**《数据采集与挖掘》课程实验教学大纲**

一、课程基本信息

课程代码：21020053

课程名称：《数据采集与挖掘》

英文名称:《Data Acquisition and Mining) 》

实验总学时：48

适用专业：会计学、财务管理、审计学

课程类别：专业选修课

先修课程：《Python基础》《初级财务会计学》、《会计学》、《数据库管理系统》、《管理信息系统》、《财务管理》、《成本会计》等基础课程

1. 实验教学的**性质、目的与任务**要求
2. 实验教学的性质和目的

大数据时代已经全面开启，带来了信息技术发展的巨大变革，并深刻影响着社会生产和人民生活的方方面面。了解大数据概念、具备大数据思维，是新时代对人才的新要求。本课程通过案例引入并探讨大数据的实质，内容深入浅出，通俗易懂，适合财会专业相关学生们学习。本课程阐述了培养具有数据素养的技术技能型人才所需要的大数据相关基础知识。在确定知识布局时，秉持的一个基本原则是，紧紧围绕大数据通识教育核心理念，培养学生的数据意识、数据思维、数据安全和数据分析能力。本实验教学的内容主要包括大数据概述，大数据思维，数据库基础知识，数据采集与挖掘相关应用，数据可视化概论，大数据安全，大数据与云计算、物联网、人工智能、数字货币、区块链等新一代信息技术的融合应用。为了避免陷入空洞的理论介绍，本实验教学在很多章节都融入了丰富的案例，这些案例就发生在我们生活的大数据时代，很具有代表性和说服力，能够让学生直观感受相应理论的具体内涵。

1. 实验教学的任务要求

1、对学生的要求

高校课程建设要进一步打破学科、专业壁垒，推动文科专业之间的深度融通，以及文科与理工农医科的交叉融合，融入现代信息技术，赋能文科教育，实现文科教育的自我革新。《数据采集与挖掘》是培养德、智、体、美、劳全面发展，具有良好的政治素质和道德修养，掌握扎实的基础理论和专业知识、良好的团队意识和协作能力，打破局限的学科认知和专业限制，提升自身综合素养，提升新文科的通识传承和知识创新能力。具备包括数据思维在内的科学思维能力、解决财会等行业数据密集型问题为主的动手能力以及大数据分析或大数据应用开发技能，在财会等相关行业数据分析系统的开发、部署与应用等方面具有就业竞争力的高素质复合型应用人才。针对学生已经具备《管理经济学》、《经济法》、《财务管理》等专业知识和能力的特点，《数据采集与挖掘》试图采用理论分析与案例讲解相结合的方法，使学生能够全面深入地理解和掌握大数据技术的发展与演进过程，通过大数据应用的场景介绍，熟悉大数据技术应用的商业价值，介绍大数据常用编程软件，了解大数据技术开源平台和工具，拓展大数据应用能力。本课程，侧重实践操作动手能力培养的一门课程。因此要求学生在深刻理解会计信息系统基本理论、基本方法和基本管理制度的基础上，侧重于具体软件的操作和应用。课程内容分为数据采集与挖掘基本理论和Phython软件应用两部分。

2、对教师的要求

教师要持续推动课程教育教学内容更新迭代，培养学生的跨领域知识融通能力和实践能力。针对学生已经具备《管理经济学》、《经济法》、《财务管理》等专业知识和能力的特点，《数据采集与挖掘》试图采用理论分析与案例讲解相结合的方法，使学生能够全面深入地理解和掌握大数据技术的发展与演进过程，通过大数据应用的场景介绍，熟悉大数据技术应用的商业价值，介绍大数据常用编程软件，了解大数据技术开源平台和工具，拓展大数据应用能力。

