Vup 项目笔记

个人研究 不太有参考性 有点主观

因为之后准备自己做 VUP，同时给我这资深的 dd 生涯来点意义，准备收集 vup 数据研究一下什么样的 vup 比较受欢迎

该笔记目前包含两部分内容，第一部分是个人收集数据与分析的尝试。第二部分是一个对冰火晚会的分析项目（但失败了）

个人收集数据与分析

B 站对 api 管可严了，一天只让爬一次好像。所以，在收集信息时候，走了很多“弯路”。最后突然发现，不对啊有人做了这个项目：

https://vtbs.moe/

所以，朋友们，文献综述的重要性。

不写文献综述盲目开干简直是愚蠢至极。

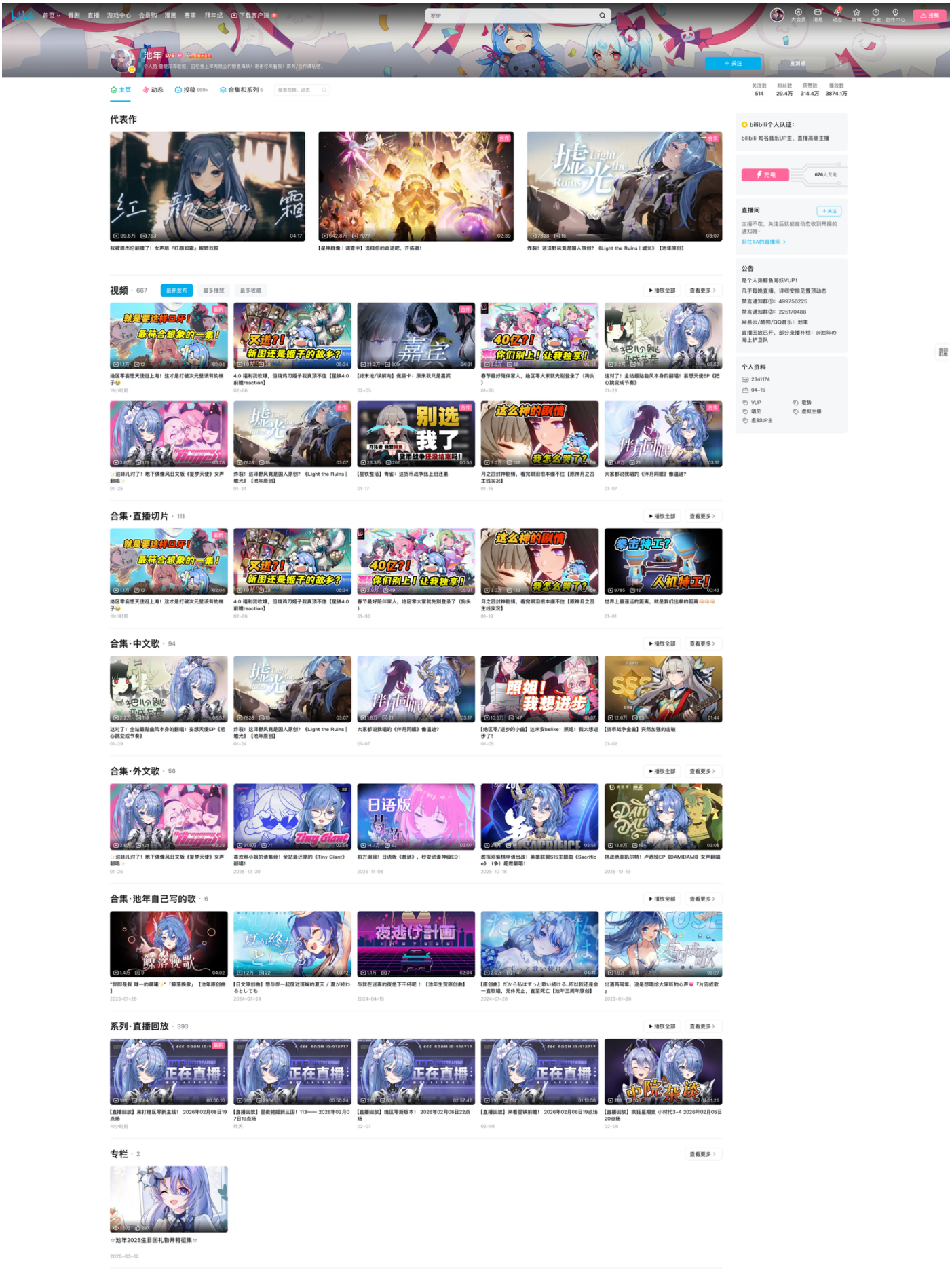
但是，还是有价值的，因为这个项目的精度没我自己做的这个高。但是，我项目充满了过于“人工”的痕迹，不适合大规模使用。

