泰迪科技数据采集项目实训

Python 汽车用户消费投诉数据爬取

培训解决方案

地址：广州市经济技术开发区开泰大道 36 号 1 栋 212

网址：http://www.tipdm.com

邮箱：services@tipdm.com

邮编：510663

联系人：

电话：

目录

1 项目介绍

1.1 项目背景

“Python 汽车用户消费投诉数据爬取”是泰迪科技专门为高校在校学员设计的一套实训项目。本项目采用了 Python 爬虫技术，主要使用 request，selenium 等库爬取数据。

本项目首先通过 Chrome 浏览器登录数据来源页，使用 Chrome 的检查功能，多次刷新，先分析要爬取的数据的规律，搜索品牌、单号、诉求问题、诉求时间、经销商等字段，然后爬取所需页面的信息进行解析，将数据存储到数据库或者其它形式的文件。使用 pandas，re 等库对数据进行清洗，去除无用的标点符号等。

通过相对较短的时间内完成一个爬虫实战中最基础最常见的任务，向学生展示数据爬取流程，使学生对数据爬虫在现实中的应用有一个整体的了解和掌握。通过本项目的学习，学员不仅能够掌握主流的爬虫方法，同时还为从事数据分析相关工作累积了方法、流程和经验。

本项目采用“技术顾问”+“项目经理”+“学员”的团队组织模式，以完全企业化的方式与学生进行交流。学员在项目进行中，能熟悉企业的工作环境，在规定时间内完成项目需求、提升专业技术、锻炼团队协作能力与沟通能力。

1、爬取结果

本项目爬取数据结果保存为 CSV 文件。

2、主流技术

本项目主要涉及技术为 selenium 库爬虫技术，运用其中的模拟登陆及数据获取，pandas 库对数据的收集整理及存储。

1.2 项目目标

本项目通过 Python 网络爬虫获取汽车消费网数据。

1.3 项目数据

http://tousu.315che.com/tousudetail/104189/

数据量大于 4 万

1.4 项目周期及时间安排

注：以下周期仅为参考，工作室可在选择该项目后自行选择项目开始和结束时间，尽量控制在规定周期内完成，以确保学生可以达到训练目的。

（1）项目周期：2 周

（2）时间安排

任务安排	内容	提交场所	提交时间
第一周	任务 1 明确项目需求与目标	邮箱	Day1 22:00
	任务 2 环境准备：numpy、pandas、requests、selenium		
	任务 3 使用 Selenium 打开网页,使用 XPath 解析网页（作业代码）（10%）		
	任务 4 获取品牌、单号、诉求问题、诉求时间、经销商等信息（作业代码）（40%）		Day5 22:00
第二周	任务 5 循环获取（作业代码）（50%）		Day1 22:00
	任务 6 数据存储（作业代码）（60%）		

	任务 7 数据清洗（作业代码）（80%）		Day3 22:00
	任务 8 完成项目报告（作业）（100%）		Day5 22:00
在线答疑		QQ 群	

1.5 项目难度

★★

2 项目任务

任务 1 明确项目需求与目标

任务 2 使用 Selenium 打开网页（作业代码）

任务 3 使用 XPath 解析网页（作业代码）

任务 4 获取品牌、单号、诉求问题、诉求时间、经销商等信息（作业代码）

任务 5 循环获取（作业代码）

任务 6 数据存储（作业代码及存储文件）

任务 7 数据清洗（作业代码）

任务 8 完成项目报告（作业）

3 项目流程

(1) 明确项目目标：阅读项目任务书，明确项目交付内容。

(2) 学习前置知识：提供项目所需知识的云课堂课程，快速掌握项目前置知识。

(3) 项目实践：获取项目需求，动手做项目。

(4) 项目验收：对项目成果进行验收，指导改进项目成果。

(5) 信息入库：参训学生信息录入泰迪人才库，为优秀结业生引荐合作公司。

4 项目核心

- Python 数据分析
- Python 爬虫

5 实现工具

Python、request、numpy、pandas、selenium。

6 实训对象

数学、统计学、计算机等相关专业学生。

7 前置知识

第 1 模块：Python 编程基础

第 2 模块：Python 数据分析与应用

第 3 模块：Python 爬虫

8 项目提交规范

8.1 项目交付内容要求

（1） 请在“项目名称（模板）.docx”基础上进行报告的编写。

（2） 请在“项目名称（模板）.pptx”基础上进行 PPT 的编写。

（3） 具体的内容以及格式要求请查看课程【工作室项目交付要求讲解】。

注：请按照课程【工作室项目交付要求讲解】中提及的要求进行文档的编写，若不按要求进行规范编写，提交的项目文件将被返回修改。

8.2 项目交付命名规范

（1） 项目交付文件包括中间数据（如有）、代码、报告、PPT。

（2） 中间数据需放入“01.中间数据（如有）”文件夹中，各中间数据文件命名为“任务 x-x”，如“任务 1-1”。

（3） 代码需放入“02.代码”文件夹中，各代码文件命名为“任务几 任务名称”，如“任务 1 预处理航空客户数据”。

（4） 报告需放入“03.报告”文件夹中，报告命名为“项目名称”，如“航空公司客户价值分析”。

（5） 报告需放入“04.PPT”文件夹中，PPT 命名为“项目名称”，如“航空公司客户价值分析”。

（6） 最终把所有文件夹放入总文件夹中，总文件夹命名为“项目名称--院校工作室名称--小组成员名称（用顿号隔开）--上交日期（如 2021.1.23）”，如“航空公司客户价值分析--重理工智能工作室--刘备、关羽、张飞--2021.1.23”。

