中国高校计算机大赛

中国高校计算机大赛——2019 大数据挑战赛

通　知

中国高校计算机大赛是由教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会和全国高等学校计算机教育研究会联合主办，面向高校学生的高水平计算机类系列竞赛，其中大数据挑战赛被列入全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2019 大数据挑战赛是在中国高校计算机大赛主办单位的指导下，由清华大学、南开大学与字节跳动公司联合主办，由亚马逊 AWS 提供资源支持，并以企业真实场景和实际数据为基础的高端算法竞赛。大赛面向全球高校在校生开放，旨在提升高校学生对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用。

本次大赛面向全球高校在校学生，鼓励高校教师参与指导。参赛队伍需要根据赛题要求设计相应的算法进行数据分析和处理，比赛结果按照指定的评价指标使用在线评测数据进行评测和排名，得分最优者获胜。

请各校积极配合，按照通知和大赛章程做好组织工作，并在指导教师工作量认可及参赛队伍经费等相关方面给予支持。竞赛详情请参见附件“2019 大数据挑战赛竞赛规程”。

教育部高等学校计算机类专业教学指导委员会
教育部高等学校软件工程专业教学指导委员会
教育部高等学校大学计算机课程教学指导委员会
全国高等学校计算机教育研究会（代章）
2019 年 4 月

2019 中国高校计算机大赛——大数据挑战赛
竞赛规程（第二版）

中国高校计算机大赛是由教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会和全国高等学校计算机教育研究会联合主办，面向高校学生的高水平计算机类系列竞赛，其中大数据挑战赛被列入全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2019 大数据挑战赛（以下简称“大赛”）是在中国高校计算机大赛主办单位的指导下，由清华大学、南开大学与字节跳动公司联合主办，亚马逊 AWS 提供资源支持以及科赛提供竞赛平台支持，并以企业真实场景和实际数据为基础的高端算法竞赛。大赛面向全球高校在校生开放，旨在提升高校学生对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用，本次大赛鼓励高校教师参与指导。

一、 参赛对象

大赛面向中国及境外在校学生（包括高职高专、本科、研究生），具体要求如下：

1. 可以单人参赛或自由组队，每个参赛队伍人数最多不超过 3 人，允许跨年级、跨专业、跨校组队。
2. 每人只能参加一支队伍（即个人参赛后不可再与他人组队参赛，或个人参加一个队伍后不可再参加另一个队伍），允许最多有一名指导老师，指导教师须为在职高校教师。
3. 报名时应具有在校学籍，已毕业的学生不具备参赛资格。
4. 参赛选手应保证报名信息准确有效，报名时应在大赛网站上提供所在学校开具的学籍证明材料。
5. 大赛主办和技术支持单位如有机会接触赛题和相关数据的人员不允许参赛，比赛期间在字节跳动实习的学生不允许参赛。

二、 赛制说明

本次大赛分为报名&预选赛、初赛、复赛和决赛四个阶段，其中预选赛是由参赛队伍根据预选赛题在本地进行算法设计和调试并在官网提交结果进行评测；初赛和复赛均要求参赛者在科赛网的 K-Lab 平台上进行数据处理、算法调试和生成结果，可使用平台提供的计算资源和工具包；决赛要求参赛者进行现场演示和答辩。

1. 报名&预选赛（4 月 28 日–6 月 15 日）

- 报名方式：登录大赛官网，完成个人信息注册并提交在校学籍证明；报名信息经审核通过后以个人形式参加预选赛，大赛不收取任何报名费用。
- 预选赛：报名成功之后，参赛选手从大赛官方网站下载数据，在本地进行算法调试，并在线提交结果。如果最终成绩达到规定分数线，选手将收到晋级初赛的通知邮件。预选赛的详细情况见附件“预选赛——文本情感分类模型”。
- 报名时间：2019 年 4 月 28 日–6 月 15 日，报名截止时间是 2019 年 6 月 15 日 16:00。
- 大赛官网：www.kesci.com/landing/bdc2019
- 大赛官方交流 QQ 群：753413531 或 758344321

2. 初赛（5 月 26 日–6 月 30 日）

参赛队伍需要在科赛网的 K-Lab 平台上完成数据处理、建模、算法调试、生成结果等，所有比赛数据均不可下载，可使用平台提供的计算资源和工具包。

- 参赛选手通过预选赛之后，可以组队参加正式比赛。组队成功之后，选手可以从大赛网站下载比赛的 sample 样本数据文件，理解和分析竞赛数据模式。
- 5 月 26 日 12:00 开始，竞赛平台提供 1 亿的初赛训练数据，用于参赛队伍训练模型以及制定预估策略；同时，平台提供 5 百万的测试数据，用于参赛队伍在比赛中的模型评估和排名。
- 初赛采用 Public-Private 榜与实时测评方式：
  （1） 每个参赛队伍每天可以进行 3 次 Public 提交，系统立即进行评测，评测完成之后返回 Public 成绩。排行榜将选择参赛队伍在本阶段的历史最优 Public 成绩进行排名展示，实时更新排行榜。
  （2） 每个参赛队伍可以选择 2 个 Private 提交，若参赛队伍在一个 Private 提交处进行了多次提交，则新提交版本将覆盖旧版本。
- 在 6 月 30 日 12:00，系统对参赛队伍的 Private 提交进行评测，并选取分数最好的 Private 成绩作为初赛成绩，公布所有队伍的初赛成绩和排名。

初赛截止时间是 6 月 30 日 12:00，TOP10 团队将获得初赛奖励，初赛成绩排名前 100 名参赛队伍将进入复赛。

3. 复赛（7 月 15 日–8 月 11 日）

复赛参赛队伍需要在科赛网的 K-Lab 平台上完成数据处理、建模、算法调试、生成结果等，所有比赛数据不可下载，可使用平台提供的计算资源和工具包。

- 7 月 15 日 12:00 开始，竞赛平台提供 10 亿的初赛训练数据，用于参赛队伍训练模型以及制定预估策略；同时，平台提供 2 千万的测试数据，用于参赛队伍

在比赛中的模型评估和排名。

- 7 月 15 日 12:00−8 月 10 日 23:00，每个参赛队伍每天可以进行 3 次结果提交，系统立即进行评测，评测完成之后返回成绩。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示，实时更新排行榜。
- 8 月 3 日 12:00 开始，竞赛平台提供 1 亿的最终成绩测试数据，此数据集仅用于复赛队伍的最终成绩评比和排名，不参与比赛过程中每天排名。参赛队伍可以随时提交该数据集的结果，如有多次提交则以当前最后提交为准。
- 8 月 11 日 12:00，系统对参赛队伍提交的最终测试集结果进行评测，并根据评测结果，公布所有复赛队伍的最终成绩和排名。

复赛截止时间是 8 月 11 日 12:00，TOP10 团队的代码需要进行审核，符合要求的排名前 10 名参赛队伍选手将受邀参加决赛。

4. 决赛（8 月下旬）

- 决赛将以现场答辩会的形式进行，晋级决赛团队需提前准备答辩材料，包括答辩 PPT、参赛总结、算法核心代码。
- 答辩现场，每支队伍面对评委有 15 分钟的陈述时间和 10 分钟的问答时间。评委将根据选手的技术思路、理论深度和现场表现进行综合评分。
- 决赛分数将根据参赛队伍的算法成绩和答辩成绩加权得出，评分权重为复赛 B 阶段 70%，决赛答辩 30%。
- 决赛地点和时间安排另行通知，受邀参加决赛的选手在北京决赛期间的食宿由大赛组委会安排，往返交通费及其他费用自理。

三、 奖项设置

初赛和决赛的奖金池总额为 30 万元人民币，所有奖金均为税前金额。

1. 初赛奖项

前 10 名队伍颁发初赛名次证书，每个队伍奖金 1000 元。该奖项由平台利用测试数据测试算法，并根据测试结果排名产生。

2. 复赛与决赛奖项

总决赛队伍获得一等奖和二等奖，入围决赛参赛队伍的指导教师获得优秀指导教师奖，复赛 11-30 名队伍获得三等奖。

奖项名称	数量	对象	奖励办法
一等奖	3	决赛前 3 名队伍	证书以及奖金：

			- 第 1 名奖金 15 万元 - 第 2 名奖金 5 万元 - 第 3 名奖金 2 万元
二等奖	7	决赛 4-10 名队伍	证书，奖金 1 万元
三等奖	20	复赛 11-30 名队伍	证书
优秀指导教师奖	10	决赛队伍指导教师	证书

3. 周周星

自大赛公布排行榜之日起，每周榜单排名前三名的参赛队伍将获得周星星。周周星以每周五中午 12 点的评分为准，取前三名，发放精美纪念礼品；对于前面已经获得周周星的队伍，不重复发放，名额按名次顺延。

4. 其他激励

- **设置实习岗位：**字节跳动特别为本次大赛开设实习岗位，具体岗位描述可以在网站 https://job.bytedance.com/intern 上通过搜索关键字“2019 大数据挑战赛”查看。参赛选手可以随时投递简历，企业相关部门将结合个人简历和初赛表现，通知部分学生进行实习面试。
- **招聘绿色通道：**复赛排名前 50 的队伍可获得字节跳动校招面试直通卡（即招聘流程省略简历筛选及笔试筛选阶段，直接进入面试阶段 ，2020 年底前有效）。

