2023年度“楚怡杯”湖南省职业院校技能竞赛

赛项规程

一、赛项名称

1. 赛项名称：Python程序开发
2. 赛项组别：高职高专组
3. 赛项归属：电子信息类

二、竞赛内容

Python程序开发赛项以企业真实项目为基础，采用市场主流软件开发架构和实际操作形式进行现场编程设计。竞赛基于Django框架搭建，采用“网络爬虫”、“数据清洗”、“数据分析与可视化”、“机器学习”4个模块。主要涉及知识和技能如下：

序号	模块名称	工作任务
1	网络爬虫	理解业务需求，编写代码实现功能
		通过requests或urllib等爬虫库发送请求
		通过Xpath或BeautifulSoup等从响应内容中解析数据
		通过Python文件操作或Pandas库等方式将爬取的数据存储到文件或数据库中
2	数据清洗	理解业务需求，编写代码实现功能
		通过Django ORM或Pandas读取脏数据
		对数据进行全方位探查分析并梳理清洗思路
		使用NumPy、Pandas等库对数据进行清洗
		通过Django ORM或Pandas将清洗后的数据进行存储
		使用DBeaver工具管理数据库
3	数据分析与可视化	对业务需求及功能的理解，规划业务流程，并编码实现功能
		使用NumPy、Pandas等数据分析工具对数据从各维度进行分析
		使用Matplotlib、PyEcharts等可视化工具将分析结果转化为合适的图表，并通过Django框架渲染出来
		使用DBeaver工具管理数据库

4	机器学习	根据业务需求编写代码实现功能
		通过 Django ORM 或 Pandas 读取数据并从数据中进行特征提取
		采取合适的手段对数据进行预处理
		基于 sklearn 库提供的接口选取算法模型、训练模型并对模型进行调优
		使用模型进行预测，并通过 Django 框架将预测结果渲染到页面上
		使用 DBeaver 工具管理数据库

三、竞赛方式

个人赛。

四、竞赛时量

总时量：240 分钟。

五、名次确定方法

以竞赛总成绩从高到低排序确定名次，不设并列名次。

成绩相同的，按照竞赛项目题型的顺序对选手得分的高低进行排序，如果第一个题型能得到排序则按此确定名次，否则对后续题型得分依次进行排序，只要能区分名次就终止后续题型的判断。如果以上都相同，按照代码行数从少到多排序（以行数少为优，空行不计入行数）。

六、评分标准

1. 评分标准

竞赛满分为 100 分。竞赛项目内容如下：

网络爬虫（15%）：爬虫应用重点考核参赛选手对开源海量信息获取能力、爬虫相关知识体系、网页结构分析及数据存储的熟练程度与编程能力。

数据清洗（30%）：数据清洗模块重点考核参赛选手处理脏数据、使用 NumPy、Pandas 等库对数据进行清洗并进行存储的能力。

数据分析与可视化（35%）：数据分析模块重点考核参赛选手对应用 Python 高级特性和数据分析软件包进行简单与复杂科学计算机数据分析以及可视化的能力，以及考察选手基于实际应用选择分析的能力。

机器学习（20%）：机器学习模块重点考核参赛选手能够使用 Python 工具读取并提取数据特征，并基于机器学习库调用、训练模型然后做出预测等。

2. 评分细则

考核环节	工作任务	分值	评分细则
网络爬虫（15 分）	理解业务需求，编写代码实现功能	2	结果评分 （1）运行结果完全达标(业务逻辑、发送请求、解析数据、数据存储)：100%
	通过 Requests 或 urllib 等爬虫库发送请求	5	

