# 第二届全国大学生开源情报数据采集与分析大赛参赛方案

## 引言

本次大赛分为初赛和决赛，初赛由开源数据收集、数据采集规则制作、数据分析报告三个比赛环节，各环节所占分数权重依次为10%、40%、50%。根据三个环节总成绩进行排名，排名前60%的参赛队晋级决赛。决赛为PPT汇报及答辩环节，采用集中式进行，团队总成绩=初赛成绩\*30%+决赛成绩\*70%，最终获奖名次根据团队总成绩进行统计排名。

从比赛内容上来看，各个环节层层递进，充分扎实的完成相应环节的任务有助于完成下一个环节的任务。在接下来的内容中，我将结合评分标准和项目任务，详细阐述我们的比赛方案。

## 初赛环节1

### 评分标准：

对提交的数据集文件进行复验，数据量超过5万条，且数据重复率低于1%，数据空值率低于1%的参赛队伍得满分。若总数据量低于5万条，则按实际比率得分（如4万条得8分，精确到小数点后两位）。若总数据量高于5万条，但数据重复率或数据空值率任意一项高于1%，则每多50条扣0.5分，最高扣5分，不满50条按50条计算。提交数据时间不影响本轮成绩。

### 项目任务：

利用八爪鱼等数据采集工具采集完整不重复数据5W条。可选择的数据类型如下表，每个团队应至少选择两种数据类型，数据量要求≥5万条，数据重复率不得高于1%，数据空值率不得超过1%。

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **数据类型** | **参考网址** | **字段** |
| 1 | 新闻数据 | 人民网、环球网、中国新闻网以及地方新闻网等 | 关键字或版块、标题、原文链接、发布时间、正文 |
| 2 | 社交平台数据 | 微博、知乎、股吧、雪球等 | 关键字或版块、发帖链接、发帖人、发布时间、帖子正文、评论数(如有)、点赞数(如有) |
| 3 | 政策数据 | 外交部、国务院新闻办、地方政府网等 | 关键字或版块、标题、原文链接、发布时间、正文 |
| 4 | 短视频平台数据 | 抖音、小红书、快手、B站等 | 关键字或版块、发帖链接、发帖人、发布时间、帖子正文、评论数(如有)、点赞数(如有) |
| 5 | 地图数据 | 百度地图、高德地图、腾讯地图等 | 关键字或版块、地标名称、详细地址 |
| 6 | 电商评论数据 | 京东、淘宝、唯品会等 | 商品名称、商品链接、评论人、评论内容、评论时间、评分或星级（如有） |

### 比赛方案：

**初赛环节1攻略：数据采集，拒绝“偏科”！**

初赛环节1看似简单，实则暗藏玄机。对于经常爬数据的参赛同学来说，我们知道不同平台、不同类型的数据爬取难度可谓天壤之别，而新闻数据的爬取无疑是其中最容易的。于是，一个“取巧”的策略浮现眼前：在5万条数据中，大部分采用新闻数据，仅留一小部分填充其他类型。

然而，比赛主办方早已预料到这种“偏科”行为，并巧妙地在初赛环节2设置了“数据采集规则制作”，这意味着你需要对八爪鱼软件了如指掌。而与其在环节2临时抱佛脚，不如在环节1就将其作为最佳练手机会！

基于此，我们团队制定了以下比赛策略：

**1. 全面覆盖，拒绝“偏科”：** 三人分工合作，每人负责一种类型数据，确保涵盖所有类型，避免单一类型数据占比过高。

**2. 精选平台，事半功倍：**针对每种类型数据，选择一个最具代表性且易于爬取的平台，集中火力攻克。

**3. 八爪鱼为主，拒绝“炫技”：** 放弃使用Python等爬虫框架，专注于八爪鱼软件的使用，将更多精力投入到数据清洗和分析中。

**4. 自建验证程序，确保数据质量：**根据评分标准，编写一个简单的验证程序，用于检查爬取数据的完整性、准确性和一致性。通过自建验证程序，可以及时发现并修正数据采集过程中的错误，避免在后续环节中出现问题。

总而言之，初赛环节1不仅是数据采集的战场，更是磨练数据采集技能的绝佳机会。拒绝“偏科”，全面发展，才能在后续比赛中游刃有余，取得佳绩。

## 初赛环节2

### 评分标准：

每个任务5-12分，一共5个任务，以团队为单位分工完成。组委会对提交的每个任务规则进行校验，能正常运行且能采集到样例数据≥100条，样例数据字段须与赛题完全一致，样例数据前100条空值率不得高于10%，有任何一项不符则该任务不得分。分数相同的，则提交时间早的排名靠前。

