**HD竞品信息爬取和Carro前台信息爬取Q&A**

1. **推荐页面每次登入都会刷新，记录结果并不固定，需要如何确定提取的时间。**
2. 能否在美国消费者购物集中的时间段自动做2次爬取前台资料（能否不需要人工操作）：

西雅图时间：12：00-14：00和17：00-20：00这两个时间段；（可行）

1. 提取后的数据能否自动保存并上传至google云端（能否不需要人工操作），然后我们可以不定时人工下载；（可行）

解决方法：我会创建一个共享的Google Drive并把链接发到群里，每天按时更新表格，各位有需要可自取。

1. **HD的网页解析有问题，提取产品详细数据时会有报错的情况，是否需要人工操作。**

（1）请列一下具体哪些问题？（下划详见“问题”）

（2）另外我看到Ben导出的csv文档只有250个产品，我们手动数了一下，HD前5页应该一共有280-283个产品的。（已解决）

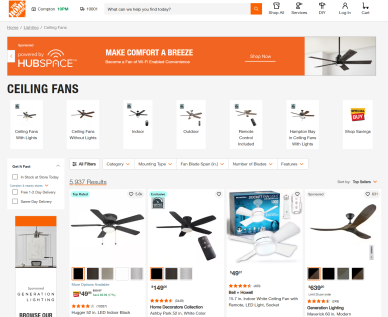
解决方法：页面加载过慢导致爬虫未能爬取每页最后五个产品，已通过调整等待时间解决。

（3）我已经在csv表格上新增了G,H,I列内容需求，还有将表头的顺序调整了一下。如附件一。（下划详见“问题”）

1. **筛选和排序，是筛选先还是排序先，是否过滤’sponsored‘和HD自家产品。**

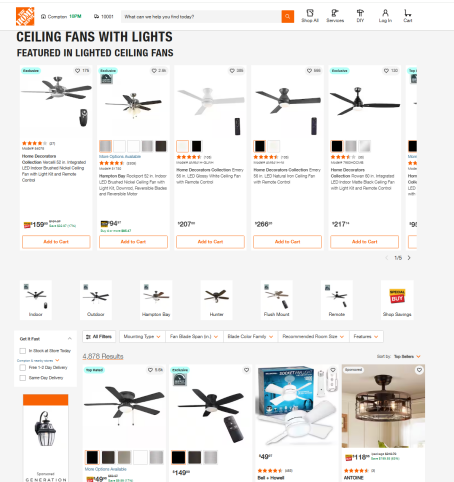
（1）以下为从homedepot.com前台自然搜索类目CEILING FAN后的结果(不需要筛选和排序，同时不需要过滤Sponsored和HD自家产品)，需要每周操作一次：

https://www.homedepot.com/b/Lighting-Ceiling-Fans/N-5yc1vZbvlq?catStyle=ShowProducts



（2）以下为从homedepot.com前台自然搜索类目CEILING FAN WITH LIGHTS后的结果(不需要筛选和排序，同时不需要过滤Sponsored和HD自家产品) ，需要每周操作一次：

<https://www.homedepot.com/b/Lighting-Ceiling-Fans-Ceiling-Fans-With-Lights/N-5yc1vZcjnu>

