

全国新职业和数字技术技能大赛

人工智能工程技术人员项目技术工作文件

全国新职业和数字技术技能大赛组委会技术组

2022年2月

目录

一、技术描述	3
(一) 项目概要.....	3
(二) 考核目的.....	3
(三) 参赛选手要求.....	3
(四) 选手需具备的能力.....	3
(五) 相关文件.....	5
二、竞赛模块	5
(一) 竞赛模块.....	5
(二) 模块简述.....	5
(三) 命题方式.....	6
三、评分标准	6
(一) 评价分（主观）	6
(二) 测量分（客观）	6
(三) 评分流程说明.....	6
(四) 统分方法.....	7
(五) 参赛队排名.....	7
(六) 裁判构成和分组.....	7
1. 裁判长.....	7
2. 裁判员的条件和组成.....	7
四、竞赛细则	7
(一) 选手要求及约定.....	7
(二) 裁判要求及约定.....	8
五、赛场、设施设备等安排	9
(一) 赛场规格要求.....	9

(二) 场地布局图.....	9
(三) 设施设备清单.....	10
1. 赛场提供设施、设备清单表.....	10
2. 竞赛场地禁止自带使用的设备和材料.....	11
(四) 赛场每个比赛工位所需设施.....	12
(五) 场地公需设施.....	12
六、安全、健康要求	13
附件 1 样题	13
附件 2 评分标准.....	15

一、技术描述

（一）项目概要

项目参考人力资源社会保障部等部委发布的“人工智能工程技术人员”新职业信息的要求，结合当前新职业从事的工作领域和技能要求，参照人社部发布的《人工智能工程技术人员国家职业技能标准》，同时结合全国新职业和数字技术技能大赛的技术工作要求设计本次竞赛的工作内容和考评标准。

（二）考核目的

为深入实施人才强国战略，推进“十四五”时期人工智能领域技术人才队伍建设，增强人工智能工程技术人员社会认同度，促进就业创业，培养选拔更多高素质人工智能工程技术人员，进一步改善新职业人才供给质量结构，为加快发展现代产业体系，推动经济高质量发展提供有力人才保障。同时，使参赛选手、裁判员、赛场工作人员等进一步熟悉人工智能工程技术人员职业标准要求，加深对人工智能工程技术要求的认识与理解。

（三）参赛选手要求

1. 人数要求：

本赛项每支参赛队由 1 名选手组成，每个省可以有 2 支队伍参赛，每省的参赛人员为 2 人。

2. 职业要求：

参赛选手为相应职业从业人员（职工身份，包括教师、博士后研究人员），按户籍属地原则（或在当地工作满 1 年以上），思想品德优秀，身心健康，具备相应职业（专业）扎实基本功和技能水平，且有较强学习领悟能力及应变能力的人员均可报名参加。

（四）选手需具备的能力

赛项主要考察人工智能工程技术人员职业能力包括：机器学习技术应用、自然语言处理技术应用、计算机视觉技术应用等方面的职业能力。本项目要求选手根据竞赛题目进行业务需求分析，利用提供的人工智能工程技术人员竞赛平台，通过企业级工作实施任务考察选手的综合职业能力。参赛选手应当具备的知识点和技能点如下表 1 所示：

表 1. 知识点和技能点

相关要求		权重比例 (%)
1	工作组织和管理	2
基本知识	-安全生产操作相关的技术规范要求和相关术语，以及针对特定岗位的特别要求 -精益生产的基础知识 -诚实与正直 -自我激励、团队合作问题解决、自我防护压力下有效的工作 -健康与安全法规、义务和文件与技能相关的最佳实践 -安全用电工作的原则	
工作能力	-在相关环境和其他因素中专业的工作 -在本地和远程环境中与同事及团队协同作业 -向团队或客户提出想法，响应客户需求	