### 项目任务：

利用八爪鱼等数据采集工具完成赛题要求的数据采集规则制作，制作完成后导出规则文件及样例数据提交。

* 选手下载数据采集工具。
* 根据赛题任务制作对应的采集规则文件，顺序不限；
* 完成某一赛题规则后进行试采集，采集样例数据符合赛题要求后导出样例数据文件和任务规则文件（otd后缀文件）。
* 命名规则：导出文件按照题目1.otd，题目1.xls方式命名，压缩包命名为参赛队编号+初赛环节2。

### 比赛方案：

在初赛环节1，我们团队就展现出了强大的协作能力和对工具的熟练运用。我们深知，数据采集规则的制作是数据爬取的关键，而八爪鱼软件则是我们手中的利器。

**1. 分工协作，全面覆盖**： 我们三人分工明确，每人负责不同类型数据的采集规则制作。我们精选了最具代表性的平台，集中火力攻克，确保每种类型数据都能高质量地采集。

**2. 实战演练，精益求精：** 在完成初赛环节1的过程中，我们不断进行实战演练，模拟各种可能出现的情况，并及时调整规则，确保其稳定性和可靠性。

**3. 交叉验证，提升鲁棒性：** 为了进一步提高我们制作出的采集规则的鲁棒性，在制作完采集规则之后，我们会安排一位同学在另一台电脑上运行导出的.otd文件，进行交叉验证。通过这种“双保险”的方式，我们能够及时发现并修正潜在问题，确保采集规则在不同环境下都能稳定运行。

得益于初赛环节1的充分准备，我们在赛时能够游刃有余地应对各种数据爬取需求。我们只需根据题目给出的明确需求，让相应的队员快速完成任务即可。这种高效的协作模式和强大的技术储备，为我们最终取得的优异成绩奠定了坚实的基础。

## 初赛环节3

### 评分标准：

选手提交数据进行分析的方案报告,其中：分析工具、思路和方法10分，分析过程15分，分析结果和结论15分，对策建议10分。提交报告由多位专家评委盲测评分，取专家评分平均分为最终得分。分数相同的，则提交时间早的排名靠前。

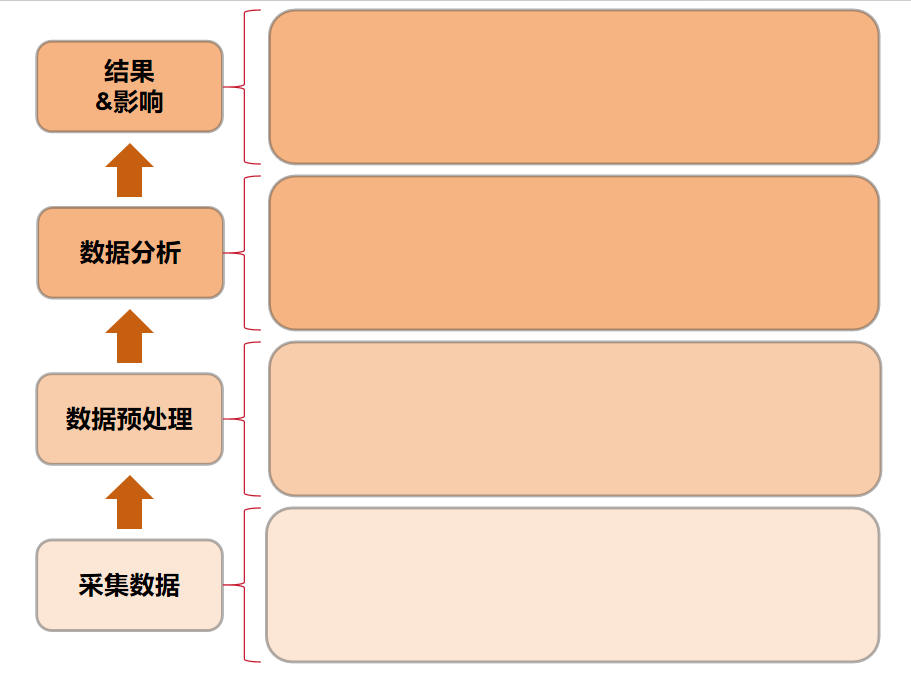
### 项目任务：

选手对所采集的数据进行分析并形成方案报告, 方案报告需要能够准确、清晰描述整个分析过程和分析结果，包含：使用的分析工具、分析思路、分析方法；分析过程、分析结果及结论；提出有针对性的对策和建议。

