



长沙职业技术学院
CHANGSHA VOCATIONAL & TECHNICAL COLLEGE

《数据导入与预处理》课程标准

教 学 单 位 : 长沙职业技术学院

执 笔 人 : 凌敏

教 研 室 主 任 审 核 : 何政山

分 院 (教 学 部) 负 责 人 审 核 : 桑子华

教 务 处 审 核 : 王聪

教 学 副 校 长 审 定 : 罗慧玲

制 订 日 期 : (2020.7.30)

修 订 日 期 :

长沙职业技术学院教务处 编制

二〇一九年〇七月

【课程信息】

表 1 课程信息表

课程名称	数据导入与预处理	开课院部	经济贸易与信息技术学院	
课程代码	113200339	考核性质		考试 <input checked="" type="checkbox"/> 考查 <input type="checkbox"/>
总学时	72	课程类型	公共基础课	是 <input type="checkbox"/>
			专业基础课	是 <input type="checkbox"/>
			专业核心课	是 <input checked="" type="checkbox"/>
			专业方向课	是 <input type="checkbox"/>
			专业拓展课	是 <input type="checkbox"/>
			其它	
适用专业	大数据技术与应用			

表 2 课程标准开发团队名单

序号	姓名	工作单位	职称/职务
1	凌敏	长沙职业技术学院	讲师
2	吴细花	长沙职业技术学院	副教授
3	殷正坤	长沙职业技术学院	副教授
4	何政山	长沙职业技术学院	副教授
.....			

目 录

一、课程概述.....	- 1 -
二、课程定位.....	- 1 -
三、教学目标.....	- 1 -
四、课程设计.....	- 2 -
五、 课程内容与要求.....	- 2 -
六、教学建议.....	- 5 -
七、教学评价与考核.....	- 8 -
八、附录.....	错误！未定义书签。

《数据导入与预处理》课程标准

一、课程概述

（一）课程性质

《数据导入与预处理》是一门专业核心课，指导学生进行网络数据、日志数据获取，并利用相关大数据与处理技术和工具，根据业务需要进行数据的清洗与转换。

（二）与其它课程的联系

本课程的先修课程为《Linux 操作系统》、《网页美化与布局》、《Python 程序设计》，为后续课程《大数据分析技术》《大数据运维》等课程的学习提供了专业知识和技能基础，也为《毕业综合实践》、《顶岗实习》、《毕业设计》等打下良好的基础，更为职业生涯发展奠定基础。

二、课程定位

本课程是大数据技术与应用专业的专业核心课程，主要目标是培养大数据技术与应用专业学生的数据采集、数据预处理职业能力、具有大数据意识职业素养和创新能力。

通过本课程的学习，使学生具备数据采集和预处理平台搭建、网络爬虫、日志数据采集和数据预处理的职业能力和职业素养，能胜任数据采集技术员等岗位工作。本课程以岗位技能的项目化以及大数据导入与预处理任务的序列化为载体，通过与长沙中电软件园建立密切合作关系，将企业的教育资源融入到教学体系中，确保学生学习到最先进和实用的数据导入与预处理的新技术。学完本课程后，学生可以参加大数据分析与应用职业技能等级证书认证考试，为将来走向工作岗位奠定坚实的基础。

三、教学目标

（一）素质目标

1. 具有工作责任心和服务意识。

2. 具有安全工作的意识。
3. 形成诚实、守信、善于沟通、合作和富有爱心的思想品质。
4. 具备创新、创意的思维和团队协作精神

（二）知识目标

1. 掌握数据采集和预处理平台搭建。
2. 掌握网络爬虫技术。
3. 掌握日志数据采集和数据预处理。
4. 掌握 scrapy、Flume、pig、kettle、Pandas、openrefine 和 urllib、selenium 基本库和 BeautifulSoup 解析库等的相关技术知识。
5. 熟悉数据导入与预处理的常用与典型操作。

（三）能力目标

1. 能利用所学的相关技术，能根据企业具体大数据分析业务，结合采集的海量数据，针对性的进行数据的预处理。
2. 会数据导入与预处理的常用与典型操作。

四、课程设计

本课程遵循应用型本科和高等职业教育规律，以大数据技术与实际应用工作岗位需求为导向选取课程内容，完成了数据采集和预处理平台搭建、网络爬虫实践、日志数据采集实践和数据预处理实践等完整的数据导入与预处理应用案例，课程目标是培养学生具备“大数据分析”应用项目所需数据导入与预处理的综合职业能力；坚持开放性设计原则，吸收企业专家参与，构建以“工作任务”为载体的“项目化”课程结构；课程教学实施教、学、做一体，坚持理论为实践服务的教学原则，通过模拟企业大数据导入与预处理应用业务进行组织，锻炼学生的实践操作能力。