3、培养目标

本课程旨在实现以下几个培养目标：

（1）引导学生步入大数据时代，积极投身大数据的变革浪潮之中；

（2）了解大数据概念，培养大数据思维，养成数据安全意识；

（3）掌握数据采集的基本流程、方法和基本技术；

（4）熟悉使用Python语言进行数据采集，能够独立设计、开发、测试和维护爬虫程序等；

（5）养成善于动脑、勤于思考的学习习惯；养成规范的软件代码编写习惯；具有良好的沟通能力和团队协作精神等。

4、对实验条件的相关要求

实验课程对计算机等教学环节的软硬件条件要求如下。

#### **硬件要求：**

##### (1)计算机，P4-2.4，8G内存，1TG硬盘；

##### (2)自主上机机房管理信息系统；开放机房；

##### (3)网络环境，每个小组局域网互通，使用教学系统管理软件上课；

##### (4)网络硬盘FTP学生个人空间，方便学生存取个人账套；

##### (5)资源环境，专题实验教学资源网站和实验教学辅助系统。

#### **软件要求：**

#### Python 3.8 (或更高版本)

#### Internet Explorer 6.0 + SP1 及更高版本

#### IIS 5.0 及更高版本

#### NET Framework 2.0 Service Pack 1

##### Dockers 和 Spark

1. 实验教学内容
2. 教学方法

本课程以课程理论教学为主，并安排课堂讨论，以深化学生对知识的理解。在理论教学层面，高屋建瓴地探讨大数据，超脱技术讲解技术，内容深入浅出，通俗易懂，适合非计算机专业学生学习；同时，在课堂上为学生展示丰富的实际应用案例，激发学生学习兴趣，开拓学生视野，培养学生大数据思维。

1. 教学内容

本课程的理论课程相关教学内容如下表所示：

|  |  |  |
| --- | --- | --- |
| **项目（或章）** | **主要内容** | **学时安排** |
| 第1章 大数据概论与python数据分析 | 数据的概念、大数据时代到来的背景、大数据的发展历程、世界各国的大数据发展战略、大数据的概念与影响、大数据的应用以及大数据产业； Python的安装； numpy、pandas库的学习 | 9 |
| 第1次讨论课 | numpy库、pandas库在财务上的应用 | 1 |
| 第2章 大数据思维；爬虫、可视化初体验 | 物联网的概念和应用、云计算的概念和应用、大数据和云计算以及物联网的相互关系；爬虫学习，爬取京东商品评论并可视化；爬取巨潮资讯网公司相关信息并可视化 | 9 |
| 第2次讨论课  | 信息获取后的加工处理；小组展示 | 1 |
| 第3章 爬虫进阶 | 大数据分析全流程所涉及的各种技术；爬取巨潮资讯网上市公司的股市行情pyecharts库使用——制作股票价格折线图 | 9 |
| 第3次讨论课 | 小组展示 | 1 |
| 第4章 可视化进阶 | 大数据在各行各业的主要应用；其他数据网站的爬取； pyecharts库进阶使用——制作3D图 | 8 |
| 第4次讨论课 | 小组展示 | 1 |
| 第5章 实战训练 | 大数据安全问题、数据保护原则、数据安全和 隐私保护对策；例子：商协会政策查询服务平台；chatGPT在数据采集与挖掘上的应用 | 8 |
| 第5次讨论课 | 小组展示 | 1 |
| 合计 | 48 |

本课程的实验教学部分设计了部分实验项目，具体如下：

实验项目一

实验名称：Anaconda和Python安装与配置

实验内容：安装Anaconda和Python

实验性质：验证性实验

实验学时：3学时

实验目的与要求：Hadoop单机和集群模式的配置 或者 jupyterhub集群

实验条件：计算机，P4-2.4，8G内存，1TG硬盘；自主上机机房管理信息系统

研究与思考：（1）安装Anaconda和Python；（2）安装系统的安装包； （3）课程的思政内容：立足学科分类和课程评价体系改革，系统推进各学科体系课程思政的科学化、专业化路径建设，为课程思政改革奠定学科和课程基础。能为师生提供远程高效计算存储平台。学生可以在线做练习写代码，不需要硬件，网页即可支持R、java和Python语言，能省不少硬件成本 。

实验项目二

实验名称：应用python爬取某政府部门规范性文件

实验内容：在Hadoop平台对某政府的部门规范性文件进行爬取

实验性质：验证性实验

实验学时：3学时

实验目的与要求：熟悉Python 的NumPy、SciPy、Pandas和Matplotlib各种程序库,使其在数据分析领域的广泛应用。能安装爬虫所需环境，创建简单的爬虫项目。成功完成爬虫核心实现。