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **选题方式** | **题目或方向** | **详细内容** |
| 1 | 开放式选题 | 社交媒体情报分析 | 通过采集社交媒体平台上的用户发贴数据，收集和分析用户发布的信息，了解公众对特定话题或事件的意见和情感倾向，以帮助政府、企业等了解公众舆情和市场动向。 |
| 2 | 网络犯罪情报分析 | 通过分析犯罪组织在互联网上的宣传、招募和策划活动，获取有关其组织结构、领导层、战术等方面的情报，以便制定反恐策略和行动。 |
| 3 | 金融情报分析 | 通过分析金融市场的公开信息，如公司财报、股价、市场交易等，获取有关公司、行业和市场趋势的情报，以支持投资决策和风,险管理 |
| 4 | 商业情报分析 | 通过采集商业公司的公开信息，如新产品发布、市场营销活动、人才招聘等，了解其战略意图和动向，以改进自己的产品和服务。 |
| 5 | 网络安全情报分析 | 通过采集互联网上的激进言论、潜在网络威胁活动，获取有关潜在网络威胁的情报，以加强网络安全防御和事件响应能力。 |
| 6 | 教育资源情报分析 | 通过采集网络上教育资源的公开信息，比如采集各个地区教育资源（学校、师资、教学设施等）信息，采集就业市场信息，分析各个地区资源配置的情况、人才培养需求和就业市场供需关系等。 |
| 7 | 政府治理情报分析 | 通过采集政府相关数据进行数据分析，了解政府治理情况，以支持政策效果评估、公共服务优化、政府决策支持等。 |
| 8 | 医疗健康情报分析 | 通过采集医疗健康领域的数据进行数据分析，以支持疾病流行趋势分析、不同地区的医疗资源分配优化、个性化医疗服务等。 |
| 9 | 固定式选题 | 主要竞争国对华投资情况分析 | 日本，美国，德国，英国及台湾地区在内地投资控股企业信息采集和分析（参赛团队选一个国家或地区进行采集分析），根据采集数据完整度及分析的深度进行评分。 |
| 10 | 低空经济领域无人机与载人低空城际通航的数据洞察与未来趋势分析 | 利用开源数据对全球或特定地区的无人机市场进行分析，要求参赛者通过数据挖掘揭示行业动态，评估潜在危险因素，并提出建设性策略和建议，同时预测市场发展障碍以及未来发展趋势。 |
| 11 | 2024年国际竞争新形势分析 | 结合2023-2024年主要竞争国（如美国、日本、德国、英国等任选一个或多个国家）对华经济、文化、立法、科技等方面的数据进行采集和分析，并预测未来可能产生的影响和趋势。 |
| 12 | 南海战略分析 | 分析南海声索国政府和民间对中国的舆论态度，评估认知战在南海争议中的作用和影响，结合分析结果提出促进南海地区舆论理解和认知战态势改善的策略。 |
| 13 | 亚洲军事演习的社交媒体评价分析与区域安全态势研究 | 采集2023-2024年亚洲军事演习相关数据，分析评估军事演习对区域安全态势的影响，探索社交媒体在塑造公众对军事和安全议题认知中的作用。 |
| 14 | 全球南方国家的多边主义合作分析 | 采集金砖国家、非洲联盟等组织以及社交媒体相关的数据，分析全球南方国家在多边主义合作中的作用和贡献，提出建设性的策略和建议以促进全球南方国家通过多边合作应对全球性挑战。 |
| 15 | 全球范围内极端天气和自然灾害带来的全球气候危机分析 | 采集全球范围内关于自然灾害（如极端天气、洪灾、旱灾、地震等）以及受灾数据。对灾害发生频率、影响范围、受灾人口、经济损失、难民流动趋势等进行分析，提出针对自然灾害的预防、减灾、应急响应和长期解决方案，并结合国际合作、地区发展、社会治理等方面，提出政策建议，以提高社会对自然灾害的抵御能力。 |

### 比赛方案：

基于选题《2024年国际竞争新形势分析》，我们在团队内进行了充分讨论，并初步将工作划分为数据采集和数据分析两个部分。



围绕这些工作内容，我们制定了如上图所示的四个步骤流程：采集数据、数据预处理、数据分析，以及结果与影响分析。

具体来说，我们分配了两位同学负责采集数据和数据预处理，另外一位同学负责查找相关文献资料并构思数据分析的方式。我们约定了一个具体的时间来规定完成这两个任务的期限。

在第二次碰面时，团队提出了新的思路。负责数据采集的同学介绍，他们已收集到部分官方新闻数据和社交媒体微博数据，这些数据大多是静态的。然而，如果时间允许，我们完全可以采集动态数据，比如每小时跟踪某条新闻的转发量、评论数和点赞数，通过动态数据的拟合，能够更深入地挖掘信息。