五、课程内容与要求

表 1 课程内容与进程安排一览表

序号	项目（模块）	任务（单元）	教学内容			目标要求	学时安排	
			知识点	技能点	思想素质		讲授	实践

1 1	数据导入与预处理概述	认识数据采集技术，熟悉数据采集平台	熟悉数据采集的来源和采集的方法。	学会数据采集平台的搭建。	大数据意识	熟悉数据采集的来源和采集的方法。	2	2
		认识数据预处理技术	了解原始数据存在的主要问题；熟悉数据预处理的作用和工作任务。	学会数据处理工具的搭建。	工作责任心和服务意识、安全工作意识。	了解原始数据存在的主要问题；熟悉数据预处理的作用和工作任务。	2	2
2	网络爬虫	使用urllib爬取北京公交线路信息	熟悉urllib基本库和BeautifulSoup解析库的使用方法	使用urllib基本库和BeautifulSoup解析库，学会北京公交线路相关信息的爬取。	工作责任心和服务意识、安全工作意识。	熟悉urllib基本库和BeautifulSoup解析库的使用方法	2	2
		使用selenium爬取淘宝网站信息	熟悉Selenium基本库和pyquery解析库的使用方法	使用Selenium基本库和pyquery解析库，学会淘宝网站信息的爬取。	工作责任心和服务意识、编程意识。	熟悉Selenium基本库和pyquery解析库的使用方法	2	2
		使用scrapy爬取北京公交信息	熟悉Scrapy的使用	学会使用Scrapy爬取北京公交信息	工作责任心和服务意识、安全工作意识。	熟悉Scrapy的使用	2	2
		创新与拓展	熟悉urllib基本库和BeautifulSoup解析库的使用方法	使用urllib基本库和BeautifulSoup解析库，学会起点中文网中相关信息的爬取	工作责任心和服务意识、安全工作意识。	熟悉urllib基本库和BeautifulSoup解析库的使用方法	2	2

3	日志数据采集	Flume 的安装和配置	熟悉 Flume 的相关基础知识	学会 Flume 的安装和不同应用场景下的配置	工作责任心和服务意识、安全工作意识。	熟悉 Flume 的相关基础知识	2	2
		Flume 采集数据上传到集群	熟悉 Flume 的相关基础知识	学会将采集的日志数据转存到 HDFS 的方法；学会将采集的日志数据转存到 HBase 的方法	工作责任心和服务意识、安全工作意识。	熟悉 Flume 的相关基础知识	2	2
		创新与拓展	熟悉 Flume 日志信息的采集过程；熟悉 Nginx 的安装和基础语法。	学会使用 Flume 搭建日志采集系统，并学会日志采集；学会使用 Nginx 和 Flume 搭建日志采集系统，并完成日志采集；学会使用 Kettle 采集 Excel 表格中的数据；学会使用 Kettle 进行多复合来源数据的采集与关联	工作责任心和服务意识、安全工作意识	熟悉 Flume 日志信息的采集过程；熟悉 Nginx 的安装和基础语法。	2	2
4	数据预处理	用 Pig 进行数据预处理	熟悉 Pig 的相关基础知识	使用 Pig 完成“北京公交线路信息”数据的预处理	工作责任心和服务意识、安全工作意识	熟悉 Pig 的相关基础知识	2	4

用kettle进行数据预处理	熟悉 Kettle 的相关基础知识	使用 Kettle 完成“北京公交线路信息”数据的预处理	工作责任心和服务意识、安全工作意识	熟悉 Kettle 的相关基础知识	2	4
用Pandas进行数据预处理	熟悉 Pandas 的相关基础知识	使用 Pandas 完成“北京公交线路信息”数据的预处理	工作责任心和服务意识、安全工作意识	熟悉 Pandas 的相关基础知识	2	4
用openrefine进行数据预处理	熟悉 OpenRefine 的相关基础知识	使用 OpenRefine 完成“北京公交线路信息”数据的预处理	工作责任心和服务意识、安全工作意识	熟悉 OpenRefine 的相关基础知识	2	4
使用 Flume Interceptor 对日志信息进行数据预处理	熟悉 Flume Interceptor 的相关基础知识	使用 Flume Interceptor 完成“北京公交线路信息”数据的预处理	工作责任心和服务意识、安全工作意识	熟悉 Flume Interceptor 的相关基础知识	2	4
创新与拓展	熟悉大数据预处理工具 Kettle 的应用	学会 Flume+Nginx+Pig 数据采集预处理应用	工作责任心和服务意识、安全工作意识	熟悉大数据预处理工具 Kettle 的应用	2	4

六、教学建议

(一) 授课教师基本要求

担任本课程教学任务的专任教师要求具有具有高校教师资格；有理想信念、有道德情操、有扎实学识、有仁爱之心；具有数据科学与大数据、云计算理论与应用、智能科学与技术、计算机科学与技术、网络工程、软件工程等相关专业本科及以上学历；具有扎实的本专业相关理论功底和实践能力；具有较强信息化教学能力，能够开展课程教学改革和科学研究；有每 5 年累计不少于 6 个月的企业实践经历。