	<ul style="list-style-type: none"> -在工作场所照顾自己和他人的安全 -采取适当的预防措施，尽量减少事故及影响 -采用符合国际标准的过程记录，为开发和修正提供可追溯的保障 -解释和认识国际符号图表和其他标准机构运用的国际语言 -协助工程师编写关于测试技术、实验室设备和规程的报告与记录 -与客户有效地沟通 -训练他人使用设施设备 	
2	机器学习技术应用	
基本知识	<ul style="list-style-type: none"> -人工智能机器学习、特征工程知识掌握 -了解常见主流算法、机器学习理论知识 -了解 python、pandas、numpy、sklearn 等常见类库知识 -常用加载数据集方法 -常用特征工程方法 -常用构建模型方法 -常用模型训练、预测、调参、模型评估知识 -JupyterNotebook 平台操作使用知识 	28
工作能力	<ul style="list-style-type: none"> -掌握加载数据集能力 -掌握特征选择、特征融合能力 -掌握模型选择能力 -掌握模型调参能力 -掌握模型加载、保存能力 -掌握基于 sklearn 的模型训练能力 -掌握模型评估能力 -掌握模型调整、优化能力 -掌握模型部署能力 	
3	自然语言处理技术应用	
基本知识	<ul style="list-style-type: none"> -人工智能自然语言处理知识掌握 -了解常见主流算法、自然语言处理理论知识 -了解主流的 PyTorch、TensorFlow 等深度学习框架知识 -了解预训练模型、微调策略 -了解常用自然语言处理范式 -了解 python、pandas、json、百度飞浆，PyTorch、TensorFlow 等常见类库知识 -常用加载数据集方法 -常用构建模型方法 -常用模型训练、预测、调参、模型评估知识 -JupyterNotebook 平台操作使用知识 	30
工作能力	<ul style="list-style-type: none"> -掌握加载数据集能力 -掌握数据处理能力 -掌握标签设计能力 	

	<ul style="list-style-type: none"> -掌握文本数据编码与解码 -掌握模型调参能力 -掌握模型加载、保存能力 -掌握基于百度飞浆、PyTorch、TensorFlow 的模型微调训练能力 -掌握模型评估能力 -掌握模型调整、优化能力 -掌握模型部署能力 	
4	计算机视觉技术应用	
基本知识	<ul style="list-style-type: none"> -人工智能深度学习知识掌握 -了解常见主流算法、深度学习神经网络知识 -了解主流的百度飞浆、TensorFlow 等深度学习框架知识 -了解 python、百度飞浆、TensorFlow 等常见类库知识 -常用加载数据集方法 -常用构建模型方法 -常用模型微调训练、模型转换、模型评估知识 	40
工作能力	<ul style="list-style-type: none"> -掌握加载数据集、数据集划分能力 -掌握数据集根据训练环境进行优化能力 -掌握基于百度飞浆或 TensorFlow 的模型微调训练能力 -掌握模型转换能力 -掌握模型评估能力 -掌握 TensorFlow 或百度飞浆模型部署能力 	
合计		100

（五）相关文件

本赛项技术工作文件只包含赛项技术工作的相关信息。除阅读本文件外，开展本赛项竞赛还需配合其他相关文件一同使用。

相关文件包括：新职业和数字技术技能大赛竞赛技术规则、竞赛日程安排、以及健康、安全及个人防护规定等。

二、竞赛模块

（一）竞赛模块

竞赛模块的任务、时间及分数安排见表 2。

表 2 比赛任务、时间及分数

模块编号	模块名称	竞赛时间	分数
A	自然语言处理技术应用	180min	30
B	机器学习技术应用	120min	30
C	计算机视觉技术应用	180min	40

（二）模块简述

A、B、C 模块的考核形式都以技能实操考核为主，竞赛覆盖了 A、B、C 三个模块的考核

要求。

(1) 模块 A: 自然语言处理技术应用

考核通过自然语言处理技术,完成数据预处理、模型训练与预测等内容。数据预处理考核内容包括数据多维度统计分析、词典设计、文本长度设计等内容;模型训练考核预训练模型调用、模型选择、超参数配置、模型标签设计、文本数据编码、模型训练、模型调优、模型保存等内容;模型预测主要考核能够调用已训练好的模型对无标签数据进行预测,并将模型预测结果解码成所要求格式,保存解码结果。

(2) 模块 B: 机器学习技术应用

考核通过机器学习技术,完成数据预处理、模型训练、模型预测等内容。数据预处理考核内容包括数据多维度统计分析、数据集划分、特征筛选、融合、降维、归一化等内容;模型训练考核模型选择、超参数配置、模型训练、模型调优、模型集成、模型保存等内容;模型预测主要考核能够调用已训练好的模型对无标签数据进行预测,并成功保存预测结果。

(3) 模块 C: 计算机视觉技术应用

考核通过深度学习技术,进行算法模型的微调训练,完成视觉数据预处理、模型训练、模型测试等内容。数据预处理考核内容包括标签文件配置、数据集划分、标注信息提取;模型训练考核搭建模型、超参数配置、模型训练;模型测试主要考核能够编写代码提取图片文件作为模型输入,正确配置超参数且成功调用模型。