负责数据分析的同学指出，纯人工的数据分析不仅费时费力，且难以准确把握不同时间节点的数据变化趋势。借助大模型的能力，利用其处理和分析数据的优势无疑是一个替代方案。于是，我们设计出一个可以自动检测国际竞争新形势的平台。

在这个平台上，首先，我们编写了自动化爬虫程序，每天定时抓取相关数据，并将数据更新到数据库中。接着，我们调用大模型，对数据库中的内容生成分析报告，并根据报告为主要竞争国家在各领域的竞争力评分，实现动态监测的功能。考虑到比赛要求提交一份静态分析报告，我们基于一段时间的数据生成了静态报告，最终获得了不错的成绩，这也在一定程度上验证了自动化框架的有效性。

通过以上流程的设计和实现，我们不仅为团队任务提供了系统的支持，还为后续动态监测国际竞争形势的新形势奠定了基础。

## 决赛

| **评分内容** | **评分项目** | **评分要点** |
| --- | --- | --- |
| **内容质量与相关性**  **（15分）** | 紧扣主题 | 汇报内容是否紧密围绕比赛要求的主题，明确阐述分析目标。 |
| 数据准确性 | 所使用的数据是否准确无误，分析方法是否科学合理。 |
| 相关度与深度 | 分析是否深入，是否有独到见解或创新的方法应用。 |
| 结论与建议 | 提出的结论是否基于数据分析得出，建议是否具有实践价值和可操作性。 |
| **PPT可视化效果与呈现**  **（15分）** | 逻辑结构 | PPT布局是否清晰，逻辑条理是否顺畅。 |
| 视觉效果 | 图表、图形的选择与设计是否有助于信息传达，整体视觉是否专业吸引。 |
| 专业性和创造性 | PPT的设计是否专业，是否展现了创意。 |
| **演讲表达**  **（10分）** | 表达能力 | 演讲者是否能够清晰、准确地传达信息。 |
| 自信和仪态 | 演讲者是否展现出自信，是否有良好的肢体语言和眼神交流。 |
| 时间管理 | 是否在规定的时间内完成了演讲。 |
| **创新能力**  **（10分）** | 创新性 | 汇报或分析是否包含独到的见解和创新的方法。 |
| 结论应用价值 | 分析结果是否具有实际应用价值和解决问题的潜力。 |
| **综合素质与团队协作**  **（10分）** | 团队协作 | 团队成员之间是否有明确的分工，是否能够协作无间。 |
| 临场应变 | 面对意外情况时的处理能力，如技术故障、突发提问等。 |
| **答辩能力**  **（10分）** | 问题回答的准确性 | 对于评委的问题，回答是否准确、深入。 |
| 思维敏捷性 | 是否能够迅速理解问题并作出回应。 |