实验条件：广东财经大学大数据Hadoop平台/同心楼实验室

研究与思考：（1）对Python 的NumPy、SciPy、Pandas和Matplotlib各种数据库进行熟悉；（2）创建简单的爬虫代码；（3）运用Python对政府文件进行爬取；（4）**课程的思政内容：通过“某政府部门设立小金库的案例”，引导同学牢记初心使命、树立政治能力过硬、敢于担当作为、勇于纠错防弊的正确价值观，切实做到“清正廉洁”。python爬取某政府部门规范性文件，要学生们学习党和国家的政策，鼓励学生学好专业知识。**

实验项目三

实验名称：爬取豆瓣和猫眼电影影评及可视化分析

实验内容：用Python对豆瓣或者猫眼网站中某部电影的影评进行爬取，并对数据进行可视化分析

实验性质：验证性实验

实验学时：3学时

实验目的与要求：熟悉Python 的NumPy、SciPy、Pandas和Matplotlib各种程序库,使其在数据分析领域的广泛应用。能安装爬虫所需环境，创建简单的爬虫项目，成功完成爬虫核心实现，再对爬取出来的影评进行可视化。

实验条件：广东财经大学大数据Hadoop平台/同心楼实验室

研究与思考：（1）主要是基于豆瓣电影的数据，进行分析，所以首先要爬取相关的电影数据；（2）ETL预处理；（3）数据分析；（4）可视化；（5）**课程的思政内容：爬取数据时应该要注意网络上很多信息难辨真假，是否涉及商业机密或个人隐私，要遵纪守法，不滥用所学的技术。课程从现代影视企业（例如：博纳、阿里巴巴影视等）的发展出发，结合实证研究的成果和实际案例，探讨企业财务会计信息如何影响资本市场的有效运作以及在公司治理中扮演的角色，帮助学生认识当前财务会计报告的环境。授课过程中结合课程教学内容，提炼出“公正、诚信、责任感”三个关键的思政元素，并融入具体的案例中。**

实验项目四

实验名称：股票动态数据获取及可视化分析

实验内容：用Python对同花顺网站中某支股票的数据进行爬取，并对数据进行可视化分析

实验性质：验证性实验

实验学时：3学时

实验目的与要求：熟悉Python 的NumPy、SciPy、Pandas和Matplotlib各种程序库,使其在数据分析领域的广泛应用。能安装爬虫所需环境，创建简单的爬虫项目，成功完成爬虫核心实现，再对爬取出来的股票动态进行可视化。

实验条件：广东财经大学大数据Hadoop平台/同心楼实验室

研究与思考：（1）主要是基于股票的动态进行分析，所以首先要爬取相关的股票数据；（2）ETL预处理；（3）数据分析；（4）可视化；（5）**课程的思政内容：用Python对有关金融和经济数据收集后进行整理和分析，动态数据呈现，为学校重点学科服务，满足应用型大学人才培养。通过剖析“控制权转移中的内幕交易监管案例”，深入分析了内幕交易如何违反“公开、公平、公正”原则而对投资者和资本市场造成不良影响，使学生树立人格要正、自律要严的价值思想；通过引用“云南绿大地财务舞弊案例”，分析财务舞弊对投资者和资本市场造成的严重负面影响，告诫学生诚信从业不仅是每一位会计人必须承担的社会责任，更是会计行业赖以生存和发展的基础。**

实验项目五

实验名称：爬取京东商品评论数据并可视化

实验内容：爬取京东商品评论数据并可视化

实验性质：验证性实验

实验学时：3学时

实验目的与要求：要求学生能够熟悉Python 的NumPy、SciPy、Pandas和Matplotlib各种程序库,使其在数据分析领域的广泛应用。能安装爬虫所需环境，能够熟练的运用pyecharts库将数据可视化