担任本课程教学任务的兼职教师从本专业相关的行业企业聘任，具备良好的思想政治素质、职业道德和工匠精神，具有扎实的专业知识和丰富的实际工作经验，具有中级及以上相关专业职称。

(二) 实践教学条件

1. 校内实训室

表 2 大数据分析与应用实训室

实训室名称	大数据分析与应用实训室	面积要求	90m ²
序号	核心设备	数量要求	备注
1	SHARPXG-MB55XA 投影机	1 台	
2	计算机服务器	1 台	
3	存储设备	1 台	
4	交换机	4 台	
5	教师机电脑	1 台	
6	学生机电脑	56 台	
7	教学管理软件	1 台	
8	大数据实训平台	1 套	

2. 校外实习基地

表 3 大数据专业校外实习基地

序号	校外实习基地名称	合作企业名称	用途	合作深度要求
1	华为数据挖掘实习基地	华为技术有限公司	认识实习、跟岗实习、顶岗实习	紧密合作
2	中电软件园实习基地	长沙中电软件园	认识实习、跟岗实习、顶岗实习	深度合作
3	自兴人工智能实习基地	自兴人工智能	认识实习、跟岗实习	紧密合作
4	强智科技实习基地	湖南强智科技发展有限公司	认识实习、跟岗实习、顶岗实习	深度合作
5	拓维信息实习基地	长沙拓维信息技术有限公司	认识实习、跟岗实习	深度合作
7	北京易观智库网络科技有限公司长沙研发基地实习基地	北京易观智库网络科技有限公司长沙研发基地	认识实习、跟岗实习、顶岗实习	一般合作

8	湖南艾瓦特网络科技有限公司实习基地	湖南艾瓦特网络科技有限公司	认识实习、跟岗实习、顶岗实习	一般合作
---	-------------------	---------------	----------------	------

3. 教学方法与策略

(1) 教学方法

项目引领、任务驱动：每个项目都是多个任务的结合体。每个项目都按规范化的流程进行组织，通过案例引导：将大数据企业中大数据技术与应用岗位相关的项目引入课堂，演绎为工作中的各种操作任务案例，培养学生利用各种数据导入与预处理的工具与手段，结合大数据具体应用业务，进行大数据导入与预处理操作。

混合式教学法：通过超星学习通课前发布任务清单，学生进行自学，课中教师进行导学，课后进行督学，课中，以学生为主体，教师为主导，增加课堂活力，提高学生学习的自觉性和主动性；激发学生的课堂活力和创新思维能力；培养学生团队合作能力、获取知识的能力。

问题探究法：课堂上引导学生思考项目需要分为几部分完成。用到哪些知识点？通过学生讨论，解决在预习中存在的疑惑问题，然后老师讲解知识的重难点和项目的操作步骤，根据讨论的方法，要求学生们协作完成项目的制作。

(2) 教学策略

利用超星网络教学平台，课前发布学习任务清单，让学生了解每次课的学习任务，通过完成预习检测，疑问讨论，了解学生预习情况，调整教学内容，课中进行以实训操作为主，教师进行答疑、知识串讲和随堂指导，课后提交作业、完成项目拓展、项目实践，课后测试，项目评价等。

4. 参考教材选用

表 4 数据导入与预处理课程参考教材选用表

序号	教材名称	教材类型	出版社	主编	出版日期
1	数据导入与预处理	高职高专	人民邮电出版社	米洪、张鸽	2019年11月
2	大数据预处理技术	高职高专	人民邮电出版社	朱晓姝、许桂秋	2019年4月

5. 课程资源的利用与开发

充分参与调研，确保教学项目来源于实际应用，并围绕项目建设教学大纲、建设、配备与大数技术与应用专业有关的音视频素材、教学课件、数字化教学案例库、虚拟仿真软件、数字教材等专业教学资源库，种类丰富、形式多样、使用便捷、动态更新、满足教学，现已建有智慧商务专业群资源库，职教云《大数据技术与应用》专业教学资源库，长沙职业技术学院智慧网络教学平台。

七、教学评价与考核

本课程为考试课程，突出学生解决实际问题的能力，加强过程性考核，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（60%）和期末考试（40%）组成，其中平时成绩包括课堂表现、作业、项目完成情况、出勤，在线学习情况组成，具体如下表 5 所示：

表 5 课程考核评价表

考核项目	考核点		建议考核方式	权重 (%)
教学单元	教学内 容	知识点	预习检测	5
		技能点	参与主题讨论	5
		思想素质	参与主题讨论	5
	课堂表现		课堂互动	5
	作业		师评	10
	项目完成情况		师评、互评	10
日常行为表现	考勤		线上平台统计	5
	在线学习情况		线上平台统计	15
期末考试			机试	40