第三届开源情报数据采集分析大赛初赛第二环节赛题

赛前须知：

比赛时间：10 月 12 日晚上 19:00--21:00

提交邮箱：osintda@163.com

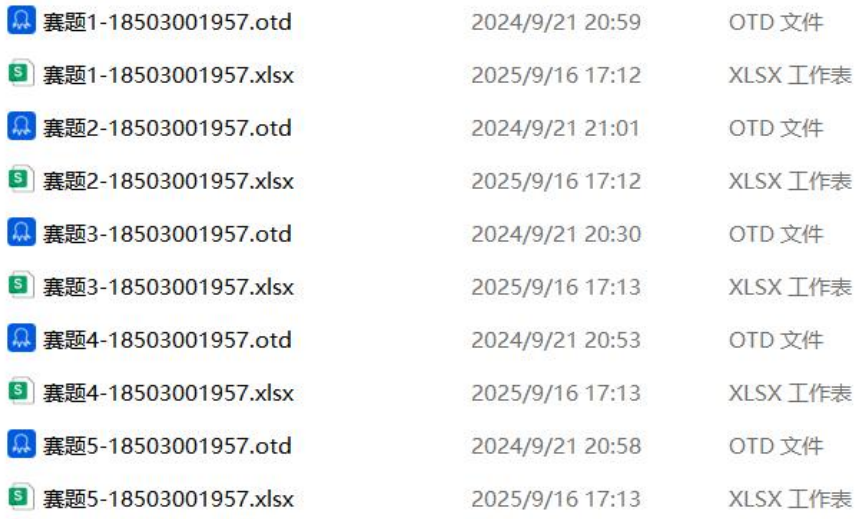
提交截止时间：10 月 12 日晚上 21:00,21:00 以后提交的数据不计入成绩。

要求：在规定时间内提交数据规则文件（otd 文件）和样例数据文件（Excel 文件），可以网上查找的资料，队伍之间不允许交流，所有赛题只验证前 100 条数据，采集规则需至少保证前 100 条数据不能出现漏数据情况，任务本身不够 100 条数据的验证所有数据，赛题要求的字段必须包含，字段顺序不作要求。部分数据需要通过格式化功能清洗数据，请仔细审题。（采集时遇到验证码可暂停采集手动跳过验证码后继续采集。）