详细收集方式：

因为反爬管的很严，所以其实运用的方式和人力收集信息一摸一样。

最开始的路径：创建一份 uid.txt，然后编写脚本使用 chrome 开发版配合 playwright 插件打开文件中每个 uid 对应的 up 主空间（这一步要使用我自己的登陆 cookie，本质上就是自己账号上去看了一下），强制加载全部内容并放大，然后截图一张全图。之后对图中信息，使用 OCR 识别转文字，再放进本地部署的 Deepseek-R1 转成可靠，合理无乱码的表格。转换是一定必要的因为 ocr 只输出文字和位置，并不理解文字对应内容。

示例，截出来的图长这样：



然后我说，不对啊，这里面这些图片太干扰了。也正是因为图片存在会让 ocr 识别出乱码，才需要扔给 ds 再清理。能不能把图片去掉？结果 playwright 插件有这个功能。同时它还支持放大显示，所以我调了一下倍率。截图示例

池年

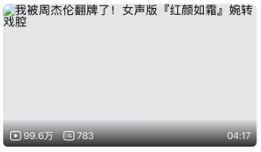
LV8

个人势 曾是深海歌姬，因怕鱼上岸再就业的鲸鱼海妖！谢谢你来看我！商务/合作请私信。

+ 关注　发消息

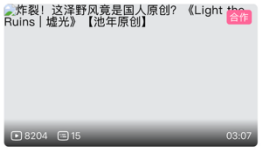
关注数	粉丝数	获赞数	播放数
514	29.4万	314.5万	3878.6万

代表作



我被周杰伦翻牌了！女声版『红颜如霜』婉转戏腔

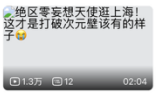
【星神群像丨调查中】选择你的命途吧，开拓者！



炸裂！这泽野风竟是国人原创？《Light the Ruins | 墟光》【池年原创】

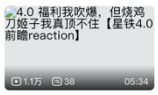
视频 · 667

最新发布　最多播放　最多收藏　播放全部　查看更多 >



绝区零妄想天使逛上海！这才是打破次元壁该有的样子

02-08



4.0 福利我吹爆，但烧鸡刀姬子我真顶不住【星铁4.0前瞻reaction】

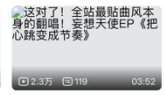
02-06

【终末地/误解向】佩丽卡：原来我只是嘉宾

02-05

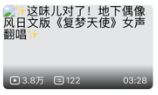
春节最好陪伴家人，绝区零大家就先别登录了（狗头）

01-30



这对了！全站最贴曲风本身的翻唱！妄想天使EP《把心跳变成节奏》

01-29



✨这眯儿对了！地下偶像风日文版《复梦天使》女声翻唱✨

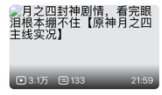
01-25

炸裂！这泽野风竟是国人原创？《Light the Ruins | 墟光》【池年原创】

01-24

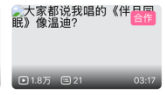
【星铁整活】青雀：这货币战争比上班还累

01-17



月之四封神剧情，看完眼泪根本绷不住【原神月之四主线实况】

01-16

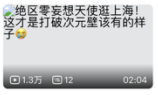


大家都说我唱的《伴月同眠》像温迪?