（7） 最后仅需把（6）提及的总文件夹的压缩包提交至邮箱 taididingyuehao@tipdm.com。

注：具体也可查看课程【工作室项目交付要求讲解】中的项目交付模板。同时，请严格按照以上方式进行上缴，如上缴方式不正确，可能影响评级。

8.3 项目提交邮件格式

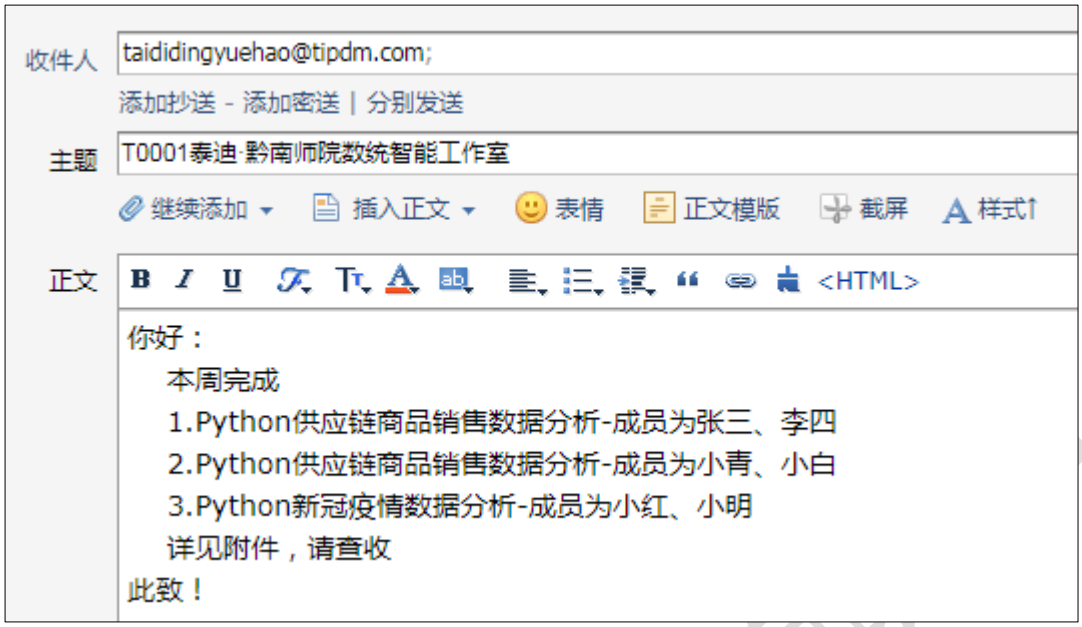
（1） 邮件收件人为：taididingyuehao@tipdm.com

（2） 邮件主题名称为：院校工作室名称

（3） 邮件正文内容为：项目名称-小组成员名称

（4） 邮件附件为：“8.2 项目交付命名规范”提及的项目总文件夹压缩包

项目提交邮件格式具体参考如下。



9 实训对应的就业岗位

9.1 就业岗位

数据分析师、爬虫工程师。

9.2 岗位分析

序号	岗位	主要业务工作	所需技能	相应课程设置
1	数据分析师	数据处理、分析建模、撰写分析报告	Python 数据分析 分析报告	Python 语言设计 Python 数据分析与应用 数据库
2	爬虫工程师	数据爬取、解析、清洗、入库	Python Xpath pandas	Python 编程基础 Python 网络爬虫实战 数据库

10 项目收获

- 足不出校门即可获得实战技能。
- 了解数据分析岗位目前的就业形式和前景，了解需要掌握的技能。
- 掌握一定的挖掘技能和工具，体验一个实际项目的全过程。
- 参训学生信息录入泰迪人才库，为优秀结业生引荐合作公司。