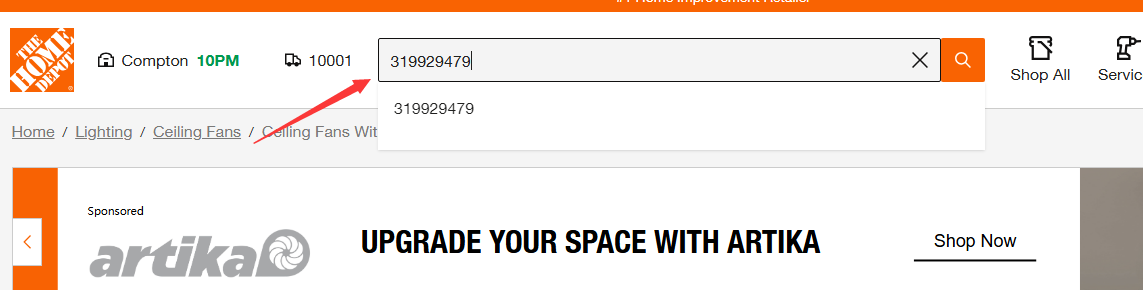


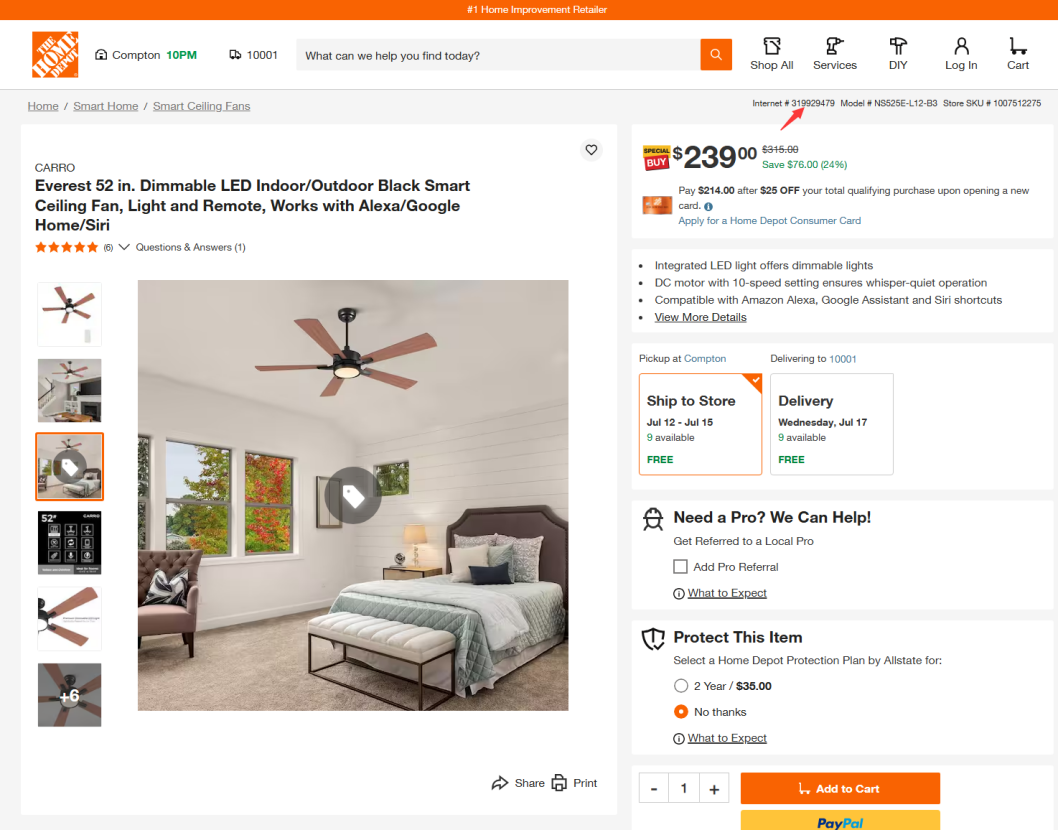
1. **新增爬取需求一：**将指定品牌（Carro,Parrot Uncle,Matrix Decro,Breezary）CEILING FAN占据（https://www.homedepot.com/b/Lighting-Ceiling-Fans/N-5yc1vZbvlq?catStyle=ShowProducts）和CEILING FAN WITH LIGHTS（<https://www.homedepot.com/b/Lighting-Ceiling-Fans-Ceiling-Fans-With-Lights/N-5yc1vZcjnu>）类目前5页排名的产品提取并生成一个表格，如附件二，看能否尽可能多地提取到跟表格一样的信息。需要每天都操作一次；（下划详见“问题”）
2. **新增爬取需求二：**使用Carro在HD的后台的SKU在前台自动搜索出该SKU所有的页面信息并保存下载。或者说能做到的自动化程度如何？（若需要整合内容成为表格，则参见“问题”部分）

例如：

登陆 [www.homedepot.com](http://www.homedepot.com)