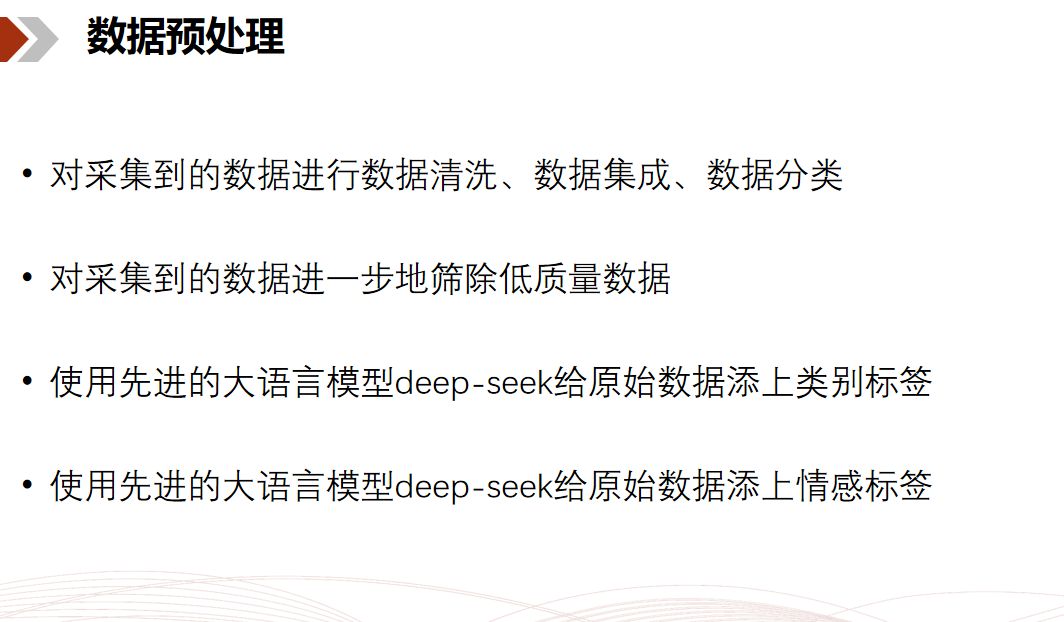
### 评分标准：

### 比赛方案：

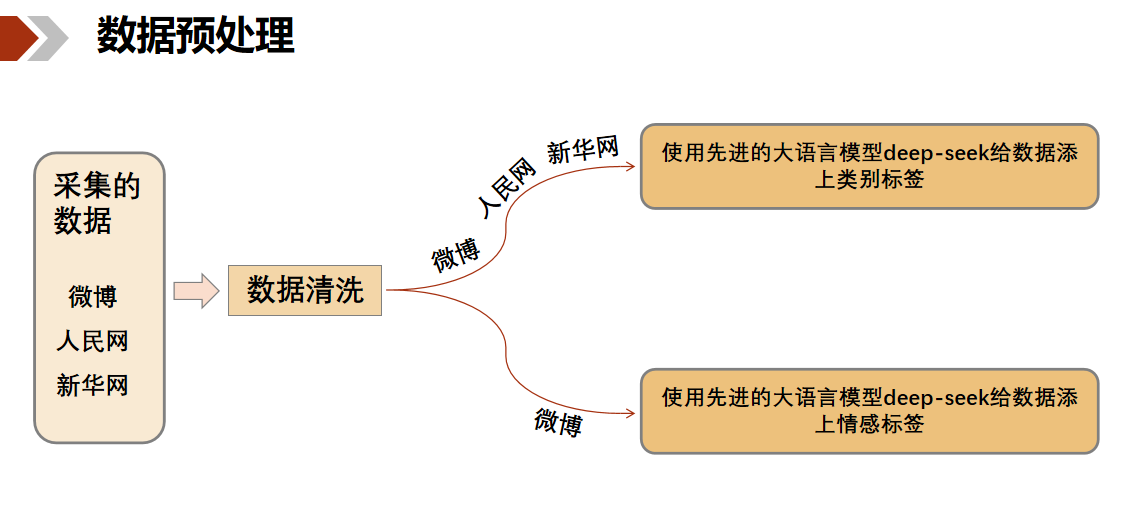
在决赛中，我们的汇报内容主要是对初赛环节3提交的报告进行PPT精简汇总。考虑到每组只有五分钟的展示时间，我们将PPT页数控制在10页左右，力求简洁明了。在队内的多次演讲练习后，我们最终选定了一名声音洪亮、表达清晰的队员负责正式汇报。

在准备过程中，我们总结出一些心得。首先，评分标准中演讲表达的占比不低，因此表达流畅和逻辑清晰尤为重要。我们在排练时特别注意语言的流畅性和逻辑的衔接，以确保信息传达准确。同时，从评委的反馈中发现，技术层面的问题往往是提问的重点。因此，对自己所使用的技术需有深入理解，避免使用一些似是而非的“高端”术语，以免在提问时显得难以自圆其说。

在PPT制作方面，我们强调了可视化的重要性。文字较多的PPT页通常难以吸引观众的注意，因此我们在内容展示时尽量以流程图、示意图等直观的图形代替长段文字。例如，初版PPT中的数据预处理步骤如下图：



而修改过后的流程图如下:



修改后的PPT采用了流程图的形式展示数据预处理的流程，使得信息流动一目了然，从“采集的数据”到“数据清洗”，再到不同的数据处理分支，形成了直观的视觉路径。这种图示方式比修改前PPT的纯文字呈现更能帮助观众快速理解每一步的处理过程。

## 总结

在本次比赛中，我们团队围绕数据采集和分析任务，制定了详尽的方案，并通过实践不断完善。针对比赛环节的要求，我们不仅高质量完成了数据采集和清洗工作，还使用了先进的大语言模型进行标签化处理，增强了数据分析的深度。通过多次演练，我们在PPT展示和表达上达到了最佳效果，顺利在比赛中脱颖而出。这次参赛经历不仅帮助我们掌握了数据处理和分析的实战技能，也提升了团队协作能力，为今后的科研与工作奠定了坚实的基础。