otd 文件导出方式如下：



提交文件命名规则如下（后缀为制作规则文件所用的八爪鱼采集器账号，若溯源后该规则不是该账号编写，则成绩无效）：



压缩包命名规则如下（邮件主题与压缩包命名一致）：



正式比赛赛题：

赛题一（6 分）

要求：

1.采集网站：https://wecruit.hotjob.cn/SU642fbf5fbef57c1e269fa798/pb/social.html

采集中国信通院社会招聘数据

2.采集字段：职位名称、工作地点、所属部门、职位类别、招聘人数、发布时间、学历要求、工作经验要求、专业要求、工作职责、任职要求、详情链接。

说明：只采集前 10 页数据



赛题二（9 分）

要求：

1. 批量采集网站：

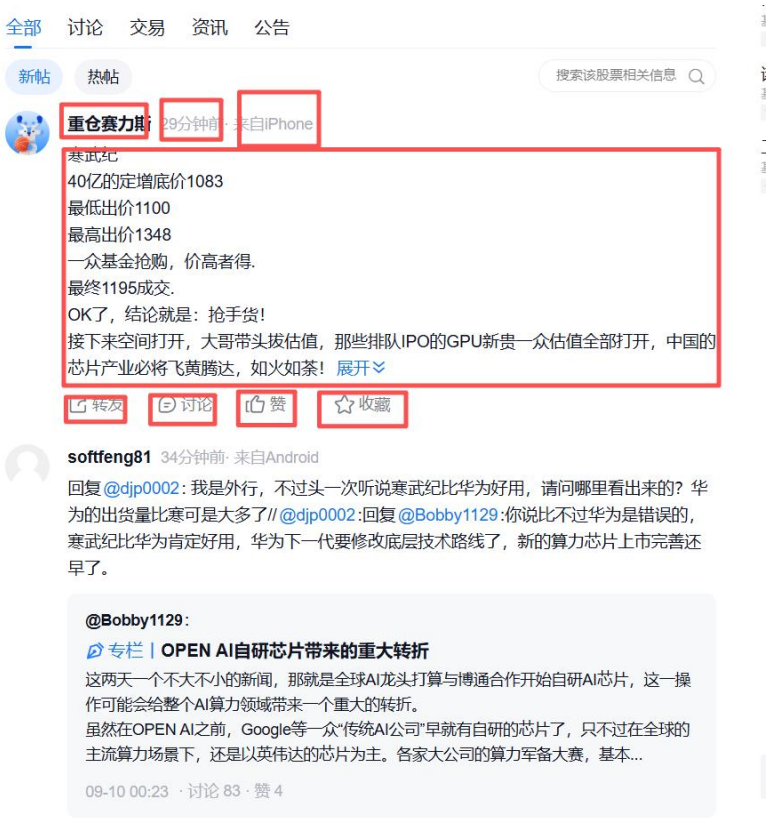
https://xueqiu.com/S/SH688256

https://xueqiu.com/S/01810

2.采集字段：股票信息、关注股票人数、发帖人昵称、正文、发布时间、发帖设备、转发数、评论数、点赞数、收藏数。

说明：正文需要采集全文，不得截断。





赛题三（9 分）

要求：

1.采集网站：

https://so.toutiao.com/search?dvpf=pc&source=input&keyword=%E5%BC%80%E6%BA%90%E6%83%85%E6%8A%A5&pd=information&action_type=pagination&page_num=0&from=news&cur_tab_title=news

3.采集字段：关键词、标题、发布时间、来源、正文、正文图片链接、评论数。

4.**只采集 2025 年发布的咨询。2025 年 1 月 1 日前发布的不采集**。

5.正文较长需设置滚动网页。

说明：没有评论的文章需默认为 0，正文如有多张图片也需要采集，统一采集到一个字段。



目前中国在远程战略预警雷达技术领域属于全球一流，中国在掌握了远程战略预警雷达和全套反导体系以后，意味着中国已经不再惧怕美国的核威慑，成为捍卫我们国家安全和权益的坚强盾牌。

举报

评论 47

请先 登录 后发表评论～

赛题四（8 分）

要求：

1. 批量采集网站：

https://bbs.hupu.com/634822124.html

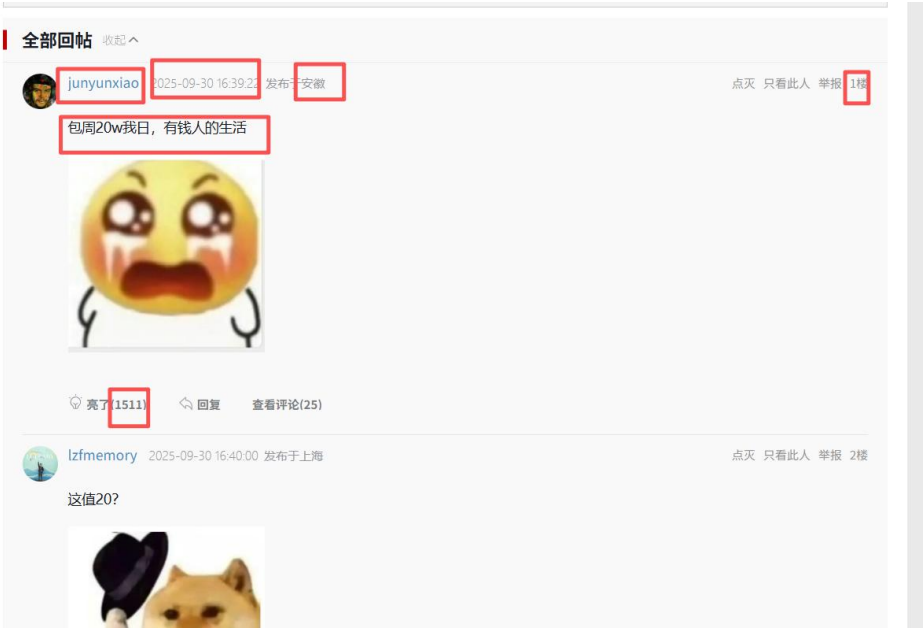
https://bbs.hupu.com/635004754.html

2. 采集全部数据

3. 采集字段：专区名称、帖子名称、帖子链接、回复数、亮数、浏览数、回帖人、发布时间、发布地点、回帖楼层、回帖内容、亮了数。

说明：回复数、亮数、浏览数、回帖楼层、亮了数均为纯数字，没有默认为 0。





赛题五（8 分）

要求：

1. 批量采集网站：

https://weibo.com/2656274875/Q76eKiq89

https://weibo.com/1699432410/Q6M9e2v0P

2. 采集两篇博文中的所有一级评论。

3. 采集字段：原文链接、发帖人、发布时间、正文、关注数、粉丝数、微博数量、评论人、评论时间、评论地点、评论内容、点赞数。

说明：只采集一级评论、评论加载方式为滚动加载。