	通过 Xpath 或 BeautifulSoup 等库从响应内容中解析数据	5	（2）运行结果部分达标：按实现结果占总要求的百分比给分。 （3）未实现：0%
	通过 Python 文件操作或 Pandas 库等方式将爬取的数据存储到文件或数据库中	3	
数据清洗（30 分）	理解业务需求，编写代码实现功能	5	结果评分 （1）运行结果完全达标(业务逻辑、数据读取、数据清洗和处理、数据存储)：100% （2）运行结果部分达标：按实现结果占总要求的百分比给分。 （3）未实现：0%
	通过 Django ORM 或 Pandas 读取脏数据	5	
	对数据进行全方位探查分析并梳理清洗思路	5	
	使用 NumPy、Pandas 等库对数据进行清洗	5	
	通过 Django ORM 或 Pandas 将清洗后的数据进行存储	5	
	使用 DBeaver 工具管理数据库	5	
数据分析与可视化（35 分）	对业务需求及功能的理解，规划业务流程，并编码实现功能	3	结果评分 （1）运行结果完全达标(业务逻辑，数据统计分析、统计图绘制、Django 视图、Django 模板、Django ORM、Django 路由)：100% （2）运行结果部分达标：按实现结果占总要求的百分比给分。 （3）未实现：0%
	使用 NumPy、Pandas 等数据分析工具对数据从各维度进行分析	15	
	使用 Matplotlib、PyEcharts 等可视化工具将分析结果转化为合适的图表，并通过 Django 框架渲染出来	15	
	使用 DBeaver 工具管理数据库	2	
机器学习（20 分）	根据业务需求编写代码实现功能	2	结果评分 （1）运行结果完全达标(业务逻辑，数据读取、特征提取、数据预处理、模型选择、模型训练、模型调优，Django 视图、Django 模板、Django ORM、Django 路由)：100% （2）运行结果部分达标：按实现结果占总要求的百分比给分。 （3）未实现：0%
	通过 Django ORM 读取数据并从数据中进行特征提取	2	
	采取合适的手段对数据进行预处理	5	
	基于 sklearn 库提供的接口选取算法模型、训练模型并对模型进行调优	8	
	使用模型进行预测，并通过 Django 框架将预测结果渲染到页面上	2	
	使用 DBeaver 工具管理数据库	1	

七、赛项相关设施设备技术参数

1. 竞赛设备

选手以个人为单位参赛，每人需要 1 台 PC 机（用于运行服务端与客户端开发）。

裁判区域：供裁判休息及工作场地。配电脑，A4 激光打印机 1 台，桌椅，饮水机，纸杯，文具用品等。

2. 硬件环境及配置

类别	部件	参数
选手工位客户端	CPU	Intel 10 代 i5 及以上
	内存	16G 及以上
	硬盘	固态硬盘 256GB 及以上
	显示器	19 寸及以上
选手工位网络	\	200Mbps 及以上
U 盘或移动硬盘	\	64GB 及以上

3. 软件环境

类别	名称	版本号	备注
竞赛平台	Python 程序开发平台	V2.0	\
选手工位	PyCharm Community Edition	Version 2021 or upper	\
	Python	Version 3.7.x	\
	Chrome	Version 90.x or upper	\
	MySQL	Version 5.7 or upper	\
	Ms Office	Version 2016 or upper	\
	django	Version 3.2.x	\
	PyMySQL	Version 1.0.x	\
	BeautifulSoup	Version 4.11.x	\
	requests	Version 2.26.x	\
	lxml	Version 4.6.x	\
	pyecharts	Version 1.9.x	\

	Matplotlib	Version 3.4.x	\
	Numpy	Version 1.19.x	\
	Pandas	Version 1.3.x	\
	Redis 数据库	Version 5.0.x	\
	redis	Version 3.5.x	\
	django_redis	Version 5.1.x	\
	sklearn	Version 1.0.x	\
	DBeaver	Version 22.2.x	\

4. 云端环境

类别	部件	参数
服务器	CPU	2 颗 Intel Xeon 银牌 4214R 以上
	内存	服务器内存 256GB 以上
	硬盘	480GB 以上 SSD*2 （Raid1）
	网卡	4*1GbE
网络	\	1000Mbps 及以上

备注：具体设备由赛点提供。

八、选手须知

1. 选手自带工（量）具及材料清单

无需选手自带工具。

2. 主要技术规范及要求

该赛项主要涉及以下国家标准，参赛选手在实施竞赛项目中要求遵循如下规范：

序号	标准号	中文标准名称
1	DB21/T 2347.3-2014	信息技术行业职业技能 第 3 部分：软件开发
2	GB/T 32423-2015	系统与软件工程 验证与确认
3	GB/T 32424-2015	系统与软件工程 用户文档的设计者和开发者要求
4	GB 8566-1988	计算机软件开发规范
5	SJ/T 10367-1993	计算机过程控制软件开发规程
6	GB/T 36475-2018	软件产品分类

7	GB/T 36964-2018	软件工程 软件开发成本度量规范
8	GB/T 37691-2019	可编程逻辑器件软件安全性设计指南
9	GB/T 25000.2-2018	系统与软件工程 系统与软件质量要求和评价（SQuaRE）第 2 部分：计划与管理
10	GB/T 28174.1-2011	统一建模语言(UML) 第 1 部分：基础结构
11	GB/T 11457-1995	软件工程术语
12	GB/T 16260.1-2006	软件工程 产品质量 第 1 部分:质量模型
13	GB/T 32421-2015	软件工程 软件评审与审核
14	GB/T 32423-2015	系统与软件工程 验证与确认
15	GB/T 30264.2-2013	软件工程 自动化测试能力 第 2 部分：从业人员能力等级模型
16	GB/T 32904-2016	软件质量量化评价规范
17	GB/T 30998-2014	信息技术 软件安全保障规范

