了 Java 程序设计课程所以对软件编程有一定的了解但是也只是处在刚接触的阶段有一些熟悉，大二接触到 Python 语言。 设置本课程的目的是：使学习者在全面了解 Python 技术、网络爬虫的基础上，系统掌握 urllib、requests、正则表达式、Beautiful Soup、XPath、pyquery、数据存储、Ajax 数据爬取等内容，接着通过多个案例介绍了不同场景下如何实现数据爬取，最后掌握 pyspider 框架、Scrapy 框架和分布式爬虫。完成本课程的学习后能够熟练地综合应用 Python 技术和网络爬虫的思想编写程序解决现实生活中的问题，最终提高程序设计水平和计算机应用能力，从而能胜任企业软件研发以及科研院所的研发、教学任务。
本学期教学的主要任务和要求	一、主要任务： 要求学生掌握爬虫概述、Ullib 实现网站下载、使用正则表达式获取网页数据、使用 beautifulsoup 工具选择数据、使用 scrapy 编写网页爬虫程序、使用 item、pipeline 实现数据序列化与存储等基础知识。学生首先了解网络爬虫的特点、发展及推荐学习方法，然后学习使 用 scrapy 实现网页递归爬取、第三方库相关知识等。同时掌握不同领域的网络爬虫技术，并能够解决实际问题。 二、要求： 鉴于该班学生的实际情况，力争实现以下目标： 1．及格率：85%左右； 2．优秀率：15%左右。

教材的重点和难点	重点： （1）爬虫程序设计理念 （2）数据提取与存储思想 （3）scrapy 爬虫框架设计思想 （4）ullib 网页下载方法 （5）正则表达式选取数据的规则 （6）beautifulsoup 工具选取数据的方法 （7）xpath、css 选择数据的方法 （8）scrapy 网页爬取的工作流程 （9）scrapy 中 item 、pipeline 数据的序列化输出方法 （10）scrapy 中 spider 的网页递归爬取技术 难点： （1）xpath、css 选择数据的方法 （2）scrapy 网页爬取的工作流程 （3）scrapy 中 item 、pipeline 数据的序列化输出方法 （4）scrapy 中 spider 的网页递归爬取技术

本学期提高教学质量的措施	1. 与院系计算机相关专业教师多相互学习、与辅导员多交流讨论不断改进教学措施。 2. 从学生的年龄特点出发，多采取游戏式的教学，引导学生乐于参与教学学习活动。 3. 在课堂教学中，注意多提一些有利于孩子理解的问题，而不是一味的求难、求广。应该考虑学生实际的思维水平，多照顾中等生以及思维偏慢的学生。提出的问题要有启发性，不愤不懈，不启不发，要让学生有迫切探究的欲望，而且有能力探究才恰到好处。 4. 布置一些比较有趣的作业，比如动手的作业，少一些呆板的练习。作业批改可以现批现改，及时纠正，让其真正理解掌握。对课后作业错误率高或难度大的写在黑板上，让全班同学共同思考、交流，让不会的学生在聆听中受到启发，相互学习，相互补充，悟出道理来。 5. 加强家庭教育与学校教育的联系，适当教给家长一些正确的指导孩子学习的方法。 6. 要认真学习新课程标准，勇于创新，坚持备课做到“六认真”。教学做到 “课课清”、“人人清”、“本本清”、“科科清”的“四清”教学目标。

学期授课计划进度计划表

累计课时	次/周	授课章节与时数		主要内容与教材分析	实践内容	作业内容或题号	备　注
		章节名称	时数				
4	1/1 2/1	第一章	4	开发环境配置			
8	1/2 2/2	第二章	4	爬虫基础			
12	1/3 2/3	第二章	4	HTTP基本原理			
16	1/4 2/4	第二章	4	爬虫的基本原理			
18	1/5	第三章	2	基本库的使用			
20	2/5	第三章	2	正则表达式			
24	1/6 2/6	开发篇	4	综合案例-爬取猫眼电影排行			
26	1/7	第四章	2	使用XPath			
28	2/7	第四章	2	使用Beautiful Soup			
32	1/8 2/8	第四章	4	使用pyquery			
36	1/9 2/9	第五章	4	数据存储			
38	1/10	第六章	2	Ajax数据爬取			
40	2/10	第六章	2	Ajax数据提取			
42	1/11	第九章	2	Selenium的使用			

44	2/11	第九章	2	Splash的使用			
46	1/12	第八章	2	验证码的识别			
50	2/12 1/13	第八章	4	图形验证码的识别			
52	2/13	第八章	2	极验滑动验证码的识别			
56	1/14 2/14	开发篇	4	综合案例二-使用代理爬取微信公众号文章			
58	1/15	第十三章	2	模拟登录并爬取GitHub			
60	2/15	第十三章	2	App的爬取			
64	1/16 2/16	第十三章	4	pyspider框架的使用			
68	1/17 2/17	第十七章	4	分布式爬虫			
70	1/18		2	复习			
72	2/18		2	考试			