(三) 命题方式

命题流程按照全国新职业和数字技术技能大赛的命题方式要求进行。赛题设计规范参照全国新职业和数字技术技能大赛赛项规则的要求进行设计,包含“试题文档”及其附属的“项目场地设施设备清单”。

竞赛样题随技术文件一起公布,竞赛试题由裁判长结合赛场设施设备、材料等实际情况,按照技术工作文件确定的试题调整工作流程和方法,对已公布的试题所涉及知识点进行不超过 30%的修改,作为竞赛正式试题,正式试题(含任务书、竞赛资源包、评分标准、参考答案等)全部保密。

三、评分标准

本项目评分标准分为:评价分(主观)、测量分(客观)。按各模块评分表分别设置评分小组,由裁判长指定各组裁判人员,分别对各模块进行评分。各评分小组负责所有选手同一指标的现场评分,并签字确认评分结果。

(一) 评价分(主观)

评价分(Judgement)打分方式:3名及以上裁判为一组,各自单独评分,再根据每组裁判人数计算出平均分为最后实际得分。

(二) 测量分(客观)

测量分(Measurement)打分方式:按任务设置若干个评分组,每组由3名及以上裁判构成。每个组所有裁判一起商议,对选手在该项中的实际得分达成一致后最终只给出一个分值,达到要求为“满分”,达不到要求为“0”分。

(三) 评分流程说明

(1) 裁判员以小组的形式进行评判工作,裁判员小组的分组和分工由裁判长执行。

(2) 在评判过程中,所有的评判结论必须由评判小组集体决定。

(3) 评判工作分为客观测量评分和主观评价评分两个部分。测量评分：针对比赛结果如选手的代码文件、数据截图按《评分表》细则进行测量评价。主观评价评分：针对选手比赛作品的主观判断进行评价，分数由裁判根据规定计算得出并记录到选手《评分表》。

(4) 技能操作部分

内容描述：在规定时间内，按任务书要求实现竞赛内容，竞赛结束，停止一切操作。

评判方式：裁判组在竞赛规定的结束时间后，分组对参赛队伍进行考评，裁判员按照评分表中要求，对提交结果进行打分评判。

(四) 统分方法

由裁判长助理进行统分，统分后由裁判长复核签字，最后由工作人员录入系统。

(五) 参赛队排名

关于成绩并列，当比赛现场出现选手总成绩并列时，裁判组首先将按照模块评分优先级不同的方式决定选手总成绩排名，评分优先级由大到小排序：模块 C > 模块 B > 模块 A，评分优先级比较仍不能区分选手总成绩排名时，由评分裁判对该组排名相同的选手提交的比赛模块所有主观评分项（评价）进行综合评价投票，投票领先的选手总成绩排名在前。

(六) 裁判构成和分组

1. 裁判长

赛场实行裁判长负责制，全面负责本赛项的竞赛执裁工作。裁判长由赛项组委会通过遴选审核确定。

2. 裁判员的条件和组成

(1) 裁判员须符合裁判员工作管理规范，赛前由赛项执委会统一组织裁判员培训。参加执裁的裁判员由赛项执委会抽取和推荐。被抽取或推荐的裁判员在大赛前可申请不参加裁判工作并放弃相应权利。一旦确认担任裁判员工作后，比赛中途不得更换人选。若裁判员不能满足裁判等技术工作需要，由裁判长按照大赛组委会相关要求处理。

(2) 裁判员应服从裁判长的管理，裁判员的工作由裁判长指派或抽签决定。在工作时间内，裁判员不得徇私舞弊、无故迟到、早退、中途离开工作地或放弃工作，否则将视其影响程度进行相应处理，直至取消裁判员资格并记录在案。

(3) 裁判员按工作需要，由裁判长将其分成加密裁判组、现场裁判组、评分组等若干小组开展工作。其中加密裁判组 2 人/组、评分组 3 名及以上/组。现场裁判组根据参赛工位和场次确定分组，原则上每组选手配 2 名裁判。各小组在裁判长的统一安排下开展相应工作。

四、竞赛细则

(一) 选手要求及约定

(1) 参赛选手应严格遵守竞赛规则和竞赛纪律，服从裁判员和竞赛工作人员的统一安排，自觉维护赛场秩序，不得因申诉或对处理意见不服而停止竞赛，否则以弃权处理。

(2) 参赛选手在赛前熟悉竞赛设备和竞赛时间内，应该严格遵守竞赛设备工艺守则和竞赛设备安全操作规程，杜绝出现安全事故。

(3) 参赛选手不得将通讯工具、任何技术资料、工具书、自编电子或文字资料、笔记本电脑、通讯工具、摄像工具以及其他即插即用的硬件设备带入比赛现场，否则取消选手比赛资格。