3. 选手注意事项

（1）参赛选手在比赛前应认真阅读赛项规程，严格按照赛项规程参与比赛，避免不必要失误。

（2）各参赛选手应在竞赛开始前一天按照规定的时间段进入赛场熟悉环境。

（3）各参赛选手不得统一着装，并不得穿有身份标识的服装。

（4）参赛选手应按照规定时间抵达赛场，凭身份证、学生证，以及统一发放的参赛证，完成入场检录、抽签确定竞赛工位号，不得迟到早退。并按工位号入座，检查比赛所需竞赛设备齐全后选手签字方可开始参赛。选手在比赛中应注意随时存盘。竞赛期间不准出场，竞赛结束后方开离场。

（5）参赛选手不得私自携带任何竞赛软硬件工具（各种便携式电脑、各种移动存储设备等）、设计资源、通信工具进入考场。

（6）参赛选手要严格遵守竞赛现场规则，如发现有冒名顶替等舞弊行为者，均取消竞赛资格。如遇到电脑或其他比赛用设备故障，可向裁判提出，获得及时解决。

（7）竞赛过程中，各参赛选手不得与其他人员讨论问题，也不得向裁判、巡视员和其他必须进入考场的工作人员询问竞赛项目的操作流程和操作方法，如有竞赛题目文字不清、软硬件环境故障问题时，可向裁判员询问。

（8）竞赛过程中除裁判和其他必须进入考场的工作人员外，任何其他非竞赛选手不得进入竞赛场地。

（9）竞赛结束后，参赛选手要确认成功提交竞赛要求的文件，裁判员与参赛选手一起签字确认，参赛选手在确认后不得再进行任何操作。

（10）其它未尽事宜，将在赛前向各领队做详细说明。

4. 竞赛直播

1.赛点提供全程无盲点录像。

2.可在赛点指定区域通过网络监控观摩比赛。

九、样题（竞赛任务书）

2023 年度“楚怡杯”湖南省职业院校技能竞赛高职高专组电子信息类 Python 程序开发赛项

[时量：240 分钟，试卷号：　]

（样卷）

竞赛任务书

场次号：_____　机位号（工位号、顺序号）：_____。

2022 年 12 月　日

一、注意事项

1. 请根据大赛所提供的竞赛环境，检查所列的硬件设备、软件清单、材料清单是否齐全，计算机设备是否能正常。

2.竞赛结束前，在竞赛平台提供的虚拟机中，根据赛题将各试题代码进行完善整合，并运行；根据竞赛平台左侧的答题区进行答题，根据题目对运行代码及结果进行截图。

3.竞赛结束时，请将答题区的答卷进行提交操作，答卷在竞赛结束前可重复提交。

二、竞赛环境

1.PC 机：系统已安装 Python 相关环境、MySQL 数据库、用户名密码分别为：root/123456。

2. 根据考题说明，从竞赛平台虚拟机桌面获取程序开发项目工程代码包。桌面的工程代码可以直接使用虚拟机中的 Pycharm 导入、编译、运行和发布。

三、软件组件

1.Python 编程语言及相关开发环境（Python、PyCharm）

2.Web 框架（Django）

3.Python 爬虫组件（Requests、lxml、BeautifulSoup）

4.Python 数据分析组件（NumPy、Pandas）

5.Python 可视化组件（Matplotlib、Pyecharts）

6.Python 机器学习组件（Scikit_Learn）

四、赛题

模块一：网络爬虫（15 分）

第 1 题：视频网站数据抓取

【任务说明】

数据是很多企业的生命，可以说，没有数据，就没有一切。企业首先要解决的问题就是数据问题，那么获取数据的手段有很多种，其中爬虫就是性价比最高的一种。现有一个视频网站，网站上有大量用户对各种视频的播放、评论、点赞等数据，请根据具体要求，编写爬虫实现数据抓取。

