📌 IC网爬虫需求文档- 0218

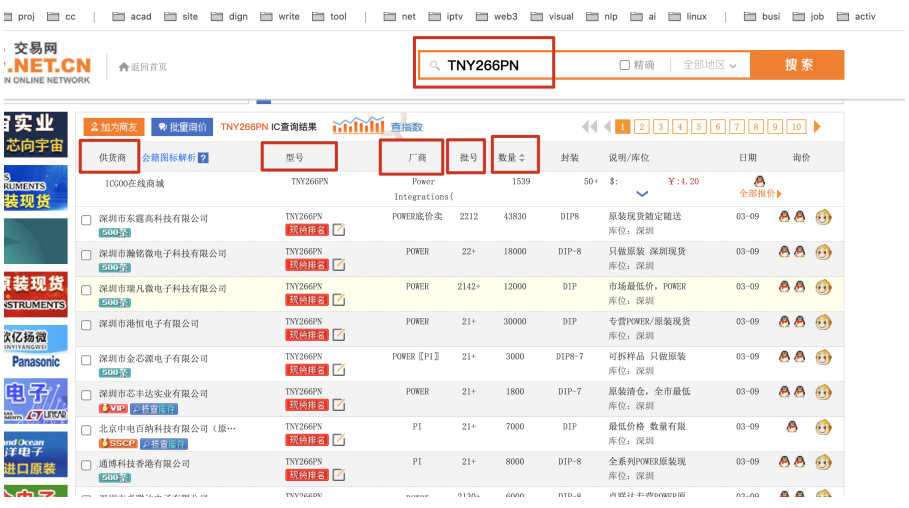
> 💡 **此网站封号很严重，我们现在也找不到规律，所以需要技术人员有耐心地进行测试，找到可行的策略 🙏**

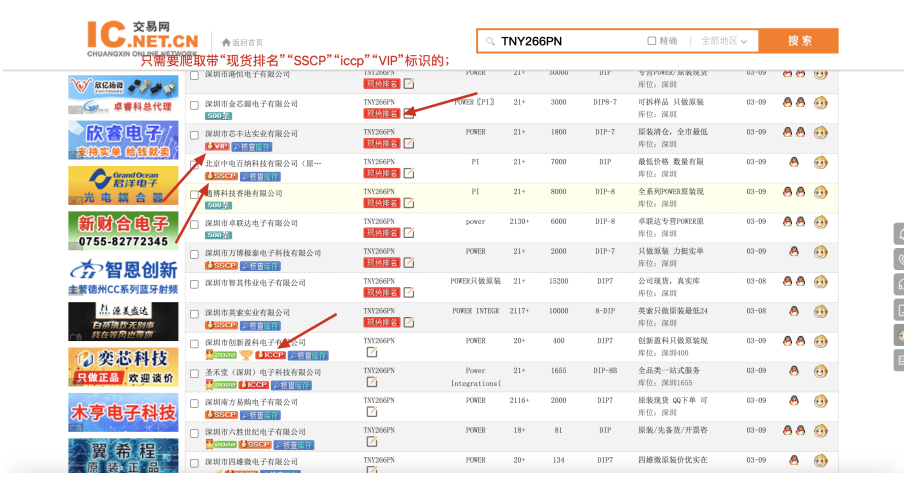
项目概述

- 本项目是一个爬虫采集数据需求，旨在从IC交易网站上爬取搜索型号、供货商、厂商、批号和数量等数据。
- 我们在此承诺所采集数据不会用于非法行为，只是用于数据分析，产出行业研究报告。
- 采用自动化的方式即可（需要保证效率）
- 测试数据（芯片型号）：https://images-ferry.oss-cn-shenzhen.aliyuncs.com/test-data.txt

爬虫目标

- 在IC交易网中，对芯片型号进行搜索并爬取数据。
- 这个平台一共有 4 个端的接口可以拿数据，数据都是一样的。（哪个端好爬就从哪个端取数据）
  - （1）H5移动端 https://m.ic.net.cn （从 H5 移动端取数据，因为请求量更少，会不会效率更高？）
  - （2）PC 移动端：https://www.ic.net.cn/
  - （3）微信小程序“IC 交易网”
  - （4）手机APP “IC 交易网”
- 需要的字段（每个型号只需要爬前面的 2 页就可以了）
  - 供货商（只需要爬取带“SSCP”“iccp”“VIP”标识的）
  - 搜索型号（只需要爬取带“现货排名”标识的）
  - 厂商
  - 批号
  - 数量
  - QQ（有多个的话，随机取一个）





爬虫流程

爬虫流程如下：

> 💡 **您这边只需要解决风控，保证拿到请求体即可**。我们这边已经有现成的数据清洗和入库逻辑的完整代码，后续可封转成接口给您直接调用入库。

1. 首先，从mysql的all_search_models 表中的取型号，现在一共有15w个型号数据，要求 10 天内跑完一轮。
2. 10 天一轮任务（定时任务），每轮任务都需要在 ic_batch_spider 表中创建一个新表。
3. 用 redis 来做队列和任务管理。

测试期间我们这边提供以下接口支持：

1. 接码平台的账号密码。
2. 从我们的代理池中获取动态 IP 的接口；（如果后续需要静态 IP 或者长效 IP 也可以提供 ）
3. 爬虫过程中会出现网易易盾的 「文字点选」 和 「空间推理」验证码，我们这边有模型，可以提供返回点击验证码坐标的接口(如使用selenium等，可能会有轨迹校验，需要自行解决)

必看的注意点！

- 使用自动化的方式进行爬虫，其实没有什么门槛。这个项目的关键点在于数据采集量较大，需要保证稳定+高效。所以请确定可以解决“稳定+高效“这个问题再接单！
- 采集数据量大的时候，可能会被平台监控到，然后无限封号（这是我们现在遇到的最大问题），所以重要的是测试出风控，给出请求的时候多 ip 和多账号的策略。
- 可能需要用到：多线程，异步，分布式来保证采集的速度。要平衡好效率：考虑需要的设备资源和爬取时间、爬虫成本；
- 长期进行数据分析，所以这边数据采集需求很多， 且此项目也需要长期维护，所以我们是奔着长期合作的目的，希望接单的技术：
  - 技术能力优秀，可以提供 github 链接 或者 过往爬虫项目的 demo代码。
  - 为人靠谱负责，好沟通，好合作。
- 此项目理想下 2 周左右完成，长期合作，预算可谈。能解决的大佬测试完成后，请带报价、时间排期和测试结果，加微信进一步沟通： 19902407422（也可直接打电话沟通），感谢！
- 为了不浪费彼此时间，非诚勿扰，谢谢！