(4) 参赛选手应严格按竞赛流程进行竞赛。

(5) 参赛选手必须持本人身份证、并佩戴组委会签发的参赛证件，按竞赛规定的时间到指定的场地参赛。

(6) 参赛选手按照裁判长指令开始、结束竞赛。

(7) 参赛选手须在赛前 30 分钟到达赛场进行检录、抽取工位号，在赛前 10 分钟统一入场，进行赛前准备，等候比赛开始指令。正式竞赛开始但尚未检录的选手，不得参加竞赛。已检录入场的参赛选手未经允许，不得擅自离开。

(8) 参赛选手按规定进入竞赛工位，在现场工作人员引导下，进行赛前准备，检查并确认竞赛设备、竞赛工位计算机、配套的工量具、相关软件等，并签字确认。

(9) 裁判长宣布比赛开始，参赛选手方可进行竞赛操作。

(10) 参赛选手必须及时备份竞赛中自己的竞赛数据，防止意外断电及其它情况造成程序或资料的丢失。并将全部数据文件存储至计算机指定盘符下，不按要求存储数据，导致数据丢失者，责任自负。

(11) 竞赛过程中，选手若需休息、饮水或去洗手间，一律计算在比赛时间内。食品和饮水由赛场统一提供。

(12) 竞赛过程中，参赛选手须严格遵守相关操作规程，确保人身及设备安全，并接受裁判员的监督和警示，若因选手个人因素造成人身安全事故和设备故障，不予延时，情节特别严重者，由裁判长视具体情况作出处理决定（最高至终止比赛）并由裁判长上报大赛监督仲裁组；若因非选手个人因素造成设备故障，由裁判长视具体情况作出延时处理。

(13) 参赛选手在竞赛过程中不得擅自离开赛场，如有特殊情况，需经裁判员同意后，特殊处理。

(14) 参赛选手在竞赛过程中，如遇问题，需举手向裁判人员提问。选手之间不得发生任何交流，否则，按作弊处理。

(15) 裁判长在比赛结束前 15 分钟对选手做出提示。裁判长宣布比赛结束后，选手应立即停止竞赛操作，现场裁判员监督参赛选手在规定时间内必须把竞赛作品、赛题、图纸、草稿纸等所有相关内容上交至现场裁判员，如选手未按规定执行，裁判有权制止，并要求选手至指定位置。

(16) 竞赛结束后，由现场裁判员和选手检查确认提交的内容，现场裁判员当选手面封装上交竞赛作品，选手在收件表上签字确认，现场裁判员确认。

(17) 比赛结束，选手应立即清理现场，包括未使用的竞赛设备、工具及周边卫生等。经现场裁判员和现场工作人员确认后方可离开工位。经裁判长统一确认后，选手统一离开赛场。此项工作将在选手职业素养环节进行评判。

(18) 参赛选手在竞赛期间未经组委会的批准，不得接受其他单位和个人进行的与竞赛内容相关的采访；参赛选手不得私自公开比赛相关资料。

(二) 裁判要求及约定

(1) 裁判员须佩戴裁判员标识上岗。执裁期间，举止文明礼貌，接受参赛人员的监督。

(2) 严守竞赛纪律，执行竞赛规则，服从赛项组委会和裁判长的领导。按照分工开展工作，始终坚守工作岗位，不得擅自离岗。

(3) 裁判员的工作分为加密裁判、现场执裁、评判裁判等。

(4) 裁判员在工作期间严禁使用各种器材进行摄像或照相。

(5) 现场执裁的裁判员负责检查选手携带的物品，违规物品一律清出赛场，比赛结束后裁判员要命令选手停止竞赛操作。

(6) 竞赛中所有裁判员不得影响选手正常竞赛。

(7) 严格执行赛场纪律，不得向参赛选手暗示或解答与竞赛有关的内容。及时制止选

手的违纪行为。对裁判工作中有争议的技术问题、突发事件要及时处理、妥善解决，并及时向裁判长汇报。

(8) 要提醒选手注意操作安全，对于选手的违规操作或有可能引发人生伤害、设备损坏等事故的行为，应立即制止并向现场负责人报告。

(9) 严格执行竞赛项目评分标准，做到公平、公正、真实、准确，杜绝随意打分；严禁利用工作之便，弄虚作假、徇私舞弊。

(10) 严格遵守保密纪律。裁判员不得私自与参赛选手或代表队联系，不得透露竞赛的有关情况。

(11) 裁判员必须参加赛前培训，否则取消竞赛裁判资格。

(12) 竞赛过程中如出现问题或异议，服从裁判长的裁决。

(13) 竞赛期间，因裁判人员工作不负责任，造成竞赛程序无法继续进行或评判结果不真实的情况，由赛项组委会视情节轻重，给予通报批评或停止裁判资格，并通知其所在单位做出相应处理。