【任务要求】

以网站首页为入口，从该页面获取各大视频分类，每个分类下面均有大量视频及相应的点击、播放等数据。现需要通过爬虫抓取相应数据，具体要求如下：

1.使用 requests 库向 url 发送请求；

2.使 BeautifulSoup 或 Xpath 从响应内容中解析数据；

3.从首页中获取视频分类名和各类别链接；

4.向各类别链接发送请求，从响应内容中获取视频具体的播放、评论、点赞等数据；

5.将抓取的数据存入 MySQL 数据库中；

【工程代码】

获取桌面“赛题/01_网络爬虫/”路径下“01_视频数据抓取”文件夹中获取相关资料，结果保存至桌面“ 提交文档/01 ”文件夹中。

模块二：数据清洗（30 分）

第 2 题：视频网站用户数据清洗

【任务说明】

数据清洗是数据分析过程中很重要的一个环节，可以说，没有高质量的数据清洗就没有高质量的数据分析。在不准确的数据基础上做出的分析，结论将变得毫无价值和意义。

现有一份某个视频网站的会员数据，请根据任务要求完成数据清洗功能。

【任务要求】

数据集中存在字段缺失、空行、单位不统一、有重复数据等问题，请你使用 NumPy 和 Pandas 对数据进行清洗，具体要求如下：

1.缺失的体重和年龄字段使用均值填充；

2.缺失的爱好字段使用高频词填充；

3.身高单位统一为 cm；

4.体重单位统一为 kg；

5.空行直接删除；

6.重复数据只保留一条；

7.将清洗后的数据通过 Django ORM 保存到 MySQL 数据库中。

【工程代码】

获取桌面“赛题/01_数据清洗/”路径下“01_视频会员数据清洗”文件夹中获取相关资料，结果保存至桌面“ 提交文档/01 ”文件夹中。

模块三：数据分析及可视化（35 分）

第 3 题：视频网站会员分布情况分析

【任务说明】

现有两份数据，一份是视频网站用户数据，其中 IP 字段为用户在观看视频时的 IP 地址，另一份数据是 IP 和各地区对应数据，请根据任务要求完成数据分析功能。

【任务要求】

读取所需数据集后，给用户数据中增加省份字段记录用户的归属地，分析该视频网站中会员用户在中国各地区的分布情况并绘制出会员分布图。绘图要求如下：

1. 使用 PyEcharts 库绘制会员分布地图；
2. 使用 Django 框架在前端页面中渲染展示会员分布图；
3. 示意图如下：

图 1 网站会员分布示意图

【工程代码】

获取桌面“赛题/02_数据分析及可视化/”路径下“02_视频网站会员分布情况分析”文件夹中获取相关资料，结果保存至桌面“ 提交文档/02 ”文件夹中。

第 4 题：视频网站用户留存分析

【任务说明】

互联网流量竞争越来越激烈，各种获客手段层出不穷，但获客成本仍在不断提升。这就是问题所在，企业不可能无限制的投入成本拉取新用户。在当前互联网存量运营的阶段，留存重要性高于获客。获客是增长的必要条件，但在大多数情况下，我们过分强调了用户拉新，而忽略了用户留存，这可能是一个致命的错误。

请你根据任务要求对该网站进行用户留存分析。

【任务要求】

读取所需数据集后，分析不同类型的用户留存情况，并绘制用户留存矩阵图，横轴为不同类型的用户留存率，纵轴为活跃用户的数量。绘图要求如下：

1. 使用 PyEcharts 库绘制留存矩阵图；
2. 使用 Django 框架在前端页面中渲染展示留存矩阵图；
3. 示意图如下：

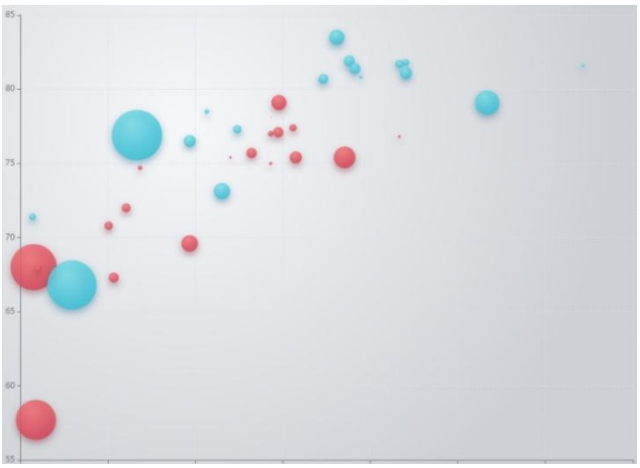

图 2 客户留存示意图

【工程代码】

获取桌面“赛题/02 数据分析及可视化/”路径下“03_视频网站用户留存分析”文件夹中获取相关资料，结果保存至桌面“ 提交文档/02 ”文件夹中。

模块四：机器学习（20 分）

第 5 题：视频网站用户活跃量预测

【任务说明】

使用该网站 5 年的用户行为日志数据，选择算法训练模型，对该视频网站的用户活跃量进行预测。

【任务要求】

1. 使用 Django ORM 读取数据库中的用户日志数据；
2. 对数据进行清洗和处理，将处理后的数据保存为 CSV 数据；
3. 根据任务要求使用 Pandas 读取 CSV 数据进行特征工程；
4. 划分训练集和测试集；
5. 构建机器学习模型；
6. 编写模型训练相关代码，完成模型训练；
7. 使用 PyEcharts 库对测试数据的预测结果和真实结果进行可视化，并使用 Django 在前端页面中渲染展示；
8. 将训练好的模型保存。