实验条件：广东财经大学大数据Hadoop平台/同心楼实验室

研究与思考：（1）主要是基于京东的json数据，进行分析，所以首先要找到存储相关json数据的网址；（2）ETL预处理；（3）数据分析；（4）可视化；（5）**课程的思政内容：爬取数据时应该要注意网络上很多信息难辨真假，是否涉及商业机密或个人隐私，要遵纪守法，不滥用所学的技术。课程从电商平台（例如：京东、淘宝等）的发展出发，结合实证研究的成果和实际案例，探讨企业财务会计信息如何影响资本市场的有效运作以及在公司治理中扮演的角色，帮助学生认识当前财务会计报告的环境。授课过程中结合课程教学内容，提炼出“公正、诚信、责任感”三个关键的思政元素，并融入具体的案例中。**

实验项目六

实验名称：商协会政策数据爬取及可视化分析

实验内容：商协会政策数据爬取及可视化分析

实验性质：验证性实验

实验学时：3学时

实验目的与要求：要求学生能够熟悉Python 的NumPy、SciPy、Pandas和Matplotlib各种程序库,使其在数据分析领域的广泛应用。能安装爬虫所需环境，能够熟练的运用python把爬取商协会政策数据

实验条件：广东财经大学大数据Hadoop平台/同心楼实验室

研究与思考：（1）熟悉Python的操作；（2）安装所需环境；（3）创建爬取商协会政策数据的代码；（4）课程的思政内容：学习和工作中对学生都是非常重要的，引导学生树立正确的职业目标，培养良好的职业素养和职业操守，努力学习，成为对祖国有用的人。

四、考核方式与标准

##### 最终成绩由平时作业成绩、平时测验成绩、期末成绩和小论文成绩等组合而成。各部分所占比例如下：

##### 平时考勤成绩：10%。出勤以及课堂情况，主要考核对每堂课知识点的理解和掌握程度。

##### 实验报告成绩：50%。主要考核数据处理分析能力。

##### 平时作业成绩：40%。主要考核对核心知识的理解程度，以小组作业或个人作业为考核依据。

##### 期末实验报告或者期末考试成绩：50%。主要考核应用Python进行数据爬取，学会简单的机器学习方法进行数据分析，懂得简单的数据库技术，最后分析结果进行可视化。期末考核采用统一命题，分散式考核，最终提交实验报告的形式。本大纲各部分所规定基本要求、知识点及知识点下的知识细目，都属于考核的内容。考试命题覆盖到各部分，并适当突出重点部分，加大重点内容的覆盖密度。实验报告中所涉及不同难度试题的分数比例适当。

**五、推荐实验教材和教学参考书**

实验教材：

1.《Python应用实战：爬虫、文本分析与可视化》 张丽，电子工业出版社，2020-04

2.《大数据技术与应用教程》，李联宁，清华大学出版社（\*2019年）

3.《大数据技术与应用》，娄岩，清华大学出版社（\*2019年）

参考书：

1. 《Python爬虫大数据采集与挖掘》，曾剑平，清华大学出版社，2020.3第1版

2. Github和CSDN论坛

3. 汪刚：《财会与商业大数据可视化智能分析》，清华大学出版社，2019

4. 杜尔森.德伦：《大数据掘金：挖掘商业世界中的数据价值》，中国人民大学出版社，2016

**六、其他需说明的**

#### 在本课程学习中，应从“了解”、“识记”、“领会”、“简单应用”、“综合应用”五个能力层次去把握：1.了解：要求概念的基本掌握，是最基本要求。2.识记：要求能够识别和记忆本课程有关知识点的主要内容，并能够做出正确的表达、选择和判断。3.领会：在识记的基础上，要求能够领悟和理解本课程中有关知识点的内涵与外延，熟悉其内容要点和它们之间的区别与联系。并能够根据考核的不同要求，做出正确的解释、说明和论述。4.简单应用：在领会的基础上，要求能够运用本课程中少量知识点，分析和解决一般的应用问题。5.综合应用：在简单应用的基础上，要求能够运用本课程中多个知识点，综合分析和解决复杂的应用问题。

大纲修订人：孔荫莹 修订日期:2023年2月19日

大纲审定者： 审定日期: