中国数字人文开放创新研究大赛

和鲸赛事结项报告

1. 项目总结

1.1. 项目概述

1.1.1. 项目名称：中国数字人文开放创新研究大赛（共 8 大赛道）

- ☐ 保守或融贯：重审《学衡》杂志中新文化运动思想与立场的历史定位
- ☐ 《建康实录》文本挖掘及六朝人物关系时空分析
- ☐ 《拉贝日记》中《日本士兵在南京安全区的暴行》文件文本挖掘处理
- ☐ 明清时期文献中的色彩知识探秘
- ☐ 南京市政府出让住宅用地及其低价的时空建模与可视化研究
- ☐ 中国历代任务传记资料库
- ☐ 识别古书中隐藏的社会偏见
- ☐ 宋元学案学术传承数据分析

1.1.2. 赛事页面

- ☐ 赛事页面链接：

  https://www.heywhale.com/home/global?search=%E6%95%B0%E5%AD%97%E4%BA%BA%E6%96%87

1.1.3. 项目周期

- ☐ 筹备：2021-8-20 至 2021-9-30，共 50 天
- ☐ 初赛：2021-9-1 至 2021-10-28，共 58 天

1.2. 目标达成情况

本次比赛邀请全球数字人文研究者和爱好者参加竞赛，并组织专家对竞赛结果进行评选，最后邀请获奖者参加此次数字人文大会发表论文并领取奖励。本次大赛鼓励海内外数字人文研究者及数据分析爱好者利用各种数字人文新技术对开放数据

进行具备人文性的探索研究与应用。

- 目标人群：在校及在职的数字人文研究者和爱好者
- 8 个赛道总报名人数：432（人）
- 8 个赛道总团队数：356（支）
- 8 个赛道总计有效提交：33 份（预提交+25 号提交+28 号提交）
- 在校生占比：56.3%
- 在职人员占比：43.7%
- 海外参赛人数：7（人）

1.3. 和鲸本次工作内容

本次比赛，和鲸主要作为赛事平台提供方协助主办方进行赛事系统搭建与配置、赛事全流程把控&运营、赛事宣推、赛事重要通知发放等相关工作。

- **大赛全流程竞赛系统全流程的支持：**包括如报名、信息收集、实名认证、组队、主观提交、主观评委与评审系统、知识库配置等
- **大赛目标人群宣传推广与定向邀请：**如推文与软文撰写、和鲸自有公众号发布、和鲸社区用户邮箱触达、外部合作公众号投放、重点用户定向邀请参赛等
- **大赛全程的赛事运营工作：**如信息与通知发布、参赛人员问题解答、重点问题与反馈整理、社群维护、讨论区维护、定期选手问题与体验调查问卷、小礼品发放、相关数据分析、作品监控与筛查等

1.4. 项目人员

1.4.1. 和鲸项目团队

部门	职位	职责范围
项目管理	项目经理	项目统筹管理、进度管理、风险管理等
	运营经理	参赛人员运营工作、重要事项通知发放
产品与技术	产品总监	竞赛系统支持与维护
品牌	文案策划	大赛相关文案撰写与渠道发布

2. 项目成果

2.1. 筹备工作

这次赛事和鲸赛事团队主要支持主办方进行赛事系统配置与测试、赛事前期与过程中宣推渠道投放规划与和鲸自有宣推渠道投放、赛程运营过程中产品功能指引和具体答疑和赛后关键数据和信息的统计与整理。

2.2. 宣传与报名

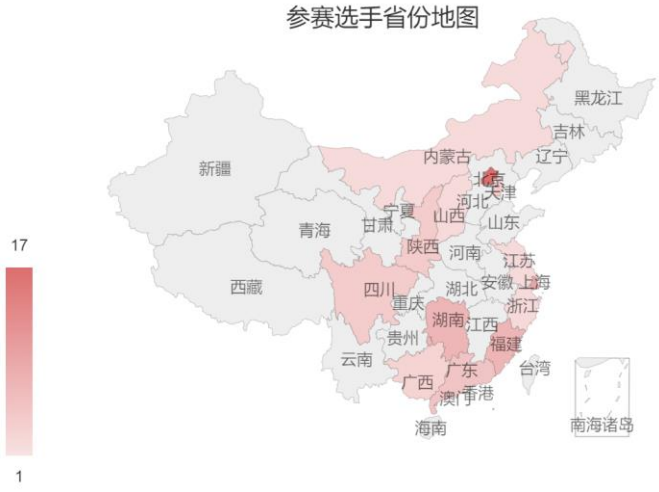
2.2.1. 报名情况概述

本次比赛 8 个赛道总计共吸引 432 位参赛选手参与，共计 356 支参赛团队，113 次提交，其中有效提交共 33 份。

2.2.2. 报名选手情况分析

□ 报名选手地域分布

参赛选手来自中国 53 个省市，top 5 的省市为北京市、广州市、武汉市、南京市和上海市。部分选手来自其他 5 个国家（美国、南非、日本、新加坡和印度）。



□ 在职/在校比例

在校生共 277 人（占比：56.3%）、在职人员共 155 人（占比：43.7%）。

□ 参赛选手学校分布

参加本次大赛的学生选手来自全世界各地的合计 79 所高校，其中参赛人数最多的五所国内高校为暨南大学（47 人）、武汉大学（18 人）、中国人民大学（18 人）、华中科技大学（12 人）和安徽财经大学（11 人）。

学生选手人数最多的 3 所海外高校为 National University of Singapore（新加坡国立大学 1 人）、Simon Fraser University（西蒙菲莎大学大学 2 人）和 Syracuse University（雪城大学 1 人）。

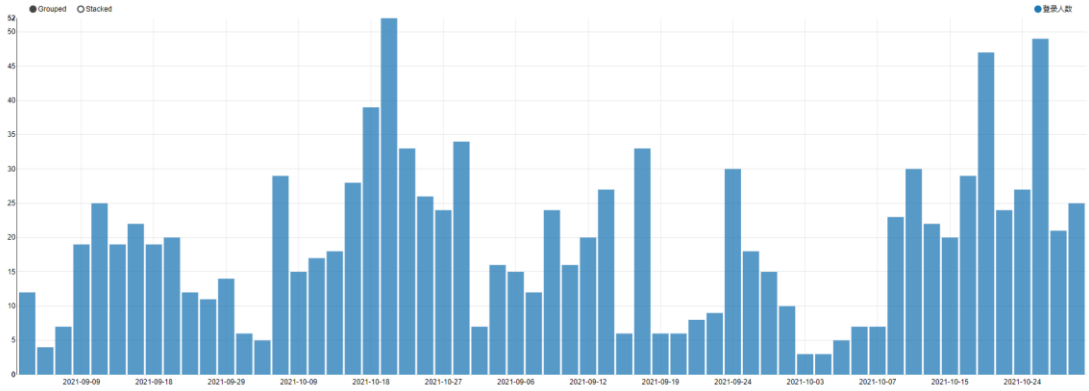
□ 在职选手单位分布

参加本次大赛的在职人员共来自业界 40 余家公司，以厦门华禹智能研究院有

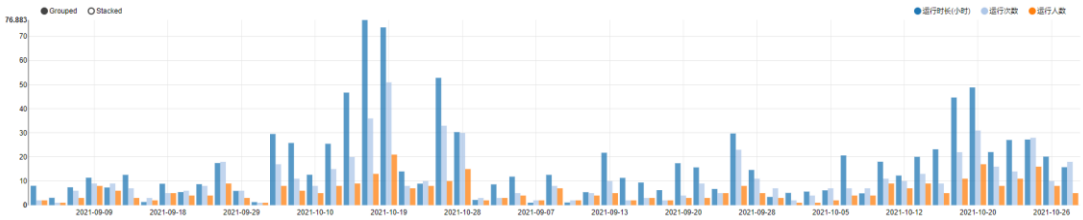
限公司、中国科学院计算技术研究院、北京科码先锋互联网技术股份有限公司、软通动力信息技术（集团）股份有限公司和中通服创发科有限责任公司等企业为本次在职人员参赛的主力军。

2.2.3. 数据分析平台 ModelWhale 使用情况

- **云计算资源使用情况**：总使用人数：105（人）（总人数 432 人，占比 24.3%）；总使用次数：633（次）（人均使用次数 6.03 次）
- **运行情况**：总运行时长：1167.5（小时）（人均使用时长 11.02 小时）
- **平台登录情况**：总登录人数：276（人）（总登录次数 1130 次，人均登录次数 4.09 次）赛中阶段 10 月 19 日登陆人数达到峰值（52 人），临近赛事提交截止前夕选手频繁登录 ModelWhale。



登录情况柱状图



运行情况柱状图

- **用户运行时长 top10**

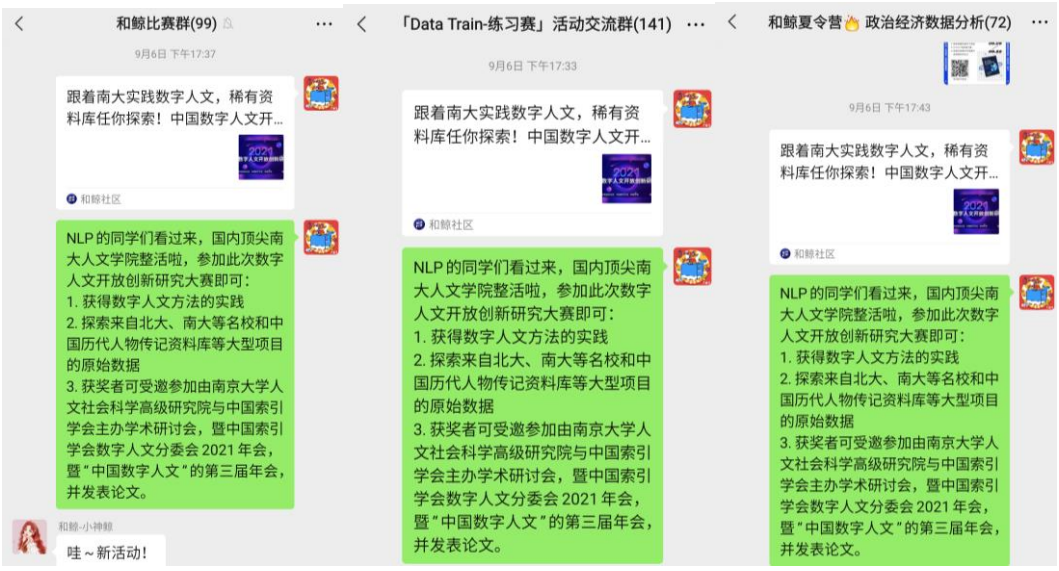
用户名	运行时长（小时）	是否提交有效作品	获奖情况	参与赛道
dfqqfb	183.5	是	二等奖	《拉贝日记》中《日本士兵在南京安全区的暴行》文件文本挖掘处理
wengmj	112.5	是	三等奖	《拉贝日记》中《日本士兵在南京安全区的暴行》文件文本挖掘处理

hooooon	83.5	是	/	南京市政府出让住宅用地及其地价的时空建模与可视化研究
宇宙的尽头是python	81	是	/	南京市政府出让住宅用地及其地价的时空建模与可视化研究
surwin	74.5	是	三等奖	宋元学案学术传承数据分析
年轻人不讲武德	48	是	二等奖	南京市政府出让住宅用地及其地价的时空建模与可视化研究（二等奖） 保守或融贯：重审《学衡》杂志中新文化运动思想与立场的历史定位
任寅瑞	43	是	/	南京市政府出让住宅用地及其地价的时空建模与可视化研究
西瓜憨熊	38	是	/	保守或融贯：重审《学衡》杂志中新文化运动思想与立场的历史定位
Mxlast	31	是	/	中国历代人物传记资料库
672859433	25.5	是	/	《拉贝日记》中《日本士兵在南京安全区的暴行》文件文本挖掘处理

2.2.4. 宣传投放回溯（和鲸自有渠道）

渠道类型	投放内容	阅读量/预计触达	投放链接/物料（备注）
和鲸社区公众号	开赛推文	2,256 阅读量	https://mp.weixin.qq.com/s/3EEXcUe2i-rxcB5J7CCZJQ
	赛题解析+直播推广	320 阅读量	https://mp.weixin.qq.com/s/paWUwpyrP_TcfL_yq1IJaw
定向邀约邮件	赛事延期提交通知	90.28%打开率	223 打开量/247 发送量
站内信	赛事延期提交通知	100% 成功发送	432 位报名选手
赛事群落	开赛推文+直播推广	10,000+（人）	微信群 50+，共宣推 2 轮

- 部分赛事群落宣推截图



2.3. 赛群运营

- **官方交流 QQ 群**（1003036870，8 个赛道一个群）

- **官方决赛通知微信群**

2.4. 最终获奖名单

南京大学人文社会科学高级研究院数字人文创研中心与和鲸科技联合举办的“中国数字人文开放数据创新研究大赛”提交、评审环节已圆满完成，非常感谢所有专家、学者和青年同仁的踊跃参与！本次竞赛共收到来自全国各地的 356 支队伍的提交项目 113 份，其中有效提交 33 份。经过 6 位评审的预先评审和 7 位评审的最终评分，去掉一个最高分、去掉一个最低分，取余下的平均分，按分数高低排名，共评选出 10 支获奖团队，名单如下：

奖项	团队名	赛题名	最终成绩

一等奖	一头倭瓜	《拉贝日记》中《日本士兵在南京安全区的暴行》文件文本挖掘处理	92.2
二等奖	年轻人不讲武德的团队	南京市政府出让住宅用地及其地价的时空建模与可视化研究	90.4
二等奖	datapie	南京市政府出让住宅用地及其地价的时空建模与可视化研究	88.4
二等奖	dfqqfb 的团队	《拉贝日记》中《日本士兵在南京安全区的暴行》文件文本挖掘处理	87.2
三等奖	wengmj 的团队	《拉贝日记》中《日本士兵在南京安全区的暴行》文件文本挖掘处理	87
三等奖	太热施肥伏特的团队	明清时期文献中的色彩知识探秘	87
三等奖	胡雪颖的团队	宋元学案学术传承数据分析	86.6
参加多个赛道，取最好成绩	年轻人不讲武德的团队	保守或融贯：重审《学衡》杂志中新文化运动思想与立场的历史定位	86.4
三等奖	挖掘机大队	识别故事中隐藏的社会偏见	85.8
三等奖	毕之爱吃鱼的团队	宋元学案学术传承数据分析	85.6
三等奖	yomorning 的团队	中国历代人物传记资料库	85.2

2.5. 优秀作品集锦

赛道名称	奖项	团队名
《拉贝日记》中《日本士兵在南京安全区的暴行》文件文本挖掘处理	一等奖	一头倭瓜

（一）文本概况

“我这不是想说这里的形势不严峻，形势的确严峻。不仅很严峻，而且会变得更加严峻。那么怎样才能对付这种严峻的形势呢？我认为，应当拿出自己的最后一份幽默，正视自己悲寂的命运。” ——《拉贝日记》

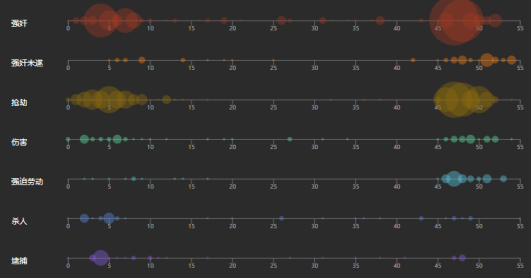
一位历史学家曾估算，如果所有南京大屠杀的罹难者手牵手站在一起，这一队伍可以从南京绵延到杭州，总距离长达200英里左右。他们身上的血液总重量可达1200吨，他们的尸体则可以装满2500节火车车厢。

而杀戮仅仅只是开端，强奸、抢劫、放火、掠夺，无时无刻在当时的南京发生。哪怕是由外国人建立的安全区，亦不安全。从拉贝提交给法庭的《日本士兵在南京安全区的暴行》，可见一斑。

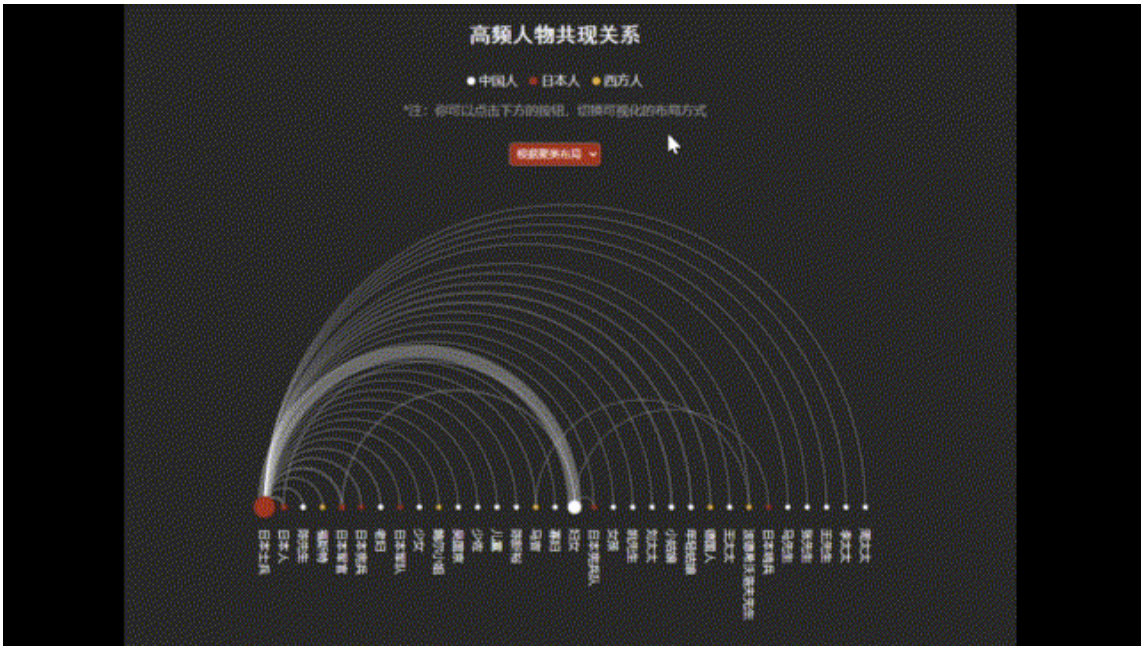
其全文约4.2万字，共计426条数据，从西方人的视角记录了发生在安全区中日本人的种种行为。这些冰冷数据背后，所代表的则是无恶不作的残暴，以及血流成河的悲痛。通过对上述文本分词，我们共得到11573个词语（包括重复词语），以及3610个不重复的词语。其中，出现频次最高的词语是“日本士兵”，共出现468次，在整个文本中的频率大约为4%。其他高频词汇还包括强奸（176次）、妇女（163次）、姑娘（118次）等。

记录从1937年12月15日开始，结束于次年2月6日，历经50余天。拉贝于1938年2月被召回，离开南京，故《拉贝日记》关于南京大屠杀的时间截点至2月26日。

各类型事件发生频次及趋势



接着，我们用堆叠可视化，来展示这些事件累加的结果。可以看到，拉贝文本中记述的事件，在前期和后期确实呈现出两个波峰，且这两个波峰对应的事件类型也是最丰富的。总体而言，强奸和抢劫占据着绝对的主导。

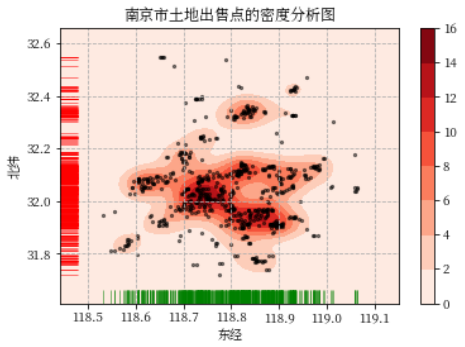


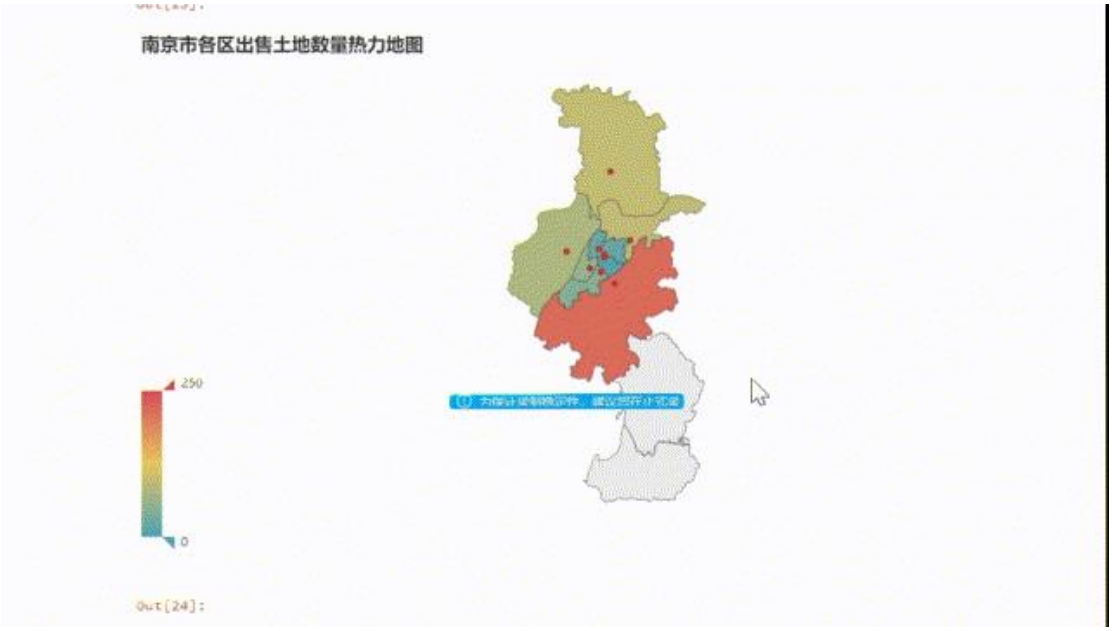
赛道名称	奖项	团队名
南京市政府出让住宅用地及其地价的时空建模与可视化研究	二等奖	年轻人不讲武德的团队

对出售点的密度进行分析。可以看出，南京市的出售地的分布具有一定的密集属性，会在某区域中出售点密度大，地理分布不均衡。

Out[17]:

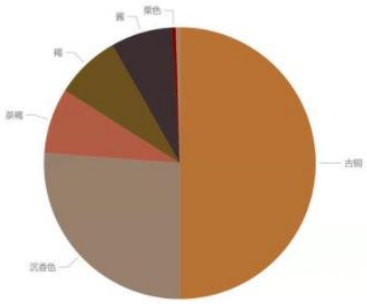
```
Text(0.5, 1.0, '南京市土地出售点的密度分析图')
```

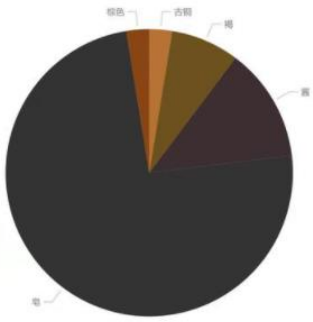




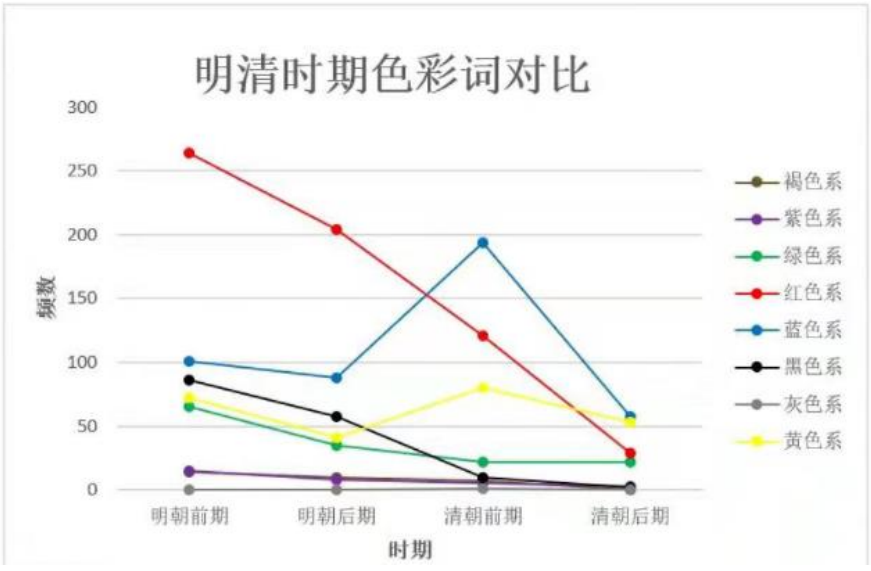
Out[24]:

赛道名称	奖项	团队名
明清时期文献中的色彩知识探秘	三等奖	太热施肥伏特的团队





（图为明代常用色彩词云）



3. 项目主要问题及改进建议

3.1. 赛题方向前沿创新，参赛选手技术水平较为薄弱，参赛门槛较高

- **问题描述：**由于人文社科研究方向较为前沿创新，大多数参赛选手为文科背景缺乏一定的数据分析能力，少数选手甚至第一次使用 Python 进行数据分析，导致参赛门槛较高，选手的赛事体验一般。
- **解决方案：**赛中特别邀请老师对各赛道进行赛题解析，组织与举办赛题解析直播，特设互动环节，选手与赛题解析老师进行互动，加深选手对赛题的理解。赛题解析均以视频方式与赛群内沉淀便于选手翻阅。
- **未来改进建议：**
  - 赛题筹备时提供更多赛题解析方面的帮助，如：赛题解析文档、赛题讲解直播、邀请优秀参赛团队破题分享。
  - 数据分析平台方提供更多的引导，主要针对第一次使用平台的用户给到更多支持。
  - 对于基础较为薄弱的同学，和鲸社区活动专区定期举办各类训练营及练习赛欢迎大家参与。

3.2. 数据种类过多、格式不同

- **问题描述：**由于数据过大、数量过多或格式特殊等问题，部分赛道的数据集在上传后难以直接在数据分析平台查看和运行，给选手的使用带来不便。
- **解决方案：**用压缩包的形式上传数据集，并在赛事页面注明，选手可将数据下载到本地后自行处理。
- **未来改进建议：**尽可能统一数据格式，统一以数据集形式挂载在赛事组织环境内，便于后续选手的读取与使用。

3.3. 因提交版本问题导致评审过程繁复

- **问题描述：**选手因提交不及时、提交了错误版本或是误删提交版本等原因造成无效提交，导致评审过程繁琐，增大评审老师的工作量，也影响评比的公平性。
- **解决方案：**询问无效提交团队后，将其信息汇总，作为特殊提交单独列出，并注明“请老师们酌情减分”。评审表发送后再提出修改或补交的团队不再予以接收。
- **未来改进建议：**在提交截止前提醒选手检查自身提交版本，并且不要在后续对此版本进行删改操作，以免出现版本失效的问题。

3.4. 前期提交作品数较少

- **问题描述：**数据分析比赛区别于一般算法赛，整体评审维度和提交物都

较为主观，无法获得及时分数反馈，故导致选手参与感较弱，大部分选手压线提交且提交作品质量较难管控。

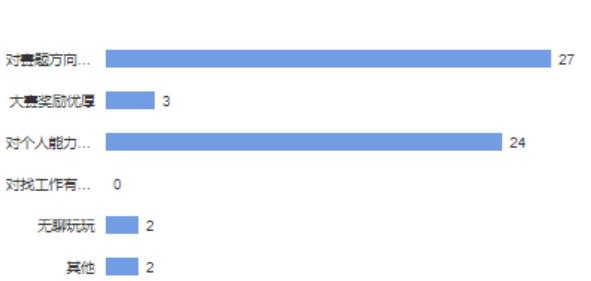
- **解决方案：**以预先提交并给到老师反馈的形式鼓励和刺激选手对目前作品预先提交，通过这种方式了解比赛选手的参与度情况、提交作品数及作品质量情况。
- **未来改进建议：**未来筹备类似数据分析类赛事时，预留充裕的时间对赛制进行合理地设计和规划，并配合运营手段（破题讲解直播、预先提交）吸引更多选手参与。

4. 选手反馈

本次问卷共收回 30 位选手的反馈，均为有效问卷。

4.1. 参加比赛的目的

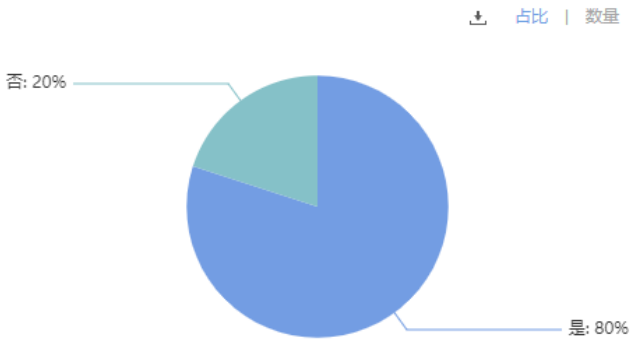
- 90%选手表示对本次赛题方向/研究领域很感兴趣
- 80%选手也希望能够通过此次本赛提升自己的个人能力



选项	数量	占比
对赛题方向和此领...	27	90.0%
大赛奖励优厚	3	10.0%
对个人能力成长有...	24	80.0%
对找工作有帮助	0	0.0%
无聊玩玩	2	6.7%
其他	2	6.7%

4.2. 比赛的最大收获

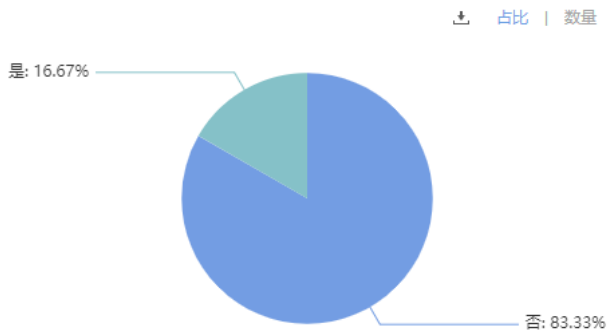
- 80%选手是第一次接触人文类选题，选手能够通过此次比赛将数字技术与人文类学科相结合，开拓了研究视野，是一次极有益的尝试。



选项	数量	占比
是	24	80.0%
否	6	20.0%

4.3. 比赛体验相关

- 83.33%选手表示是第一次参与数据竞赛，不仅在参赛过程中提高了自己的科研能力，也加强了团队协作能力。

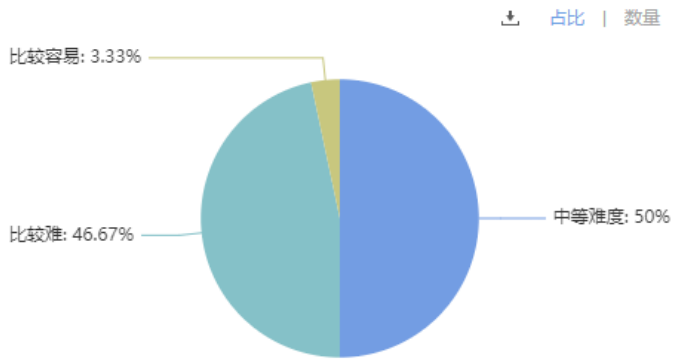


选项	数量	占比
是	5	16.7%
否	25	83.3%

- 沟通方面，主办方工作人员能够及时在社群中解答选手疑惑，提高了选手的参赛体验，也为和鲸运营方提供了很多帮助，感谢！
- 80%的选手对比赛数据分析平台（Model Whale）使用体验打到 6 分及 6 分以上，大部分选手为平台打 8 分。

4.4. 遇到的问题

- 对于第一次参加此类数据竞赛的学生，略显复杂的相关指导会使参赛门槛较高，近 96.7%的选手认为比赛赛题难度较大。同时大部分选手也是第一次使用和鲸的工作台，操作不是特别熟练。