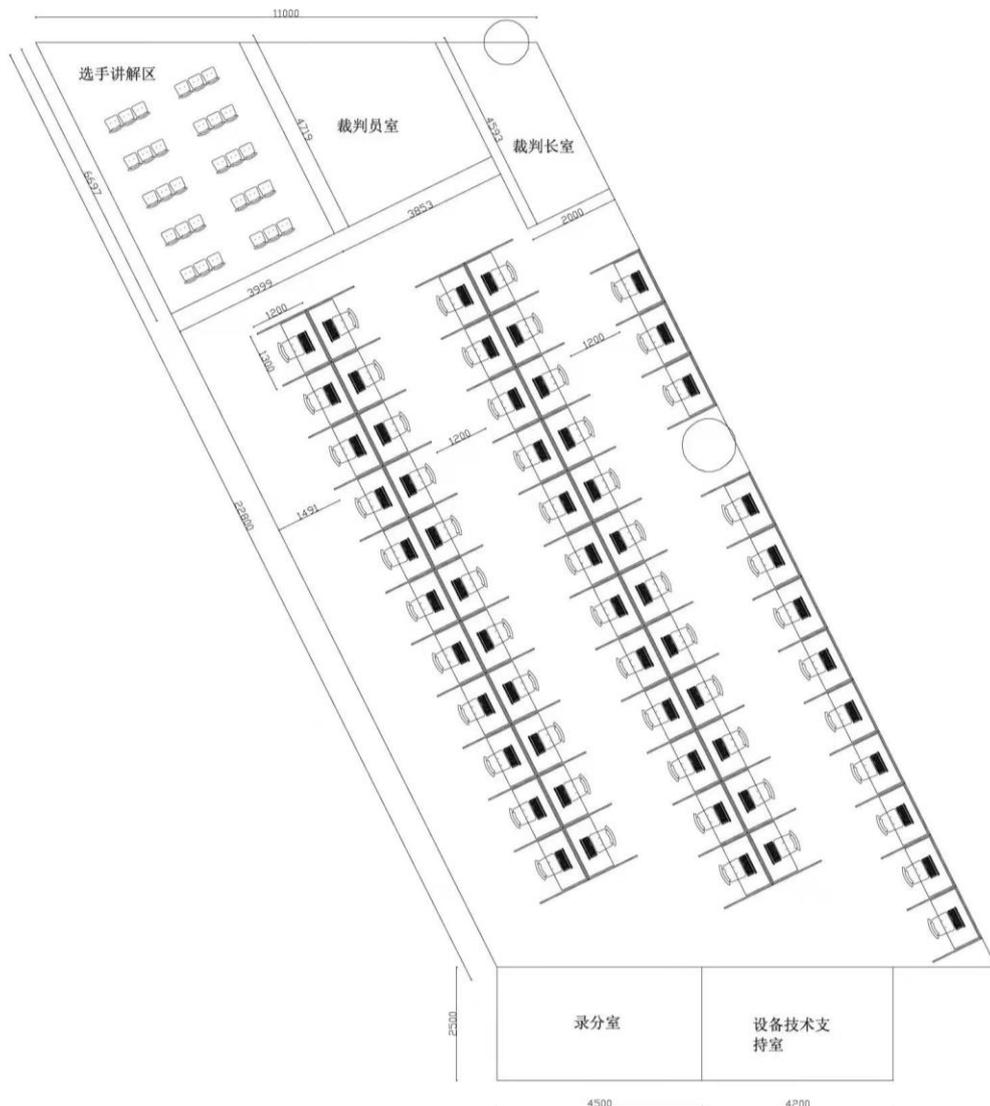
五、赛场、设施设备安排

(一) 赛场规格要求

本项目场地总面积按照赛位数量 58 个（包含备用工位）预估需要 272.5m² (20.6m*11.4m)，每个赛位面积为 1.56m² (1.2m*1.3m)，工位间隔 1.2m。

(二) 场地布局图

赛场布局请参考图 1:



(三) 设施设备清单

1. 赛场提供设施、设备清单表

本赛项所用技术平台，选用合作企业北京易华录信息技术股份有限公司&新大陆科技集团有限公司提供的“人工智能工程技术人员竞赛平台”。

全国新职业和数字技术技能大赛人工智能工程技术人员项目提供的竞赛设备由：实训软件、硬件设备及实训资源三部分组成。产品设计根据中华人民共和国人力资源社会保障部（2019）48号文件中关于“人工智能工程技术人员”的职业定义及主要工作任务，并遵循ISO、IEEE等国际组织发布的人工智能相关技术标准，涵盖了人工智能相关算法、深度学习技术、计算机视觉应用等技术。



人工智能工程技术人员竞赛平台效果图（仅供参考）

表 3. 人工智能工程技术人员项目赛场提供设施设备清单表

设备编号	场地设施清单	名称	需求规格描述	应用区域	备注
1	工位设施设备	人工智能工程技术人员竞赛平台	人工智能工程技术人员竞赛平台由实训系统、硬件设备及实训资源组成： 1. 实训系统包括模型训练框架、训练环境、模型部署环境、各类依赖库资源，可以完成模型训练、模型部署、优化与应用等工作； 2. 实训资源包括竞赛样题、实训指导书、深度学习项目案例包、计算机视觉项目案例包； 3. 硬件配置：Intel Core i7 8核 16 线程 CPU 处理器、内存 32GB DDR4，512GB 固态硬盘，NVIDIA GeForce 8GB 图形显卡。	操作区 选手工位	

2. 竞赛场地禁止自带使用的设备和材料

表 4. 禁止自带使用的设备和材料清单

序号	设备和材料名称
1	笔记本电脑、平板电脑
2	U 盘及可存储设备

3	通信设备
4	电动螺丝刀、测线仪等工具设备
5	易燃、易爆、放射及腐蚀性材料

注：未明确在选手携带工具清单中的，一律不得带入赛场。另外，赛场配发的各类工具、材料，选手一律不得带出赛场。

(四) 赛场每个比赛工位所需设施

每个比赛工位所需设施见表 5。

表 5 比赛工位所需设施

序号	设备名称	品牌/规格/型号	数量及单位	备注
1	办公桌椅	尺寸：1200*600mm	58 套	
2	电源插口	插孔要求：三孔	116 个	
3	网口		58 个	
4	插线板	插孔要求：4 口的二三插座 用电量：220V10A	58 只	

(五) 场地公需设施

场地公需设施见表 6，由实施保障单位提供。

表 6 场地公需设施

序号	设施设备名称	需求明细	单位	备注
1	办公桌椅	尺寸：1200*600mm	8 套	
2	门	带锁	4 扇	
3	会议桌	尺寸：2300*1200mm	1 张	
4	休息椅	尺寸：约 330*240mm	80 张	
5	警戒线	布局：每工位 2 米	116 米	
6	电源插口	插孔要求：三孔	18 个	
7	网口		8 个	
8	路由器	规格：千兆路由（企业级）	1 个	
9	机柜	规格：19 寸，32U	1 个	
10	桌面交换机	24 口交换机	4 个	
11	插线板	插孔要求：4 口的二三插座，用电量：220V10A	8 只	
12	打印纸	A4	5 箱	
13	签字笔	黑色	150 支	
14	订书器		5 个	
15	订书针		5 盒	
16	起钉器		5 个	
17	现场广播系统		2 个	
18	无线麦克风		4 个	
19	曲别针		5 盒	
20	便利贴		5 盒	
21	计时器		4 个	

22	饮水机		3 台	
23	饮用水		10 桶	
24	打印机		3 台	
25	电视	(带移动支架)	2 台	
26	赛场时钟	具有时、分、秒	4 套	

六、安全、健康要求

1. 大赛的安全目标——事故为零。
2. 在赛项承办单位内提供工作人员咨询服务、赛场布局图、消防设施分布情况等，张贴安全提示和赛场标识、路线标识，确定设置安保人员地点和当日现场所需的安保服务人员数量。
3. 赛项执委会须在赛前对本赛项全体裁判员、工作人员进行安全培训，并在赛前对选手进行培训，避免发生人身伤害事故，建立完善的安全事故防范制度。
4. 参赛专家、裁判、工作人员及指导教练、选手入住承办单位统一安排的宾馆、注意饮食卫生、乘坐承办单位统一安排的大巴车接送赛场及宾馆之间的往返。
5. 参赛选手公平竞赛，杜绝舞弊，遵守赛场纪律；遵守设备操作规程，安全、文明参赛；着装规范整洁，爱护设备，保持竞赛环境清洁有序。
6. 承办单位配备有医务服务、餐饮等后勤保障服务。
7. 所有人员应服从组委会管理及工作人员的指挥、调动，按照比赛秩序表提供的安排准时入场，准时参赛、准时离场。
8. 严格控制与参赛无关的易燃易爆以及各类危险品进入比赛场地，不许随便携带书包进入赛场。
9. 如遇特殊或紧急情况，按照疏散方向标识，指挥赛场人员安全有序地撤离。