01-07

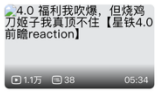
合集·直播切片 · 111

播放全部　查看更多 >



绝区零妄想天使逛上海！这才是打破次元壁该有的样子

02-08

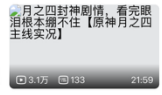


4.0 福利我吹爆，但烧鸡刀姬子我真顶不住【星铁4.0前瞻reaction】

02-06

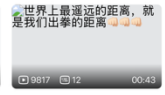
春节最好陪伴家人，绝区零大家就先别登录了（狗头）

01-30



月之四封神剧情，看完眼泪根本绷不住【原神月之四主线实况】

01-16

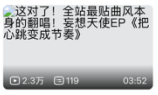


世界上最遥远的距离，就是我们出拳的距离👊👊👊

01-01

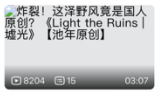
合集·中文歌 · 94

播放全部　查看更多 >



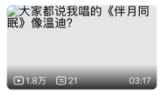
这对了！全站最贴曲风本身的翻唱！妄想天使EP《把心跳变成节奏》

01-29



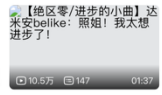
炸裂！这泽野风竟是国人原创？《Light the Ruins | 墟光》【池年原创】

01-24



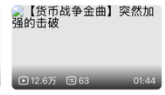
大家都说我唱的《伴月同眠》像温迪?

01-07



【绝区零/进步的小曲】达米安belike：照姐！我太想进步了！

01-05

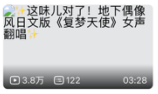


【货币战争金曲】突然加强的击破

01-02

合集·外文歌 · 56

播放全部　查看更多 >

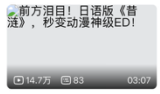


✨这眯儿对了！地下偶像风日文版《复梦天使》女声翻唱✨

01-25

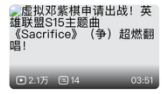
喜欢照小姐的请集合！全站最还原的《Tiny Giant》翻唱！

2025-12-30



前方泪目！日语版《普涟》，秒变动漫神级ED！

2025-11-08



虚拟邓紫棋申请出战！英雄联盟S15主题曲《Sacrifice》（争）超...

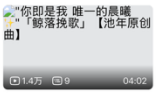
2025-10-18

挑战绝美凯尔特！卢西娅EP《DAMIDAMI》女声翻唱

2025-10-16

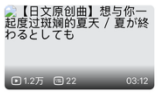
合集·池年自己写的歌 · 6

播放全部　查看更多 >



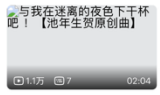
"你即是我 唯一的晨曦✨"「鲸落挽歌」【池年原创曲】

2025-01-26



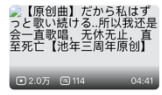
【日文原创曲】想与你一起度过斑斓的夏天 / 夏が終わるとしても

2024-07-24



与我在迷离的夜色下干杯吧！【池年生贺原创曲】

2024-04-15



【原创曲】だから私はずっと歌い続ける..所以我还是会一直歌唱，...

2024-01-26

出道两周年，这是想唱给大家听的心声💗『片羽成歌』

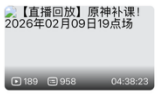
2023-01-26

系列·直播回放 · 395

播放全部　查看更多 >

【直播回放】过绝区零的活动！ 2026年02月10日18点场

7小时前



【直播回放】原神补课！ 2026年02月09日19点场

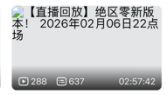
昨天

【直播回放】来打绝区零新主线！ 2026年02月08日19点场

02-09

【直播回放】星夜驰援新三国！ 113—— 2026年02月07日19点场

02-08



【直播回放】绝区零新版本！ 2026年02月06日22点场

02-07

专栏 · 2

查看更多 >

☆池年2025生日回礼物开箱征集☆

2025-03-12

---

bilibili个人认证：

bilibili 知名音乐UP主、直播高能主播

充电　676人充电

直播间　+关注

主播不在，关注后就能在动态收到开播的通知哦~

前往TA的直播间 >

预约

直播预约：米游合家欢歌回

今天 19:00直播 359人已预约

由预约有奖：100星琼*5份

预约

公告

是个人势鲸鱼海妖VUP!