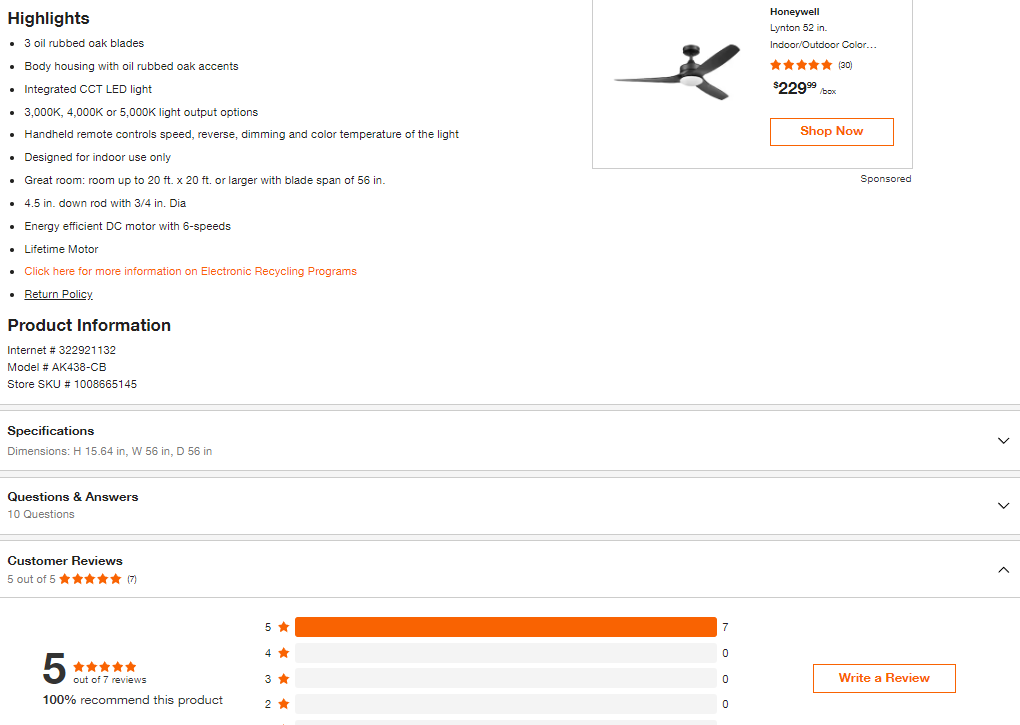
在搜索框输入HD SKU：319929479





问题详述：（适用于**需求一**和**需求二**对单个产品网页爬取的问题反馈）

该程序原理为模拟用户行为，在HD官网上进行 点击-搜索-浏览-返回网页解析数据 的步骤，因此可以批量获取在搜索界面返回的所有产品信息。然而，单个产品的详细信息，如OMSID, SKU, 第一个留评时间，顾客评分都隐藏在产品自带的网页上，需要爬虫依次对每日获取的280多个网页进行页面元素提取，如下图示例：



由于单个产品的网页的渲染时间并不固定，短时可在5秒内加载完成，长则多达数分钟，而爬虫并不能确认该页面上所有网页元素都已加载完成，因此在实际操作中，爬虫会返回约10%-20%因页面加载不出来而产生的空值。

下述列名均由单个产品的网页元素提取而来，极易产生大量空值：

SKU

OMSID

Rating

Date of latest review 与 review content则因平均耗时过长而返回大量空值，未登记在HD\_products.xlsx附件中。

-

此外，分别访问单个产品的网页耗时约为30秒左右，除去爬虫等待响应的时间，其最快返回速度也在三十分钟左右，是否可以先把残缺的数据导出来，交由人工补齐？

-7.16：1. 我们希望能够实现全自动操作、尽量减少人工操作，等待时间长一点也没关系

2. 附件一中的H列需要记录的是该产品最早期第一条评论的时间，而不是最近的。（应该是Date of the first review） （已解决）

若以附件二的表格为模板，则在每日更新的基础上存在以下困难：

1. 每次登记新的数据前，代码需要确认该产品SKU是否已在以前的爬取中登记过，若为新的SKU，则需要创建新的一列。因此当数据量过于庞大时，代码需要在每新增一行数据前，遍历以前所有的数据以比对。
2. 若以Excel表格为载体，以单日数据达280条为例，登记多日数据会引发多表格拼接，重复数据筛选遍历等占用内存过大的操作。

解决方法一：只考虑特定品牌的单个产品页面信息，缩小所需要的爬取数据量以增加单个产品页面的爬取时间。

-7.16：“只考虑特定品牌的单个产品页面信息”具体是指怎样呢？需要爬取的品牌现缩减为四个：Carro、Parrot Uncle、Matrix Decro、Breezary（详见附件二

-7.19：附件二为模板的爬取需求二是否能够实现呢？是否有样表可以看看呢？

方法二：鉴于Date of latest review，review content是固定内容，并不会随爬取次数改变，是否可以手动登记后，对爬取后的数据进行匹配。

-7.16：那如果有新增的review呢，能否在爬取时更新？ （已解决，详见共享文档）

新增问题：

sponsored的产品会在多个页面重复出现，并不能实现附件二中 “SKU-日期”的单一键值对。以7月17日爬取的数据为例，SKU为“1007173226”的产品在前五页中一共出现了四次，该如何处理同一产品有多个排序的情况？

其次，若在中午和晚上分别运行，也会出现同一SKU在同一天有多个排序，该如何进行筛选。

-7.19：重复的sku不需要处理，并不影响我们的数据参考

也就是说，如果要实现附件二中的结构，只取第一次出现的SKU的排序吗

对于在晚高峰时期进行爬取的需求，对两个搜索关键词的数据进行爬取耗时接近一个半小时，待运行完成也已远超美国这边的下班时间。是否有远程服务器或云服务可以在线运行？

-7.19：这个远程服务器或云服务我们也不太清楚，能否详尽解释一下是什么呢？

指可以在我这边工作电脑关机后，能供代码运行的服务器，这样我只需要发布定时任务，就可以在我离线后进行提取数据等操作。。

爬取时间过长是否由Date of first review和review content导致？如果是的话，是否可以通过采用上述的方法二（鉴于Date of latest review，review content是固定内容，并不会随爬取次数改变，是否可以手动登记后，对爬取后的数据进行匹配）去缩减爬取时间？或者说您有没有其他的一些解决方法？希望能够不影响到您的工作和生活。

这是由Homedepot网页架构决定的，在爬取过程中出现过页面元素一直加载，多次刷新无果的情况，或者页面只渲染了一半导致数据不完整，这和网站本身架构有关，我只能尽量通过增加等待时长来减少空值，不然就会出现之前返回的数据有10%-20%是空值的情况。一个半小时的时间是我为了最大程度减少空值设置的等待时长，短于这个时间就会有大量空值返回。

当前的解决方法是，如果公司有远程的服务器，或者国内工作电脑有配置python的环境，抑或是有购买云端服务，我就可以把这部分代码转移到上面定时运行。如果以上条件都没有，我可以在我下班后保持电脑的开机状态，让其运行代码。

云端的两个表格已经基本符合我们的需求了，若我们日后有关于这两个表格的细节优化再联系您，感谢！