

推文整理步骤（正则法）

ultramarine471

February 2022

1 推文获取

1. 登录 <https://www.allmytweets.net/> (预设了该网站的可信性，有疑虑的可以先不用)。

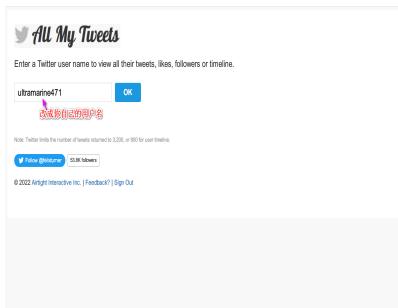


图 1: 登陆过程-1



图 2: 登陆过程-2

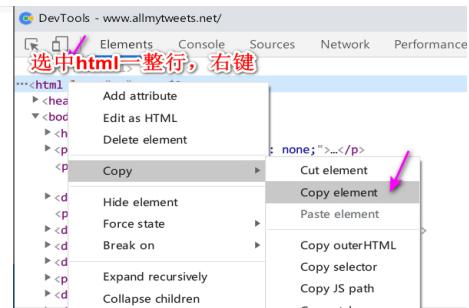


图 3: 复制元素

2. 页面从头到尾拖动几下，确保完全加载。
3. 按 f12，在最上面那个 `<html>` 处右键，然后点击复制元素。

2 信息筛选

4. 打开（除记事本以外）任意文本编辑器（本次使用 vscode 作为案例，其他编辑器应该大同小异），新建文件，粘贴先前复制的网页内容。

5. 键入 `ctrl+f` 查找，点击 使用正则表达式，上方输入框

```
<li>([\w\W]*?)</li>
```

6. 回车，点击页面空白处，键入 `ctrl+shift+l` (可见多处光标闪烁)，键入 `ctrl+f` 复制后新建文件，粘贴先前复制的初筛内容。

7. 键入 `ctrl+h` 替换，点击 使用正则表达式，上方输入框

```
<li>([\w\W]*?)<a.*href="(.*?)">\s<span class="grey">(.*?)</span>.*</a></li>
```

下方输入框（觉得 \t 比较合适就用了，也可以换别的）

```
$3\t$1\t$2
```

图 4: 步骤 4,5,6

图 5: 步骤 7

8. 键入 $ctrl+alt+enter$ 全部应用，目前还剩下一些引用链接未整理，在本文件内重复步骤 7，上方输入框

`<a.*?>(.*)?`

下方输入框（\$1 前后建议留取一定空格）

`$1`

9. 键入 $ctrl+alt+enter$ 全部应用，整理基本完成。

图 6: 步骤 8

图 7: 基本完成

3 补充信息

- 我的推文总量较小，像那种量大的、100k 以上的可能会存在加载问题，这个我不会处理，唯一能给的建议就是步骤 2。
- 步骤 5,7,8 的正则表达式在 vscode 里试过几次，没意外的话应该可以。
- 步骤 7 中 \$1,\$2,\$3 分别代表第一、二、三个 (.*?) 占据的内容。在这里 \$1,\$2,\$3 分别是正文、网址和日期，可以随意排列的，例如

“日期 \t 正文 \t 网址” \rightarrow “\$3\t\$1\t\$2”