几乎每晚直播，详细安排见置顶动态

禁言通知群①：499756225

禁言通知群②：225170488

网易云/酷狗/QQ音乐：池年

直播回放已开，部分录播补档：@池年の海上护卫队

个人资料

2341174

04-15

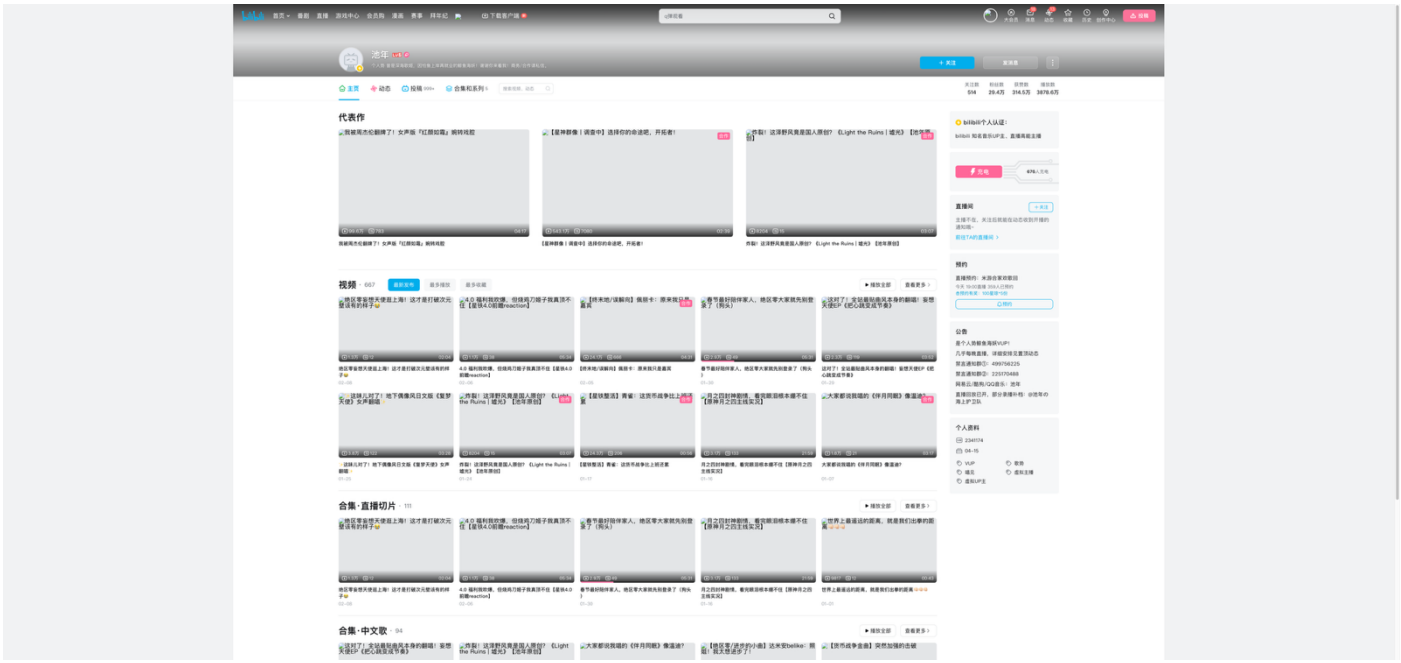
VUP　歌势

唱见　虚拟主播

虚拟UP主

返回旧版

我发现，这样的截图清晰度足够，且各个信息位置固定。所以直接在 ocr 识别时候给它写好区域，识别出来结果放到对应区域就行了。同时，我意识到，很多信息都不是经常更新的，所以我可以少截很多内容。