11 项目评价及积分

（1）公司将按照工作室对该项目的完成情况给予评分，项目评分由低到高依次为 D、C、B、B+、A、A+，B 为及格，B 以下不及格。

（2）企业将根据工作室学员的项目评分给予工作室对应的积分，积分达到一定水平，可以获取更多泰迪资源。

（3）可获得的资源包括但不限于

A.由泰迪科技颁发的 CBDA 证书

B.由泰迪科技颁发的实习证明

C.云课堂课程资源

（4）评分对应积分如下

项目评分	积分
项目评分为 B	+50
项目评分为 B+	+60
项目评分为 A	+70
项目评分为 A+	+80

（5）工作室制度详见附件一。

12 附件一 工作室管理制度

查看网页：https://kdocs.cn/l/ckdaaSEuhp0y

13 附件二 前置课程课表

Python 编程基础	Python 数据分析与应用
1 准备工作	1 Python 数据分析概述
1.1 认识 Python	1.1 认识数据分析
1.2 搭建 Python 环境	1.2 熟悉 Python 数据分析的工具
1.3 安装 PyCharm	1.3 安装 Python3 的 Anaconda 发行版
1.4 PyCharm 使用入门	1.4 掌握 Jupyter Notebook 常用功能
2 列表操作	2 NumPy 数值计算基础
2.1 第一个 Python 程序	2.1 认识 NumPy 数组对象 ndarray
2.2 Python 固定数据类型介绍	2.2 认识 NumPy 矩阵与通用函数
2.3 列表构建及索引操作	2.3 利用 NumPy 进行统计分析
2.4 列表元素的增删改查操作	3 Matplotlib 数据可视化基础
2.5 列表推导式	3.1 了解绘图基础语法与常用参数
3 程序流程控制语句	3.2 分析特征间的关系
3.1 Python 常用操作符	3.3 分析特征内部数据分布与分散状况
3.2 Python 条件判定语句	4 Pandas 统计分析基础
4 字符串操作	4.1 读写不同数据源的数据
4.1 字符串及其索引&切片	4.2 掌握 DataFrame 的常用操作
4.2 字符串的常见方法	4.3 转换与处理时间序列数据
4.3 字典的创建及索引	4.4 使用分组聚合进行组内计算
4.4 字典常用操作	4.5 创建透视表与交叉表
4.5 字典推导式	5 使用 Pandas 进行数据预处理
5 Python 文件读取操作	5.1 合并数据
5.1 Python 读取文件	5.2 清洗数据
6 函数	5.3 标准化数据
6.1 Python 函数自定义	5.4 转换数据
7 面向对象与模块	6 使用 scikit-learn 构建模型
7.1 Python 方法与函数对比介绍	6.1 使用 sklearn 转换器处理数据
7.2 Python 面向对象示例	6.2 构建并评价聚类模型

7.3 Python 模块使用 7.4 第三方库的安装与调用 8 注意事项 8.1 Python 工作路径说明 8.2 模块命名及存放路径的注意事项 8.3 结语	6.3 构建并评价分类模型 6.4 构建并评价回归模型

Python 网络爬虫实战	
1 Python 爬虫环境与爬虫简介 1.1 Python 网络爬虫实战介绍 1.2 认识爬虫 1.3 认识反爬虫 1.4 Python 爬虫环境 2 网页前端基础 2.1 概述 2.2 HTTP 请求方法与过程 2.3 常见 HTTP 状态码 2.4 HTTP 头部信息 2.5 认识 cookies 2.6 小结 3 简单静态网页爬取 3.1 静态网页爬取概述 3.2 使用 urllib3 实现 HTTP 请求 3.3 使用 requests 库实现 HTTP 请求 3.4 谷歌开发者工具介绍 3.5 正则表达式介绍 3.6 使用正则表达式获取网页标题信息 3.7 使用 XPath 进行网页解析 3.8 使用 BeautifulSoup 进行网页解析 3.9 数据存储 3.10 小结 4 常规动态网页爬取	

4.1 常规动态网页爬取概述 4.2 逆向分析爬取动态网页 4.3 使用 Selenium 打开浏览对象 4.4 Selenium 页面等待 4.5 使用 Selenium 获取图书信息 4.6 小结 5 模拟登录 5.1 模拟登录概述 5.2 查找表单数据入口及提交数据 5.3 验证码人工处理与代理 IP 5.4 使用 POST 请求方法登录 5.5 使用浏览器 cookies 登录 5.6 基于表单登录的 cookies 登录 5.7 小结 6 终端协议分析 6.1 终端协议分析概述 6.2 了解 HTTP Analyzer 工具 6.3 爬取千千音乐 PC 客户端数据 6.4 小结	

工作室邀请函

为了适应大数据与人工智能及发展的需求，顺应教育部提倡的深化校企合作的号召，更好的服务于广大数据分析爱好者，广东泰迪智能科技股份有限公司诚邀各高校相关专业老师、相关协会学会、俱乐部等组织合作成立“泰迪·智能工作室”，工作室以独立的模式运营，并以学生为中心成立，受泰迪科技监督且由其免费提供各种工作室所需资源的创新型数据智能工作室。

工作室旨在通过教育与产业之间的联动，实行“引进来，走出去”模式，引导学生学习数据科学与人工智能方法为导向，通过与企业的联系、合作、实践，激发学生的数据分析思维，全面推进数据分析与人工智能发展，提高大学生的数据分析素质，激发学生的创新创业精神，以实现创新型数据智能创业人才为培养目标。