

1. 错误次数太多, 用户连接超时(不定期出现)

原因: 太多次重复连接, 导致网页缓慢, 不能再指定时间(5秒)内连接上

解决: 将代理ip中间件换成新框架的

2. 代理池有卡顿

3. 频繁请求ip, 无论成功与否

4. 爬取速度慢:

原因: 不停在爬取重复页面, 在爬取帖子详细内容时, 重复请求该页面

解决: 目前是注释掉那句代码

5. 数据量不对:

原因: 目前爱卡更改了对发布帖子排序的操作, 增加了cookie(oderby参数)来控制是否排序

以前爬下来的是按回复排序

解决: 增加cookie 'oderby': 1,

6. 被爱卡禁止访问:

原因: 爱卡增加了cookie验证, 来做反爬

解决: 增加所需cookie, 具体如下:

'_appuv_newcar': 'f0fafccbbbed05fa57d0b404f1be0d158',

'_fwck_newcar': 'c28879b9110138c43ca7ef25c7e7f52b',

'_appuv_www': '37d34d06ad43bde460744eedc6a82c98',

'_fwck_www': 'b15320548fb602fafa7e384f4c0f568a',

7. 无法进行翻页:

原因: 获取下一页数字的代码错误, 例:

Href: forumdisplay.php?fid=741&orderby=dateline&page=4

原代码: 按 "=" 分开取第三个, 则会取出 dateline&page , 此为错误

解决: 按 "page=" 切分, 取第二个

8. 用户禁言论坛修改

能爬取禁言论坛, 修改了xcar.py文件