日常只用截取一张头部图片取上面“播放量”“粉丝数”两个关键数据就可以作为日常更新了。

但是，这个方法也有致命漏洞。每张截图固定需要时间 15s（避免 b 站反爬）。意味着一小时最多获取 240 份数据，这太少了。除非我多弄几个 b 站账号和电脑同时刷—但这也太愚蠢了。

这时候我才想起来，说不定有人已经做了类似项目呢？一搜，我人傻了。

我的项目仅剩的优势在于它可以绝对稳定的获取信息，不受 b 站影响；以及可以通过截图分析最近一个投稿视频（位置也是固定的，最新视频的第一个）与播放量的关联—长期来看就是可以分析运行期间每一个视频对播放量的影响。但这也不准确，因为 vup 基本都直播。直播获得的粉丝比较难衡量；同时这里还得引入一个概念“直播切片”。认识一个 vup 有时候都不是它本人视频认识的，而是发在别的账号的切片认识的。

总之，我的项目在收集数据效率与全面性方面完全不如找到的项目。估摸着只适合定性研究时候需要更高精度开来用一用。但是，已有的项目也不够，需要更好的数据。

我认为，为了完成我的 vup 研究，我需要 1. 更权威的身份去找 b 站要数据；2. 搭建并训练一个 ai，特化于虚拟主播数据收集（投稿和直播两方面抓：投稿方面重点要教会它识别“切片”概念，以及其他可能得会给一个虚拟主播带来粉丝的投稿；直播这边基于已有的 DD 监控室项目教它如何识别直播获得的粉丝，以及对舰团的分析。我感觉舰团这玩意水分很大，一堆企业势刷舰长挺明显的）

冰火晚会粉丝涨粉特征研究过程梳理

（基于 2026.02.08-09 冰火晚会期间多 UID 粉丝数据，聚焦 “抽奖粉丝” 与 “真实涨粉” 特征验证）

一、研究核心问题

围绕冰火晚会（所有 UID 参与）的粉丝量变化，验证 3 个关键假设，明确 “抽奖引流粉丝” 与 “真实留存粉丝” 的特征及影响因素。

二、核心概念定义（量化标准）

表格

概念	定义（基于 2026.02.08 12:00-02.09 13:00 数据）	计算逻辑
起始粉丝数均值	2026.02.08 00:00-11:59（晚会前）最近 3 条粉丝数的平均值	平滑单条数据偶然性，代表 “活动前基础粉丝量”
虚假涨粉量（FFG）	活动期内最高粉丝数 - 起始粉丝数均值	含抽奖关注的短期峰值涨粉
抽奖粉丝量（LF）	活动期内最高粉丝数 - 活动期内最低粉丝数	冲高后回落的 “抽奖取关粉丝”
真实涨粉量（TFG）	FFG - LF	活动结束后实际留存的有效涨粉
年涨粉速率	2025.02.09-2026.02.08（近 1 年）日均涨粉量	代表 “历史涨粉能力”

三、研究数据处理流程

1. **原始数据**：多 UID 粉丝数历史 JSON 文件（含时间戳、粉丝数等字段）；
2. **数据清洗**：
    - 筛选核心时段：2026.02.08 12:00-02.09 13:00（活动期）、2025.02.09-2026.02.08（年历史期）；
    - 计算衍生指标：按上述定义生成 FFG、LF、TFG、年涨粉速率等；
    - 过滤异常值：剔除 FFG≤0（无涨粉）、LF<0（无回落）、数据量 < 2 条的 UID；
3. **最终分析数据**：整理为 OD.csv，含 UID、各核心指标及辅助字段（如数据点数量、速率计算状态）。

