



CCF BDCI | CCF BIG DATA & COMPUTING
INTELLIGENCE CONTEST

2021 CCF BDCI 大数据与计算智能大赛

参赛指南

2021.09-2021.12



CCF BDCI CCF BIG DATA & COMPUTING
INTELLIGENCE CONTEST

目录

01 大赛介绍

02 赛程赛制

03 活动与奖励

04 赛题详情

05 问题答疑



CCF BDCI CCF BIG DATA & COMPUTING
INTELLIGENCE CONTEST

01 大赛介绍



CCF大数据与计算智能大赛（CCF Big Data & Computing Intelligence Contest, 简称CCF BDCI）是大数据与人工智能领域的算法、应用和系统大型挑战赛事。大赛面向重点行业和应用领域征集需求，以前沿技术与行业应用问题为导向，以促进行业发展及产业升级为目标，以众智、众包的方式，汇聚海内外产学研用多方智慧，为社会发现和培养了大量高质量数据人才。

大赛迄今已成功举办八届，累计吸引全球1500余所高校、1800家企事业单位及80余所科研机构的12万余人参与，已成为中国大数据与人工智能领域最具影响力的活动之一，**是中国大数据综合赛事第一品牌**。

2021年第九届大赛以**“数引创新，竞促汇智”**为主题，立足余杭、面向全球，于9月至12月举办。大赛将致力于解决来自政府、企业真实场景中的痛点、难点问题，邀请全球优秀团队参与数据资源开发利用，广泛征集信息技术应用解决方案。

主办单位	中国计算机学会
支持单位	杭州市余杭区人民政府
承办单位	浙江杭州未来科技城管理委员会、CCF大数据专家委员会、CCF自然语言处理专业委员会、 CCF高性能计算专业委员会、CCF计算机安全专业委员会、CCF计算机应用专业委员会、 教育部易班发展中心、LandInn、数联众创
独家战略合作	百度
创新合作	华为
赛道合作	360集团未来安全研究院、中原银行、中国中文信息学会
赛题合作	明朝万达、高德地图、竹云、爱奇艺、OneFlow一流科技、 中国联通、GeaGraph、浙江省发展和改革委员会、锐安科技
合作媒体	学术世界
官方竞赛平台	DataFountain

国际级权威赛事

中国计算机学会 (CCF) 主办，杭州市余杭区人民政府支持，浙江杭州未来科技城管理委员会、CCF大数据专家委员会、CCF自然语言处理专业委员会、CCF高性能计算专业委员会、CCF计算机安全专业委员会、CCF计算机应用专业委员会，教育部易班发展中心、LandInn、数联众创联合承办，DataFountain为官方竞赛平台。

顶级专家评审指导

梅宏、李国杰、倪光南、徐宗本等**七大院士**指导，国内外百余名学术专家评审，百度、华为、中原银行等十余家名企技术专家参评。

价值上亿的免费算力

作为本届大赛的独家战略合作伙伴，百度携飞桨平台为所有赛题提供了**价值上亿的免费GPU算力**，成功报名大赛任意赛题即可领取。

以练备赛，以赛促练

大赛开启赛练结合模式，**大奖赛和训练赛兼具**，以练备赛，以赛促练。鼓励选手在训练赛道开源分享，帮助初学者锻炼基础能力，辅助高校开展相关专业及课程教学实践工作。

真诚靠谱的赛事激励

百万奖金池，入围决赛团队即可获得招聘绿色通道、政府及知名投资机构的融资及孵化扶持，还有各合作单位提供的项目成果应用与试点机会，以及周冠军、邀请榜等赛事花样活动大礼包收到手软。

丰富的线下活动

入围总决赛的优秀团队，将与百余名领域顶尖专家共度两天一夜“**决赛嘉年华**”，**数据科学家之夜、大数据及计算智能高峰论坛、颁奖盛典**应有尽有，差旅食宿全安排。

携手精英共建开源

所有决赛入围作品，将在条件允许情况下统一开源，借由全球技术精英的力量，惠及所有大数据及人工智能参赛者、学习者及从业者。

全流程反作弊净化赛圈

全自动反作弊机制贯穿全程，违规禁赛黑名单长期开放，人工持续复核违规行为，晋级代码云端逐行复核，特设仲裁委员会公正判决，全方位构建公平公正竞赛环境。



梅 宏

中国科学院院士
中国计算机学会理事长



李国杰

中国工程院院士
中科院计算所研究员



倪光南

中国工程院院士
中科院计算所研究员



徐宗本

中国科学院院士
西安交通大学教授



吴建平

中国工程院院士
清华大学教授



陈 纯

中国工程院院士
浙江大学教授



管晓宏

中国科学院院士
清华大学西安交通大学
双聘教授

权威专家 业内大咖 为作品把关



CCF BDCI | CCF BIG DATA & COMPUTING
INTELLIGENCE CONTEST

02 赛程赛制

09.16

赛题正式上线
开启线上报名

11.20

截止正式赛报名

11.23-12.05

大赛决赛 (线上)

09.22-11.20

初赛A榜阶段

11.22

初赛B榜阶段

12中旬

大赛总决赛 (线下)

*具体赛程视赛题而定, 详情请见大赛官方网站

正式赛

数据算法赛道

通过算法模型，就某场景或行为的特定指标进行优化。

数字方案赛道

提供数据和开放式赛题，自由发挥创意，提供完整解决方案。

自主平台赛道

通过使用国产AI开源框架，进行相应算法赛题的研究。

先进系统赛道

通过解决方案，提升对指定数据或环境的系统优化能力。

创新创业赛道

提交围绕大数据、人工智能、相关技术及其在制造、健康、交通、零售、家居、教育、金融等行业场景的创新项目。

训练助力赛道

根据公开数据集，出具自然语言处理、图像识别、分类预测等不同方向的低难度赛题，辅助参赛者学习。

专题赛

数字安全专题

以“悬赏未来安全最优解”为主题，提出人工智能安全、数据安全、工业互联网安全等数字化场景中的新型风险问题，设置数据算法赛题，促进安全数据开源共享。

信息检索专题

全国信息检索学术会议（CCIR）发起的技术评测比赛。致力于围绕实际问题探索解决方案，并为信息检索领域科研人员提供必需的研究数据支持。



CCF BDCI | CCF BIG DATA & COMPUTING
INTELLIGENCE CONTEST

03 活动与奖励

选手奖项

- 1、单赛题奖：**单赛题将评选出一二三等共五支团队，各获奖团队队员均可获得奖金、证书、面试绿色通道等奖励；参与任意赛题并获得排行榜成绩，即可获得电子证书。
- 2、CCF BDCI综合奖：**CCF BDCI综合奖将评选出综合特等奖一名、最佳单项奖三名，参赛团队队员均可获得奖金、证书及相关荣誉。
- 3、参赛特别奖：**在线上赛过程中将根据参赛团队周榜、邀请等活动可获得特别奖项。
- 4、飞桨特别奖：**百度飞桨为参加第九届CCF BDCI大赛的开源生态贡献者设立的专属奖池，总金额高达50万元人民币。

除奖金奖项外，参加大赛的优秀团队有机会获得：

- 1、绿色通道：**优秀团队可获得相关赛题出题单位实习、就职绿色通道。
- 2、创业孵化：**优秀团队可获得由政府、投资机构提供的融资、孵化支持。
- 3、落地应用：**优秀团队可获得由大赛各合作单位提供的项目成果应用、试点机会。

其他奖项

- 1、优秀指导老师奖牌：**各赛题一等奖指导老师，及对大赛辅导、宣传、组织提供重要支持的老师可获得。
- 2、特别贡献奖奖牌：**在大赛组织工作中做出突出贡献的个人或单位可获得。
- 3、优秀合作伙伴奖杯：**对大赛组织做出重要支持的合作单位可获得。

辅助学习活动



助学特别计划

于9月至11月，联合百所高校共同就训练赛题进行系统学习、训练，辅助教学、能力测试。



《CCF BDCI大咖说》系列专题报告

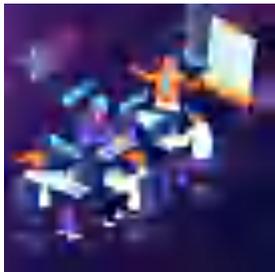
于9月至11月，邀请CCF各专委专家进行相关技术方向的线上报告，讲述大数据及人工智能的算法及产业应用。



走进高校

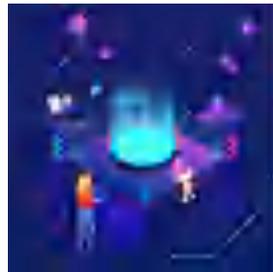
于9月下旬至11月，走进20余所高校，邀请大数据及人工智能知名专家及企业，做大赛相关技术专题报告。

行业交流活动



产学研线上沙龙

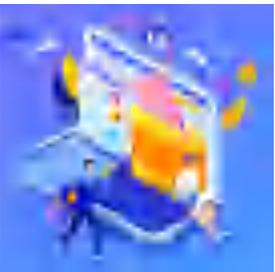
于9月下旬至11月，邀请企业、相关领域专家、参赛者通过视频直播进行人才交流、赛题答疑、线上互动。



决赛嘉年华

于12月中下旬，邀请决赛入围团队现场答辩评审，举办数据科学家之夜、大数据及计算智能高峰论坛、颁奖盛典等活动。

赛中激励活动



“邀请有礼”活动

9月24日-11月20日（北京时间），已报名用户可在赛题页“邀请有礼”区域参与活动，实时查看已邀请新用户数量并赢取奖励。



周榜单活动

9月27日-11月20日，算法类赛题每周将开放周榜单（具体时间以赛题活动时间为准），每周一公布上周测评排行榜，排行榜的第一名将获得相应奖品。



CCF BDCI CCF BIG DATA & COMPUTING
INTELLIGENCE CONTEST

04 赛题详情

出题单位



赛题名称

技术方向

图像识别

赛题奖金

10万元



扫码报名

奖项	数量	金额
一等奖	1	4万
二等奖	2	2万
三等奖	2	1万

1

所属赛道 自主平台赛道

赛题背景

目前的研究数据普遍缺少细粒度语义信息，导致现存的分割或识别任务缺少时空细粒度动作语义模型，时空细粒度语义的人体动作类内方差大、类间方差小等特点，将导致产生一系列问题，而利用粗粒度语义的识别模型进行学习也难以获得理想的结果。

赛题任务

参赛选手利用比赛提供的训练集数据，构建基于骨骼点的细粒度动作识别模型，完成测试集的动作识别任务。

出题单位

CCF BDCI X Bai 百度 飞桨

赛题名称

千言-问题匹配
鲁棒性评测

技术方向

文本分类

赛题奖金

10万元

扫码报名



奖项	数量	金额
一等奖	1	4万
二等奖	2	2万
三等奖	2	1万

2

所属赛道 自主平台赛道

赛题背景

在与训练集同分布的测试集上，当前大多数工作采用单一的指标评测模型的好坏。这可能过高的估计了模型能力，并缺乏对模型鲁棒性的细粒度评估。

赛题任务

给定一组问题对，判断问题对在语义上是否匹配(等价)。

出题单位



赛题名称

基于BERT的大模型容量挑战赛

技术方向 性能优化

赛题奖金 5万元



扫码报名

奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

3

所属赛道 自主平台赛道

赛题背景

参赛者需要在系统层面上进行优化使得可以进行参数量尽可能多的大模型的训练。参赛者最终需要提交一个BERT的实现模型，通过提高隐含层参数量 (hidden size) 提升模型的参数量，并在基于保证参数量的基础上尽可能提升系统的吞吐率。

赛题任务

参赛者需要将实例化的模型置于 GPU 上，并基于 BERT 模型进行分布式训练方面的改进（常用的技术包括数据并行、模型并行、ZeRO 和流水并行等）。参赛者的模型禁止使用 fp16 混合精度训练以及模型集成，最多可以在 2 卡上对模型进行训练，每张卡显存占用不超过 12G。

出题单位



赛题名称

基于MindSpore AI框架
实现零售商品识别

技术方向

图像分类

赛题奖金

5万元



扫码报名

奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

4

所属赛道 自主平台赛道

赛题背景

在高效管控货架排面的场景下，利用AI零售商品识别技术辅助人工完成陈列审查可以在时间、精准度、数据化程度等多方面得到提升，助力企业赢得消费者的选择。

赛题任务

参赛团队将设计算法对零售商品进行识别和分类，选手需要通过分析训练集中标注数据的特点，构建AI模型，对测试集中测试数据进行识别和分类。

出题单位



赛题名称

产品评论观点提取

技术方向 文本预处理

赛题奖金 5万元



扫码报名

奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

5

所属赛道 金融创新专题

赛题背景

随着互联网的不断深入普及，越来越多的用户在体验产品和服务时选择将感受与评论分享在互联网上。这些评价和反馈信息对企业针对性地改善产品和服务有极强的指导意义，但互联网信息的海量性让人工成本高昂。

赛题任务

通过自然语言处理的语义情感分析技术判断出银行产品评论文本的情感倾向，并能进一步通过语义分析和实体识别，标识出评论所讨论的产品关键信息。

出题单位



赛题名称

个贷违约预测

技术方向 异常检测

赛题奖金 5万元



扫码报名

奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

6

所属赛道 金融创新专题

赛题背景

因为缺乏对新客群的了解，对新生细分客群的风控处理往往成为金融普惠的重要阻碍。

赛题任务

利用已有信贷数据辅助目标业务风控模型的创建，基于迁移学习捕捉不同业务中用户基本信息与违约行为之间的关联，帮助实现对新业务的用户违约预测。

出题单位



赛题名称

技术方向

情感分析

赛题奖金

4万元

扫码报名



奖项	数量	金额
一等奖	1	1.6万
二等奖	2	8千
三等奖	2	4千

7

所属赛道 数据算法赛道

赛题背景

剧本角色的情感识别主要是对剧本中对白和动作描述中涉及到的角色从多个维度进行分析并识别出情感。相对于通常的新闻、评论性文本的情感分析，有其独有的业务特点和挑战。

赛题任务

对剧本场景中每句对白和动作描述中涉及到的每个角色的情感从多个维度进行分析和识别。该任务的主要难点和挑战包括：1) 剧本的行文风格和通常的新闻类语料差别较大，更加口语化；2) 剧本中角色情感不仅仅取决于当前的文本，对前文语义可能有深度依赖。

出题单位



赛题名称

基于UEBA的用户
上网异常行为分析



技术方向

异常检测

赛题奖金

5万元



扫码报名

奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

8

所属赛道 数据算法赛道

赛题背景

为了有效保护企业敏感数据，践行企业安全操作行为准则，杜绝由异常操作行为导致的企业敏感数据泄露安全事件发生，用户异常行为分析与识别成为重难点技术之一。

赛题任务

利用机器学习、深度学习，UEBA等人工智能方法，基于无标签的用户日常上网日志数据，构建用户上网行为基线和上网行为评价模型，依据上网行为与基线的距离确定偏离程度。

出题单位



赛题名称

POI名称生成



技术方向

图像分割

赛题奖金

5万元

扫码报名



奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

9

所属赛道 数据算法赛道

赛题背景

图像资料是制作POI数据的重要来源，依据图像数据生成POI名称是其重要环节。用人工作业的方式处理POI数据，成本高，效率低。如何自动化从图像中提取POI的名称是亟需解决的一大难题。

赛题任务

根据提供的图像、挂牌检测结果、和OCR识别结果，设计算法，给图像中指定POI挂牌生成一个完整的名称。本次评测旨在算法：生成的名称正确、通顺、完整、能够体现被制作对象的起名意愿。

出题单位

赛题名称



技术方向

文本摘要

赛题奖金

5万元



扫码报名

奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

10

所属赛道 数据算法赛道

赛题背景

客服中心每天都需要接通大量的客户来电，客户来电需要进行语音转文本，同时对文本进行概括，提取客户核心诉求，但是人工总结会增加客服工作量，降低工作效率，因此期望使用AI算法进行自动的文本摘要生成。

赛题任务

赛题任务是对客户通话数据进行摘要提取，因属于特定领域的通话数据，所以同宽泛性文本摘要提取存在一定的异同，主要的难点如下：1、语音通话通过第三方服务转写为文本内容，存在一定的转写错误；2、文本长度不固定，长短不一，可能存在文本长度过长的现象；3、因为是具体领域的客服通话文本，专业词汇可能较多。

出题单位

赛题名称



技术方向

异常检测

赛题奖金

5万元

扫码报名



奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

11

所属赛道 数据算法赛道

赛题背景

IAM厂商参考UEBA、用户画像等行为分析技术，来探索一种既能确保用户体验，又能提高身份鉴别强度的方法。而在当前IAM的探索进程当中，目前最为可具落地的方法，是基于规则的行为分析技术，因为它的可理解性很高，且很容易与身份鉴别技术进行联动。

赛题任务

基于用户认证行为数据及风险异常标记结构，构建用户认证行为特征模型和风险异常评估模型，利用风险评估模型去判断当前用户认证行为是否存在风险。1、利用用户认证数据构建行为基线。2、采用监督学习模型，基于用户认证行为特征，构建风险异常评估模型，判断当前用户认证行为是否存在风险。

出题单位



赛题名称

openLookEng
跨域数据分析
性能提升



技术方向 性能优化

赛题奖金 5万元

扫码报名



奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

12

所属赛道 先进系统赛道

赛题背景

openLookEng的跨域能力为“东数西算”工程及其场景提供了关键技术方案，在免数据搬迁的情况下实现跨地域跨数据中心的数据融合分析。但面对跨地域场景下存在的带宽有限、网络时延等诸多问题，如何实现高性能分析，是openLookEng社区一直在探索的关键技术方向。

赛题任务

基于openLookEng社区所提供的比赛分支，进行代码优化，实现性能提升。

出题单位



赛题名称

13

所属赛道 先进系统赛道

赛题背景

目前图计算在金融场景的运用最为成熟，贷前审批、贷后管理、反欺诈、反洗钱等业务均对图计算能力有要求，包括但不限于k度邻居、找环、社区发现。

赛题任务

赛题使用简化的金融仿真数据，数据为带有时间戳和金额的账户间转账数据。基于此数据查找满足条件的交易环路，用于洗钱、反欺诈等场景的分析。

技术方向 图计算

赛题奖金 5万元



扫码报名

奖项	数量	金额
一等奖	1	2万
二等奖	2	1万
三等奖	2	5千

出题单位



赛题名称

泛在感知数据关联融合计算



技术方向

性能优化

赛题奖金

2万元

扫码报名



奖项	数量	金额
一等奖	1	8千
二等奖	2	4千
三等奖	2	2千

14

所属赛道 先进系统赛道

赛题背景

近年来针对人像、车辆、设备等数据的泛在感知采集正在得到广泛的应用。这些感知手段往往相互独立，难以在前端直接对人、车、物的数据进行有效的关联融合。但通过设置多个感知点位，经过一定时间的数据积累，可以在后台进行大数据的关联分析，计算得到人、车、物之间的关联关系，实现泛在感知数据的关联融合。关联计算的准确率和海量数据的计算性能，是实际工作中的主要难点。

赛题任务

本赛题模拟提供两种感知手段分别采集的要素数据，由参赛者通过统计学方法、数据挖掘、图计算等相关算法，将两种数据进行关联融合计算，最终给出两种手段采集到的关键要素之间的对应关系。

出题单位



赛题名称

教育基础设施数量规划
容量预测分析

技术方向

回归预测

赛题奖金

2万元

扫码报名



奖项	数量	金额
一等奖	1	8千
二等奖	2	4千
三等奖	2	1千

15

所属赛道 数字方案赛道

赛题背景

数字社会建设作为浙江数字化改革聚焦的五大关键点之一，有力支撑全生命周期公共服务跨部门系统。在此背景下，各类提供公共服务的基础设施的规划部署与容量预测成为数字社会建设的重中之重。

赛题任务

先依据各学校信息推断各学校在地图坐标系下的具体坐标，给出文本地址映射为具体坐标的解决方案，形成地图坐标系下的学校地址位置数据；再结合浙江省对应设市区内历年统计数据，预测分析浙江省重点设市区未来五年教育基础设施规划方案。



以上赛题杰出作品还可参与CCF BDCI综合奖评选

奖项	数量	金额
综合特等奖	1	3万
最佳算法能力奖	1	1万
最佳商业价值奖	1	1万
最佳创新探索奖	1	1万

*具体评测规则参见各赛题官网

出题单位

赛题名称



技术方向 医疗；图像分割

奖品奖项



扫码报名

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

16

所属赛道 训练助力赛道

赛题背景

目前临床对血管瘤的病灶的分割，主要由专家人工勾画，受临床经验水平的影响，分割结果无法避免人为误差。而采用人工智能，精准且海量的数据样本成本较高，如何利用小样本数据集训练血管瘤超声图像的自动精准分割，成为目前热门的研究方向之一。

赛题任务

利用深度学习方法，在小数据集训练出一个血管瘤的优秀分割模型，以达到辅助医生进行治疗血管瘤的目的。

出题单位



赛题名称

基于序列信息的
microRNA和gene
关系对预测



技术方向

回归预测

扫码报名



奖品奖项

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

17

所属赛道 训练助力赛道

赛题背景

microRNA等生物小分子在生物过程中发挥着重要的作用，可以通过调控gene进而调控下游的生物过程。预测潜在的micro RNA和gene的关系可以帮助科学家更好的理解疾病的发生机理，为精准医疗提供更多潜在的靶点。

赛题任务

依据真实的microRNA和gene的序列信息，提取序列中的信息，结合已知的关系对信息，利用机器学习相关技术，建立microRNA和gene的关系预测模型，判断某一对microRNA和gene是否有关系。

出题单位



赛题名称

机器翻译 领域适应



技术方向

机器翻译



扫码报名

奖品奖项

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

18

所属赛道 训练助力赛道

赛题背景

机器翻译的性能不仅依赖于大规模的双语数据，还取决于训练和测试数据之间的领域匹配程度。具有丰富数据资源领域的机器翻译性能不断提高，但是由于数据资源获取困难等原因，某些特殊领域的翻译效果还不够理想。如何利用富资源领域的的数据帮助低资源领域提升翻译质量是一个热点研究问题。

赛题任务

参赛队伍需要基于提供的训练样本进行中到英机器翻译模型的构建与训练，并基于口语、专利、医药三个领域测试集分别提供翻译结果。

出题单位



赛题名称



技术方向

目标检测



扫码报名

奖品奖项

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

19

所属赛道 训练助力赛道

赛题背景

出租车发票在日常财务发票报销中较为常见，由于这类发票样式丰富，区域性特点明显，并且包含大量模糊字迹和错位字迹，因此准确的定位发票文字字段，准确的识别文字和结构化输出显得十分重要。

赛题任务

本赛题任务是利用图像处理、机器学习、深度学习等方法训练出租车发票的文字检测，识别模型，并实现识别结果的结构化输出。

出题单位



赛题名称

船舶航行 轨迹预测



技术方向

回归预测

扫码报名



奖品奖项

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

20

所属赛道 训练助力赛道

赛题背景

目前，大多数船舶轨迹预测模型还停留在基于特定运动学方程的模型上，这种模型受限程度较高，已经很难应对当今繁杂的海上交通情形。利用机器学习相关技术可以挖掘船舶历史轨迹数据中潜在的船舶运动模式，建立高精度的船舶航行轨迹预测模型，提高船舶海上风险应对能力。

赛题任务

依据真实船舶历史航行轨迹数据，利用机器学习相关技术，建立船舶轨迹预测模型，预测船舶下一时间段的航行轨迹。

出题单位



赛题名称

渔船作业方式识别



技术方向

回归预测

扫码报名



奖品奖项

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

21

所属赛道 训练助力赛道

赛题背景

随着大数据发展和轨迹数据挖掘的深入研究，渔船轨迹数据为渔业数据挖掘及相关应用提供非常好的机会。海洋捕捞渔船作业类型分为多种,主要有拖网、围网、张网、刺网等方式,违规捕捞作业将会对海洋生态环境和渔业资源产生严重的影响,同时也给渔业管理带来困难。因此,需要依据真实渔船轨迹数据(已脱敏)利用机器学习相关技术,建立稳健的渔船作业方式识别模型,用于准确识别渔船的作业方式,便于渔业管理的同时也为后续渔业科技研究带来便利。

赛题任务

依据真实渔船轨迹数据(经纬度、速度、方向、时间等),利用机器学习相关技术,建立稳健的渔船作业方式识别模型,有效识别渔船的作业方式。

出题单位



赛题名称

法律领域篇章级多事件检测

技术方向

关键抽取

扫码报名



奖品奖项

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

22

所属赛道 训练助力赛道

赛题背景

为帮助司法办案人员快速理清案件的发展状况，掌握法律案件中包含哪些类型的事件，需要依据真实存在的法律案件信息（人名、时间等信息已重造）利用机器学习等相关技术，建立稳健的事件检测模型，用于判断法律案件中所包含的各个事件对应的事件类型，为后续抽取各事件所涉及的元素提供有利信息。

赛题任务

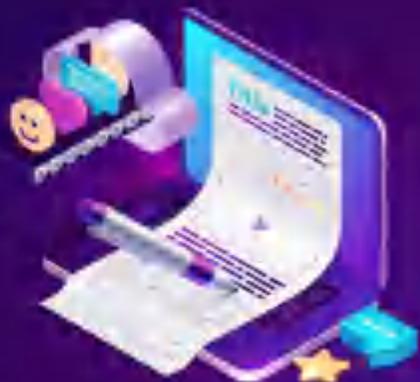
利用机器学习、深度学习等相关技术，建立稳健的篇章级多事件检测模型，解决事件检测任务中事件触发词不明显等问题，有效判断一个真实发生的婚姻家庭类法律案件中包含多少事件以及对应的事件类型。

出题单位



赛题名称

新闻摘要 自动生成



技术方向

文本摘要



扫码报名

奖品奖项

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

23

所属赛道 训练助力赛道

赛题背景

随着互联网、社交媒体的迅速发展，各类新闻文章层出不穷，读者在面对海量新闻信息时，难以有效发现哪些是自己感兴趣的新闻内容。为了帮助读者快速了解新闻文章，并找到其感兴趣的新闻内容，本训练赛题旨在研究构建高质量摘要生成模型，通过为新闻文档生成简短摘要，帮助读者快速理解新闻内容。

赛题任务

依据真实的新闻文章，利用机器学习相关技术，建立高效的摘要生成模型，为新闻文档生成相应的内容摘要。

出题单位



赛题名称



技术方向

RS推荐系统

扫码报名



奖品奖项

名额	奖项	颁布时间
前3名	1个CCF会员名额+纪念奖牌	12月6日
前50名	电子证书	12月6日

24

所属赛道 训练助力赛道

赛题背景

推荐系统正是在互联网快速发展之后的产物。为帮助电商系统识别用户需求，为用户提供其更加感兴趣的信息，从而为用户提供更好的服务，需要依据真实的图书阅读数据集，利用机器学习的相关技术，建立一个图书推荐系统。用于为用户推荐其可能进行阅读的数据，从而在产生商业价值的同时，提升用户的阅读体验，帮助创建全民读书的良好社会风气。

赛题任务

依据真实世界中的用户-图书交互记录，利用机器学习相关技术，建立一个精确稳定的图书推荐系统，预测用户可能会进行阅读的书籍。

出题单位



赛题名称

技术方向

异常检测

赛题奖金 10万元

扫码报名



奖项	数量	金额
一等奖	1	5万
二等奖	1	3万
三等奖	1	1万
优胜奖	2	5千

25

所属赛道 数字安全专题

赛题背景

恶意软件作者为逃避检测，在组件中引入了多态性。虽然属于同一恶意软件“家族”的恶意文件具有相同形式的恶意行为，但由于编写者使用了各种策略不断修改和/或混淆，原本隶属同一家族的文件看起来像许多不同的文件，给检测识别的准确性带来了高度挑战。

赛题任务

设计算法对恶意软件进行识别和家族分类，选手需要通过分析训练集中各家族恶意软件的特点，构建AI模型，对测试集中各个家族的恶意软件进行区分。设市区未来五年教育基础设施规划方案。

出题单位



赛题名称

技术方向 异常检测

赛题奖金 9万元

扫码报名



奖项	数量	金额
一等奖	1	4万
二等奖	2	2万
三等奖	2	1万

26

所属赛道 数字安全专题

赛题背景

恶意短信一直是黑灰产引流的重要渠道，不法分子为逃避检测，通过使用变体字发送恶意短信绕过拦截规则的情况越来越多。如何精准和高效地还原变体字文本，提高非法信息的抽取能力，以及新型变体字还原的泛化性和时效性，就成为了解决这一难题的“关键之钥”。

赛题任务

对训练集中的短信样本进行分析，采用深度学习建模的方法将测试集中新出现的短信变体字还原为正常信息文本，即不含有变体字、干扰字符，所有变体字部分应使用常见简体汉字、字符来表示，同时需要保证不包含变体字的正常文本不受影响。

出题单位

赛题名称



技术方向

异常检测

赛题奖金

9万元

扫码报名



奖项	数量	金额
一等奖	1	4万
二等奖	2	2万
三等奖	2	1万

27

所属赛道 数字安全专题

赛题背景

在工业生产环境中，工作人员违规使用手机，屡屡造成安全生产事故，引起人员伤亡。因此，如果引入人工智能技术，对设置在生产工区内摄像头采集的视频图片进行分析，快速、准确识别员工违规使用手机的行为，并进行提示和规范，可有效加强安全生产监管，实现降本、提质、增效，加速数字化转型进程。

赛题任务

结合提供的数据训练集，识别判断图片上的人物是否存在使用手机行为，提高识别的正确率。方法不限，分类方法、目标检测方法等均可。

出题单位

赛题名称

技术方向



文本预处理

赛题奖金

2.5万元

扫码报名



奖项	数量	金额
一等奖	1	1万
二等奖	2	5千
三等奖	2	2.5千

28

所属赛道 信息检索专题

赛题背景

希望通过考察模型的知识量，驱动业界对于模型知识表示的研究，实现预训练模型向下游任务更有效的知识转移。本次预训练模型的知识量评估侧重于考察模型的记忆能力，特别是在few-shot或zero-shot情况下的记忆能力和多知识组合能力，后期会加入更多的推理类问题，以及考察模型本身的泛化能力。

赛题任务

本赛题构建了完型填空形式的英文测评数据集，评估预训练模型在9个领域、两大知识类型（事实知识、常识知识）、不同难度任务上的知识含量。

出题单位

赛题名称



技术方向 对话系统

赛题奖金 2.5万元

扫码报名



奖项	数量	金额
一等奖	1	1万
二等奖	2	5千
三等奖	2	2.5千

29

所属赛道 信息检索专题

赛题背景

本赛题对NLU领域的意图识别及槽位填充任务进行考察。期望通过小样本学习任务减少产品对大量新类别标注数据的依赖，通过域外检测任务识别未知意图，摆脱对已知意图的干扰，同时达到尽可能好的学习效果。

赛题任务

根据用户与系统的单轮对话，识别对话用户意图并进行槽位填充。除基础的意图识别及槽位填充任务外，本赛题额外包括2个子任务：人机交互-NLU-1（小样本学习任务）：根据基础意图类别数据及少量含标注的新意图类别样本，完成新意图类别的识别及槽位填充任务。人机交互-NLU-2（域外意图检测任务）：除识别出训练数据中已知的意图类别外，对于未知意图类别数据进行检测。

出题单位

赛题名称



技术方向

文本预处理

赛题奖金

2.5万元

扫码报名



奖项	数量	金额
一等奖	1	1万
二等奖	2	5千
三等奖	2	2.5千

30

所属赛道 信息检索专题

赛题背景

本赛题针对命名实体任务开展鲁棒性评测。本次评测提供给参赛者的原始数据集训练模型，而验证集和测试集由鲁棒性评测工具 TextFlint 产生，以最终排名依据为新的鲁棒性验证集和测试集的性能表现。

赛题任务

给定未分好词但有实体标记的中文文本训练数据，设计算法对验证集中词语的实体类型进行预测，其中部分文本经过了鲁棒性评测工具TextFlint的扰动。预测的任务是判断哪些词语是人名(PER)，地理位置(LOC)，机构(ORG)或行政单位(GPE)，并以(B,M,E,S)的格式给出词语的边界。



CCF BDCI CCF BIG DATA & COMPUTING
INTELLIGENCE CONTEST

05 问题答疑

● Q：电脑端如何参与比赛报名？

1、点击链接进入大赛专题页：<https://www.datafountain.cn/special/BDCI2021>

2、点击感兴趣的赛题



3、点击“报名参赛”



● Q: 手机端如何参与比赛报名?



1、手机微信扫码进入大赛专题页



2、点击“查看赛题”



3、选择相关赛题



4、点击“立即报名”

● Q：竞赛届小白参加比赛有些害怕？不知如何下手？

A: **无需担心!** 每道赛题我们都提供了基础的baseline供大家参考。同时也可以扫码添加DF小助手官方微信，加入赛题群聊，和“大神”们一起讨论。无论你是“身经百战”，还是“初出茅庐”，都欢迎你报名大赛，加入大家，共同遨游数据之海!

● Q：关于飞桨特别奖

A: 飞桨特别奖是百度飞桨为参加第九届CCF BDCI大赛的开源生态贡献者设立的专属奖池，**总金额高达50万元人民币。**

各赛题决赛top5且参赛技术方案使用飞桨框架并开源于AI Studio平台与GitHub平台的队伍，即可获得税前金额5,000元人民币/队的飞桨特别奖。

各赛题非top5队伍也不要灰心，百度飞桨为参赛技术方案使用飞桨框架并开源于AI Studio平台与GitHub平台的各队送上惊喜大礼包，不仅有100h GPU Tesla V100算力，更有飞桨周边奖品等你来拿。

● Q：如何组队？组队需要注意什么？何时截止组队？

A：所有报名参加同一赛题的参赛者，可进行组队操作。每队1-5人，不可重复组队。报名之后，默认自己为一队。在“参赛队伍”一栏中，可查看我的队伍、参赛队伍、组队申请。**每队1-5人，不可重复组队，一旦组队不可撤销。为保证每支参赛团队享有相对平等的提交机会，组队需满足组队成员在该赛题中的提交总次数 \leq 开赛天数*赛题每日最高提交次数。组队截止日期为11月20日。**

● Q：什么是AB榜？

A：A/B榜是指在对结果进行评分排名时，将测试数据分割为A/B两份，分别评分并生成对应排行榜。A榜和B榜的存在是为了防止参赛队伍对测试数据过拟合，分割方式可能会采用随机分割方式、时序分割方式或者其他专业分割方式。A榜在“提交开放阶段”为每支参赛队伍的提交结果自动评分并排名,生成A榜；B榜在“提交截止阶段”为参赛队伍选定的候选文件进行自动评分并排名,生成B榜。

● Q：非大陆用户如何实名认证？

A：非大陆用户可以通过“海外实名认证”渠道人工完成实名认证。



赛题持续上线中，敬请期待.....



欢迎扫码添加DF小助手

诚挚解答竞赛相关问题



更多赛事信息

尽在DataFountain



CCF BDCI | CCF BIG DATA & COMPUTING
INTELLIGENCE CONTEST

Thanks