五. 违规处理

参赛者应本着诚实、公平的态度参加比赛，如在以下情况出现违规，大赛组织委员会（简称“大赛组委会”）有权取消参赛者所在队伍的参赛资格，情节严重者将通报参赛者所在高校并追究其违法责任。

1. 账号使用：参赛者有义务保证账号信息的真实性和有效性，且账号仅限于参赛者本人使用；参赛者禁止使用多账号参赛，同一参赛者不可使用多个账号进行提交、刷分操作；如根据判断认为参赛账号存在异常或违背正常使用条例，组委会可以单方面暂停或终止该账号登录大赛平台。
2. 比赛成果：严禁参赛队伍之间相互抄袭。如不同参赛队伍提交结果高度相似，经判定存在抄袭行为的，组委会将取消相关参赛队伍的参赛资格，相关参赛成绩无效。另外，参赛者应保证其在比赛过程中所产出的所有成果未侵犯任何第三方的知识产权、商业秘密及其他合法权益。如第三方因为参赛者侵权行为提出索赔、诉讼等，参赛者应承担由此产生的全部责任及损失。

3. 数据使用：对于大赛提供的数据（数据集），参赛者须仅在比赛场景下使用，同时不得以任何形式使用比赛之外的任何数据参赛。对于不提供下载的比赛数据，参赛者不得以任何形式擅自复制、下载或获取。参赛者如发现任何出现数据未授权访问的可能，应立即通知组委会并积极提供相关信息。
4. 代码分享：在大赛举办期间，未经组委会同意，参赛者禁止公开分享与赛事相关的数据、模型和代码；大赛结束之后，参赛者可以在拥有模型和代码的知识产权的情况下自行选择公开分享，但需要确保此类公开共享不会侵犯任何第三方的知识产权、商业秘密及其他合法权益。
5. 参赛者若在参赛过程中发现相关规则漏洞或技术漏洞，有义务及时告知资委会相关漏洞的信息，组委会将对提供相关信息的参赛者表示相关感谢；若参赛者利用相关漏洞进行参赛，经判断查证后，成绩将会被判断为无效成绩。

六. 申诉与仲裁

1. 参赛团队或选手对不符合大赛规定的设备、工具和软件，有失公正的评判和奖励以及工作人员的违规行为等，均可向大赛组委会提出申诉。组委会负责受理比赛中提出的申诉并进行调解仲裁，以保证大赛的顺利进行和大赛结果的公平公正。组织委员会作出的仲裁结果为终局决定。
2. 申诉报告应明确申诉内容，指定一名成员作为联系人，并要有参赛队伍成员的签名，否则申诉将不予以受理。
3. 组织委员会将在收到申诉之日起 5 个工作日之内受理，并认真核查和处理。

七. 其他

本大赛规程的最终解释权归“中国高校计算机大赛——大数据挑战赛”组织委员会所有。

“中国高校计算机大赛——大数据挑战赛”组织委员会

2019 年 5 月

附：预选赛——文本情感分类模型

本预选赛要求选手建立文本情感分类模型，选手用训练好的模型对测试集中的文本情感进行预测，判断其情感为「Negative」或者「Positive」。所提交的结果按照指定的评价指标使用在线评测数据进行评测，达到或超过规定的分数线即通过预选赛。

一. 比赛数据

数据样本格式：

NO	列名	类型	字段描述
1	ID	int	文本唯一标识
2	review	string	文本
3	label	string	文本的情感状态

其中，训练集的样本规模为 6328，测试集的样本规模为 2712。为保证比赛结果的真实性，每周系统将替换新的测试集数据。

二. 提交结果

选手提交.csv 的结果文件，文件名可以自定义，但文件内的字段需要对应。其中，ID 表示文本唯一标识，pred 表示预测该条文本的情感状态是否为「Positive」。

结果文件的内容示例：

ID	Pred
1	0.123456
2	0.654321
3	0.799212

三. 评估方法

选手提交结果的评估指标是 AUC（Area Under Curve），本次测评采用 Public-Private 榜与实时测评方式。每个选手总共有 10 次 Public 提交和 6 次 Private 提交，若选手在 Private 提交截止时间前没有选择提交到 Private 的结果文件，系统将自动选择其 Public 最高分所对应的结果文件进行 Private 测评。选手只要选择 Private 提交结果，系统直接反馈通过或不通过的结果。

附：赛题描述——文本点击率预估

搜索中一个重要的任务是根据 query 和 title 预测 query 下 doc 点击率，本次大赛参赛队伍需要根据脱敏后的数据预测指定 doc 的点击率，结果按照指定的评价指标使用在线评测数据进行评测和排名，得分最优者获胜。

一. 比赛数据

1. **sample 样本：**query 以及对应网页数据的的明文数据，并已经分词为 term，term 之间空格分割，样本行数 3000；这份样本供参赛队伍了解基本的参数数据模式。
   说明：所有数据按列分割，分隔符为”,"，为不带表头的 CSV 数据格式。

   sample 样本格式：

列名	类型	示例
query_id	int，一个 query 的唯一标识	1
query	字符 string，term 空格分割	"字节跳动"
title	字符 string，term 空格分割	"字节跳动 - 百科"
label	int，取值{0, 1}，有点击为 1，无点击为 0	1

2. **training 样本：**脱敏后的 query 和网页文本数据，并已经分词为 term 并脱敏，term 之间空格分割，样本规模初赛 1 亿，复赛增加到 10 亿；这份数据主要用于参赛队伍训练模型，制定预估策略。

   training 样本格式：

列名	类型	示例
query_id	int	3
query	hash string，term 空格分割	1 9 117
query_title_id	title 在 query 下的唯一标识	2
title	hash string，term 空格分割	3 9 120
label	int，取值{0, 1}	0

3. **test 样本：**脱敏后的 query 和网页文本数据，并已经分词为 term 并脱敏，term 之间空格分割，样本规模初赛 5 百万，复赛 2 千万（与初赛不重复）；这份数据用于参赛队伍比赛中模型评估和排名，初赛队伍根据初赛队伍根据初赛排名择优进入复赛；复赛 test 数据仅用于复赛排名，不计入复赛最终成绩。

   test 样本格式：

列名	类型	示例
query_id	int	7
query	hash string，term 空格分割	1 11 132
query_title_id	title 在 query 下的唯一标识	2
title	hash string，term 空格分割	3 2 27 14 1123

4. **最终成绩 test 样本：**脱敏后的 query 和网页文本数据，样本规模 1 亿，样本数据格式与 test 样本格式相同；这份数据用于复赛队伍的最终成绩评比和排名，不参与比赛过程的排名；数据放出日期为复赛结束前一周。

二. 选手提交结果

对于测试数据“query_id, query, query_title_id, title”，选手提交数据是“query_id, query_title_id, prediction”，提交格式说明如下：

列名	类型	示例
query_id	int	1
query_title_id	title 在 query 下的唯一标识	4
prediction	对应 title 点击率的预测值，范围∈(0,1)	0.5

例如：

对于测试数据：1,1 2 3,4,1 4 9

提交样例为：1,4,0.5 （其中 1 位 query_id，4 为 query_title_id, 0.5 为 prediction）。

说明：提交数据需要与 test 样本顺序一致，格式为无表头 CSV 数据，列之间仅有逗号分隔，没有空格。

三. 评估标准

选手提交结果的评估指标是 qAUC，qAUC 为不同 query 下 AUC 的平均值，计算如下：

$$qAUC = \frac{sum(AUC_i)}{query\_num}$$

其中 AUCi 为同一个 query_id 下的 AUC（Area Under Curve）。

最终使用 qAUC 作为参赛选手得分，qAUC 越大，排名越靠前。

2019“大数据挑战赛”组织机构名单

(按照界别和姓名拼音排序)

一、 竞赛指导委员会

主　任：孙家广（清华大学）

副主任：

陈　纯（浙江大学）

龚　克（南开大学）

廖湘科（国防科技大学）

马维英（字节跳动）

二、 竞赛专家委员会

主　任：王建民（清华大学）

副主任：

杜小勇（中国人民大学）

周傲英（华东师范大学）

乔　木（字节跳动）

委　员：

崔　斌（北京大学）

高铁杠（南开大学）

蒋树强（中国科学院计算技术研究所）

金　海（华中科技大学）

李占怀（西北工业大学）

梁吉业（山西大学）

马　帅（北京航空航天大学）

彭智勇（武汉大学）

王国仁（北京理工大学）

王宏志（哈尔滨工业大学）

王晓阳（复旦大学）

郑子彬（中山大学）

严　林（字节跳动）

苏俊阳（字节跳动）

周　珣（字节跳动）

三、 竞赛组织委员会

主　任：丁贵广（清华大学）

副主任：

卢　苇（北京交通大学）

张玉志（南开大学）

田　野（字节跳动）

委　员：

贝毅君（浙江大学）

郭东伟（吉林大学）

邝　坚（北京邮电大学）

廖明宏（厦门大学）

龙明盛（清华大学）

骆　斌（南京大学）

舒　坚（南昌航空大学）

王　泉（西安电子科技大学）

王志强（深圳大学）

吴　卿（杭州电子科技大学）

臧斌宇（上海交通大学）

张志政（东南大学）

朱　青（北京工业大学）

朱国斌（电子科技大学）

敖　雪（字节跳动）

唐　菲（字节跳动）

范向伟（科赛网）