四、3 个核心假设与验证逻辑

表格

假设（H）	假设内容	验证方法	支持 / 拒绝标准
H1	所有 UID 的 “抽奖粉丝量（LF）” 数值接近（同批抽奖粉）	描述性统计（变异系数 CV）+ 分布检验	CV<0.1（数值高度集中）→ 支持；反之拒绝
H2	真实涨粉量（TFG）与活动前 “历史粉丝基数” 无关	Spearman 相关性检验（适配非正态数据）	相关系数 <0.1 且 p>0.05（无线性关联）→ 支持；反之拒绝

假设（H）	假设内容	验证方法	支持 / 拒绝标准
H3	真实涨粉量（TFG）与“年涨粉速率”（历史能力）有关	Spearman 相关性 + 线性回归	相关系数 ≥0.2 且 p≤0.05（显著正关联）→ 支持；反之拒绝

五、验证工具与输出

1. **工具**：Python（数据清洗、指标计算）+ R Studio（统计检验、可视化）；
2. **关键输出**：
   - 统计结果：各指标均值 / 中位数 / 标准差、相关性系数、p 值；
   - 可视化图表：LF 分布直方图 + 箱线图（H1）、TFG 与历史基数散点图（H2）、TFG 与年速率散点图（H3）；
   - 结论报告：Hypothesis_Verification_Summary.csv，含每个假设的验证结论、有效样本量、图表路径。

六、研究核心目标

1. 区分 “抽奖短期引流” 与 “真实有效涨粉”，量化冰火晚会的 “粉丝质量”；
2. 明确影响 “真实涨粉” 的关键因素（是否与历史基数 / 历史涨粉能力相关）；
3. 为类似活动的 “涨粉效果评估” 提供可复用的量化方法（如 LF、TFG 的计算逻辑）

理想是这样的，但是现实很骨感啊，数据太少了。主要原因是获得数据的网站更新较慢。如果能精确到 10 分钟级别会好很多。

现在表格长这样

UID	2.8 中午前起始粉丝数均值	标准化起始基数（消除量级）	近 1 年日均涨粉速率（粉丝/天）	活动期最高粉丝数	活动期最低粉丝数	活动期最后粉丝数	Fake_Fan_Growth(FFG)	Lottery_Fa
11073	4927624.0	6.5151	446.38	4928290	4927770	4928290	666.0	520
510047	494695.0	0.6541	-48.07	503820	502824	502824	9125.0	996
922573	685661.0	0.9066	-13.5	685811	685700	685811	150.0	111
1950658	1119999.0	1.4808	82.52	1120073	1120035	1120073	74.0	38
3046429	2314596.0	3.0603	35.8	2314790	2314679	2314790	194.0	111
14387072	504353.0	0.6668	35.07	505491	505123	505123	1138.0	368
15641218	216279.0	0.286	58.53	216352	216279	216352	73.0	73
43272050	1127997.0	1.4914	484.01	1128258	1128072	1128258	261.0	186

73857541	54643.0	0.0722	56.38	54701	54643	54701	58.0	58
401480763	2164492.0	2.8618	855.53	2165268	2164549	2165268	776.0	719
484322035	1655801.0	2.1892	649.32	1656118	1655727	1656118	317.0	391
595407557	1891676.0	2.5011	1868.14	1893482	1891677	1893482	1806.0	1805
631070414	1539107.0	2.0349	463.32	1541214	1538808	1541214	2107.0	2406
672328094	1791772.0	2.369	226.5	1793011	1791861	1793011	1239.0	1150
672342685	925691.0	1.2239	585.91	927467	925825	927467	1776.0	1642
672353429	660110.0	0.8728	206.62	660162	659958	660162	52.0	204
698029620	2248781.0	2.9732	507.94	2250566	2248862	2250566	1785.0	1704
1217754423	195703.0	0.2587	54.83	195811	195775	195811	108.0	36
1878154667	217778.0	0.2879	56.16	217852	217778	217852	74.0	74
1900141897	195787.0	0.2589	59.09	196022	195849	196022	235.0	173
3546569288714792	976301.0	1.2908	1296.93	978317	976130	978317	2016.0	2187

乱到难以形容的数据，还少

我大概知道数字民族志为什么真的很好用了