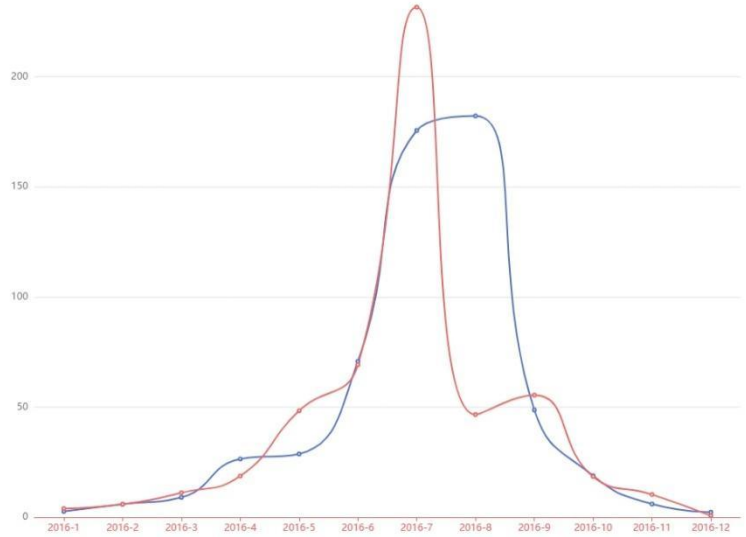

图 3 用户活跃量预测示例图

【工程代码】

获取桌面“赛题/03_机器学习/”路径下“04_视频网站用户活跃量预测”文件夹中获取相关资料，结果保存至桌面“ 提交文档/03 ”文件夹中。

第 6 题：视频网站每日广告收益预测

【任务说明】

使用视频网站的历史广告收益数据，构建机器学习模型，实现对该网站每日的广告收益进行预测。

【任务要求】

1. 使用 Django ORM 读取数据库中的广告收益数据进行数据清洗及处理；
2. 将清洗处理后的数据缓存到 Redis 数据库中；
3. 从 Redis 中读取数据进行特征工程；
4. 数据集划分；
5. 构建机器学习模型；
6. 编写模型训练相关代码，完成模型训练；
7. 使用 PyEcharts 库对测试数据的预测结果和真实结果进行可视化，并使用 Django 在前端页面中渲染展示；
8. 将训练好的模型保存。

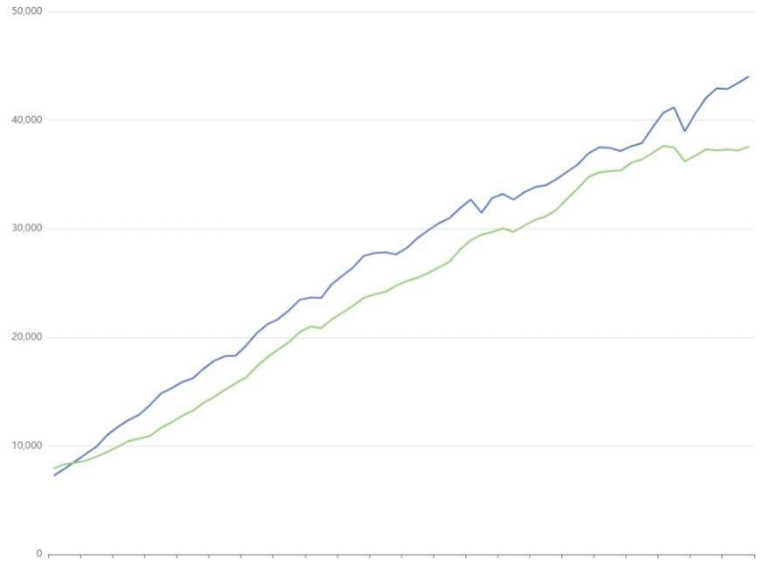

图 4 每日广告收益预测示例图

【工程代码】

获取桌面“赛题/03_机器学习/”路径下“05_视频网站每日广告收益预测”文件夹中获取相关资料，结果保存至桌面“ 提交文档/03 ”文件夹中。