选项	数量	占比
非常难	0	0.0%
比较难	14	46.7%
中等难度	15	50.0%
比较容易	1	3.3%
非常容易	0	0.0%

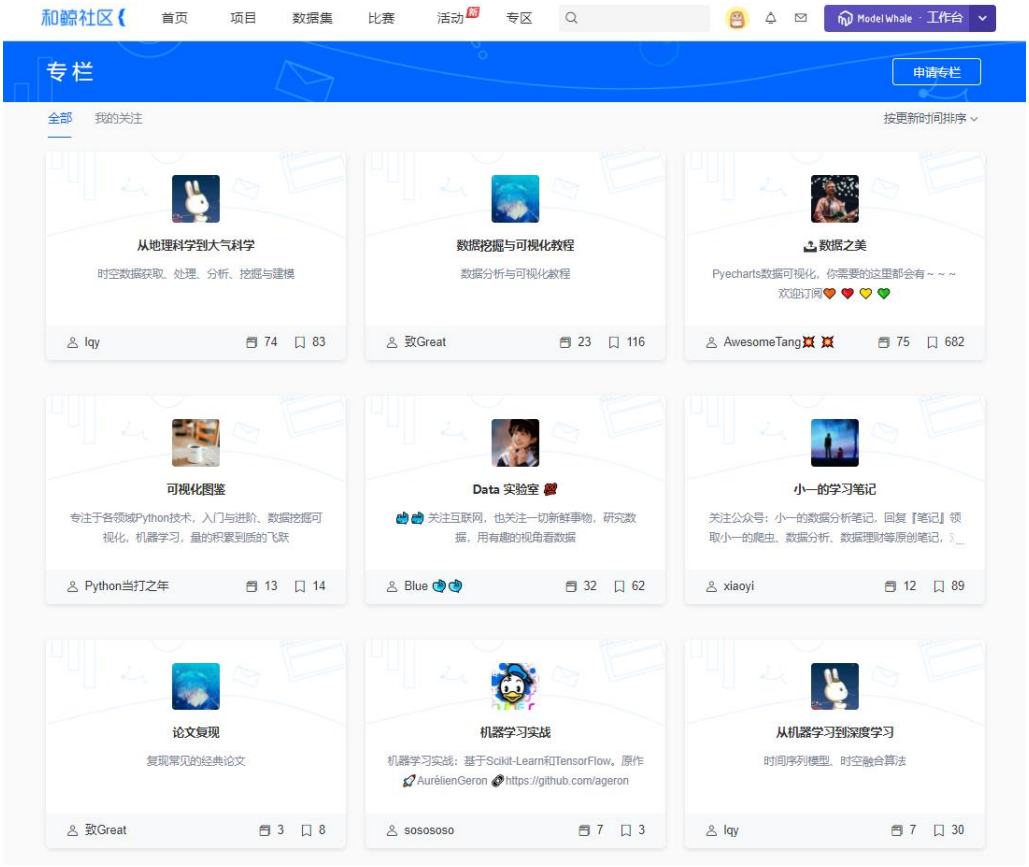
4.5. 未来的建议

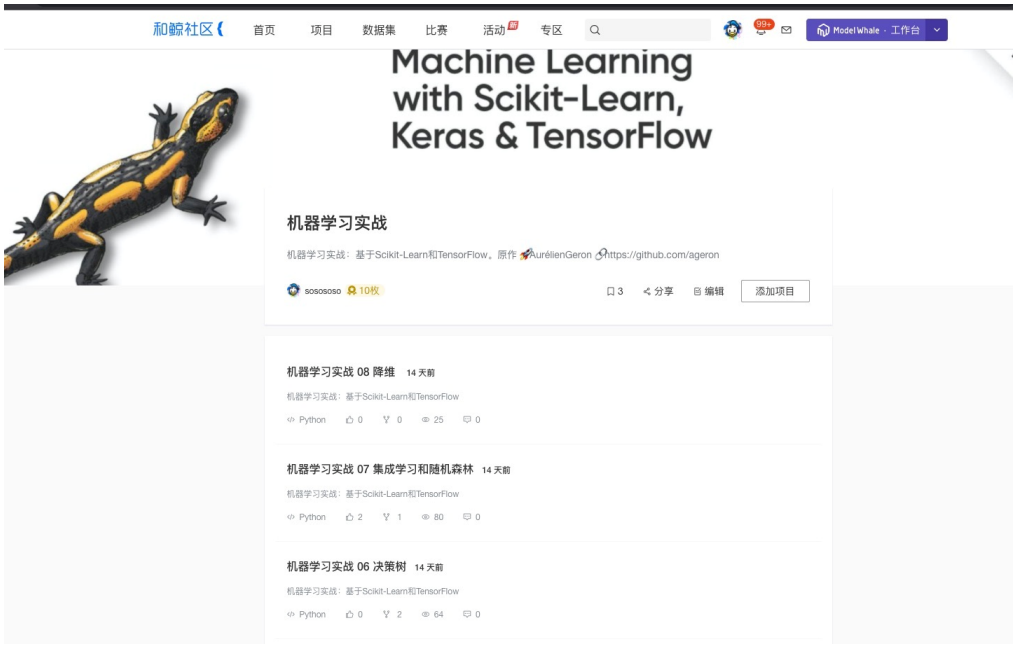
- 和鲸会尽可能提供更多的工作台上手引导，不断完善产品，优化选手的参赛体验。
- 希望主办方可以在之后的比赛中提供更加丰富和简单的赛题指导，帮助学生打开思路，完成比赛。

6. 后续工作安排

6.1. 沟通优秀作品如何沉淀及具体方式

- 建议以【专栏】形式沉淀本次比赛优秀作品，该专栏后期可持续沉淀其他南京大学人文社科高级研究员数字人文创研中心的优秀数据分析成果物。（专栏展示效果如下）
- 【专栏】需要一位专门的管理员，进行专栏内容运维。管理员享有该专栏的编辑、上传、更新等权限。





特别鸣谢

特别感谢南京大学人文社科高级研究院数字人文创研中心各位老师，感谢各方的赛题与数据提供方，感谢参与作品评审的各位专家老师。感谢赛事运营管理人员。