附件 1 样题

(一) 模块 A、自然语言处理技术应用

1. 数据探索

进入本小题对应的文件夹，数据已采集好并存放于指定路径。数据文件中的每条数据有两个属性：“text”，对应每条数据的文本；“label”，对应每条数据的标签。

请打开相应文件，在该文件内编写代码完成以下操作：

- 1.1 使用 json 读取相应文件，编码格式为 utf-8，并输出指定数据；
- 1.2 统计数据总数，并使用指定类型输出；
- 1.3 统计每类标签对应的数据数量，请使用 dict 字典类型表示（字典内无序，格式为 标签:数据数量）；
- 1.4 请统计相应文件中除空格外的全部字符(汉字同样视为字符)，并输出去重后的字符数；
- 1.5 请统计全部数据的文本长度，并输出最大文本长度。文本长度指数据 text 中除空格外的全部字符数量；

2. 模型训练与验证

进入本小题对应的文件夹。文件夹中提供了训练模型验证的相关内容，请打开指定文件，在该文件内编写代码完成以下操作：

- 2.1 词表映射，请利用 dict 字典类型将指定文件中除空格外的全部字符(汉字同样视为

字符)以 1 起始使用数字编码, 编码后添加' <PAD>' (占位符), 并将其数字编码设置为 0;

2.2 利用上一步骤的结果编码文件的全部文本, 并根据标签特点生成每条数据的标签编码;

2.3 数据转换, 依次将映射后的文本编码与标签编码转化为模型可处理的格式;

2.4 搭建文本分类模型, 根据经验使用 tensorflow/keras 工具搭建文本分类模型, 模型搭建完成后, 请使用 `model.summary()` 打印模型结构;

2.5 划分数据集, 请将指定文件转换后的数据按照比例划分, 其中训练数据占比 X (全部数据的前 X), 验证数据占比 Y (全部数据的后 Y), 二者间数据不重复;

2.6 模型训练, 使用处理过的训练集与标签进行模型训练;

2.7 保存模型, 保存模型至当前目录;

2.8 模型验证, 使用处理过的验证集进行模型验证, 并输出模型在验证集上的 macroF1 值 (需大于 0.85)。

(二) 模块 B 样题: 机器学习技术应用

1. 数据探索

进入本小题对应的文件夹, 数值数据已采集好并存放于指定路径, 该数据有多列, 其中第一列为“id”, 是每条数据的编号; 最后一列为“label”, 是每条数据的标签; 其余列为特征数据。请打开指定文件, 在该文件内编写代码完成以下操作:

1.1 读取数据文件, 编码格式为 utf-8, 并输出指定数据;

1.2 统计数据总行列数将其转换为指定类型表示并输出;

1.3 统计每类标签对应的数据数量, 请使用 dict 字典类型表示 (字典内无序, 格式为 标签名: 数量);

1.4 按照相应的比例将数据划分为训练集与验证集, 二者间数据不重复, 切分过后, 请将训练集与验证集依次保存至当前目录;

2. 模型训练与预测

进入本小题对应的文件夹。文件夹中提供了训练模型的相关内容, 请打开指定文件, 在该文件内编写代码完成以下操作:

2.1 特征数据处理, 读取划分好的训练集和验证集以及测试数据, 提取文件中的特征数据, 并将其转换为模型可用的数据格式;

2.2 模型训练, 根据经验选择合适的模型, 并使用处理过的数据进行模型训练;

2.3 验证模型效果, 利用验证集验证步骤 2.2 生成的模型效果, 并输出 F1 宏平均值;

2.4 保存模型, 保存为当前目录下的相应文件;

2.5 模型预测, 利用训练好的模型对测试集进行预测, 并使用列表类型输出前 10 条预测结果。

(三) 模块 C、计算机视觉技术应用

1. 数据探索

进入本小题对应的文件夹, 训练数据存放在指定文件夹下, 文件夹下的图片以一定规则命名, 例如图片“1-2.jpg”, “1”代表该图片的标签; “2”代表该图片的编号。

请打开相应文件, 在该文件内编写代码完成以下操作:

1.1 请根据经验利用相关工具包读取相应文件中的图片数据;

1.2 统计文件夹内的图片数量, 并使用指定类型输出;

1.3 统计文件夹内图片的每类标签及其对应的图片数量, 请使用 dict 字典类型表示 (字典内无序, 格式为 { 图片标签: 图片数量 });

- 1.4 请根据上一步的结果输出图片的标签总数;
- 1.5 请统计图片的尺寸, 并使用 tuple 元组输出, tuple 元组的第一个元素是图片的宽 (单位像素), tuple 元组的第二个元素是图片的高 (单位像素);

2. 模型训练与验证

进入本小题对应的文件夹, 验证数据存放在指定文件夹下, 文件夹下的图片以“一定规则命名, 例如图片“1-2.jpg”, “1”代表该图片的标签; “2”代表该图片的编号。

请打开指定的文件, 在该文件内编写代码完成以下操作:

- 2.1 请设置所使用 GPU 编号并验证 GPU 是否可使用;
- 2.2 请根据经验利用相关工具包读取指定文件夹中的图片数据;
- 2.3 请将读取后的图片数据组织成模型可使用的格式;
- 2.4 根据经验使用 tensorflow/keras 工具搭建 vgg 系列模型, 模型搭建完成后, 请使用 model.summary() 打印模型结构;
- 2.5 模型训练, 使用处理过的训练集与标签进行模型训练;
- 2.6 保存模型, 保存模型至当前目录;
- 2.7 模型验证, 使用处理过的验证集进行模型验证, 并输出模型在验证集上的 macroF1 值。

附件 2 评分标准

模块 A、自然语言处理技术应用

模块内容	评分方面描述	测量或评价内容	最高分
数据探索 (10分)	打开答题文件, 查看步骤 1.1 打印结果	答题文件中的相应内容	2.00
	打开答题文件, 步骤 1.2 打印结果	答题文件中的相应内容	2.00
	打开答题文件, 步骤 1.3 打印结果	答题文件中的相应内容	2.00
	打开答题文件, 步骤 1.4 打印结果	答题文件中的相应内容	2.00
	打开答题文件, 步骤 1.5 打印结果	答题文件中的相应内容	2.00
模型训练 与验证 (20分)	打开答题文件, 步骤 2.4 有打印结果, 能看到模型结构	答题文件中的相应内容	5.00
	指定文件夹下含有模型文件	指定文件夹下的模型文件	5.00
	打开答题文件, 查看步骤 2.8 打印结果	答题文件中的相应内容	10.00

模块 B 样题: 机器学习技术应用

模块内容	评分方面描述	测量或评价内容	最高分
数据探索 (10分)	打开答题文件, 查看步骤 1.1 打印结果, 对应位置数字正确即可。	答题文件中的相应内容	2.00
	打开答题文件, 查看步骤 1.2 打印结果	答题文件中的相应内容	2.00
	打开答题文件, 查看步骤 1.3 打印结果	答题文件中的相应内容	2.00
	指定文件夹下含有相应文件, 含有一个文件得 2 分, 满分 4 分	指定文件夹下的有相应的文件	4.00
模型训练 与预测 (20分)	打开答题文件, 查看步骤 2.3 打印结果	答题文件中的内容	5.00
	指定文件夹下含有模型文件	指定文件夹下的模型文件	5.00
	打开答题文件, 查看步骤 2.5 打印结果, 对	答题文件中的相应内容	10.00

	应位置的数字有一个相同的得 1 分，满分 10 分，数量对不上的得 0 分。		
--	--	--	--

模块 C、计算机视觉技术应用

模块内容	评分方面描述	测量或评价内容	最高分
数据探索 (15 分)	打开答题文件，查看 1.2 打印结果	答题文件中的相应内容	3.00
	打开答题文件，查看步骤 1.3 打印结果	答题文件中的相应内容	5.00
	打开答题文件，查看步骤 1.4 打印结果	答题文件中的相应内容	2.00
	打开答题文件，查看步骤 1.5 打印结果	答题文件中的相应内容	5.00
模型训练 与预测 (25 分)	打开答题文件，查看步骤 2.1 有打印结果	答题文件中的相应内容	5.00
	打开答题文件，步骤 2.4 有打印结果，能看到模型结构	答题文件中的相应内容	5.00
	指定文件夹下含有模型文件	指定文件夹下的模型文件	5.00
	打开答题文件，查看步骤 2.7 打印结果	答题文件中的相应内容	10.00

注： 未尽事宜以竞赛现场规定或竞赛组委会规定为准