**Redfin 爬虫存在的主要问题及解释**

“Redfin资料需求.xlsx” 表格中的数据可以分为两类，分别是**静态内容**和**动态内容**。静态内容是网页中直接包含在HTML文件中的内容，无需额外的用户交互或后续的脚本执行即可显示，页面加载后这些内容是固定的，可以直接解析。表格中以下内容为静态内容：

A screenshot of a computer

Description automatically generated

For Sale框内的Beds, Baths, Sq Ft, Est price.

A screenshot of a web page

Description automatically generated

About this home的内容

反之，动态内容是指需要通过 JavaScript 执行、用户交互（如点击或滚动），或异步请求（如 AJAX）从服务器获取并显示的内容。页面初次加载时，这些内容通常不直接包含在 HTML 中，而是由客户端（浏览器）在页面加载后生成或请求。表格中以下内容为动态内容：

A screenshot of a computer

Description automatically generated

Sale and tax history

A screenshot of a web page

Description automatically generated

Property details

如果用户不进行交互，如不点击“tax history”或不展开Exterior的模块，那么里面的内容就不会显示，这就要求爬虫具备网页交互的能力。

Redfin的爬虫协议支持获取以上的**静态内容**，但是会严格检测用户端浏览器的交互行为，如果用户存在快速交互和频繁开关浏览器等可疑行为，就会违反协议并被禁止访问所有网页的内容。

以下是我的爬虫遭到拦截返回的截图，实测间隔在1-5分钟的访问和交互均会被认定为恶意行为。

A screenshot of a phone

Description automatically generated

因此，如果要实现多个地址的数据爬取并规避上述检测行为，爬虫有以下要求：

1. 仅爬取静态页面的数据
2. 设置爬取间隔，也就是等待5分钟以上再进行下一个地址数据的爬取，或者在同一个页面下，每隔5分钟以上进行一次交互（如展开下拉目录）。