

Data Collection from social media

Xiaohongshu:

1. Do have official API, but it is not open to the public. These APIs are mainly used for internal applications and partners: like E-commerce
2. Some third-party companies provide API
3. Scraper: <https://github.com/XiaochenCui/Zhihu-Scraping/blob/master/zhihuScraping.py>

```
1  {
2      "source_platform": "xhs",
3      "url": "https://www.xiaohongshu.com/explore/6486a9440000000027001034",
4      "title": "15种让你变好的水果，建议姐妹收藏🍎 - 小红书",
5      "content": "水果大家都爱吃，而且不分年龄段，只是每个人吃水果的爱好不一样。那么，对于我们女人来说，想要皮肤好，排du养颜、抗衰老，多吃哪些水果呢？👉 接下来让我带大家👉 一起来看一下吧。 1. 苹果 🍏 2.
6      "img_urls": [
7          "http://sns-webpic-qc.xhscdn.com/202406091744/6d504a712ae885b3d9a28cea36a95ff/1800g0882kspk2t2im06g501puoj08msi4iuiwif0ind_prv_wlteh_jpg_3?imageView2/format/jpg!1",
8          "http://sns-webpic-qc.xhscdn.com/202406091744/6b644cc416a86deb1872553d82d034464/1800g0882kspk2t2im06g501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
9          "http://sns-webpic-qc.xhscdn.com/202406091744/6d504a712ae885b3d9a28cea36a95ff/1800g0882kspk2t2im06g501puoj08msi4iuiwif0ind_prv_wlteh_jpg_3?imageView2/format/jpg!1",
10         "http://sns-webpic-qc.xhscdn.com/202406091744/6b644cc416a86deb1872553d82d034464/1800g0882kspk2t2im06g501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
11         "http://sns-webpic-qc.xhscdn.com/202406091744/506dc0a809ec2191fd259dc114dd25d2/1800g0882kspk2t2im06g501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
12         "http://sns-webpic-qc.xhscdn.com/202406091744/4948ae85b5088262086d1a2d335409e/1800g0882kspk2t2im050501puoj08msi4iuiwif0ind_prv_wlteh_jpg_3?imageView2/format/jpg!1",
13         "http://sns-webpic-qc.xhscdn.com/202406091744/506dc0a809ec2191fd259dc114dd25d2/1800g0882kspk2t2im050501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
14         "http://sns-webpic-qc.xhscdn.com/202406091744/4948ae85b5088262086d1a2d335409e/1800g0882kspk2t2im050501puoj08msi4iuiwif0ind_prv_wlteh_jpg_3?imageView2/format/jpg!1",
15         "http://sns-webpic-qc.xhscdn.com/202406091744/901fbc11dfbac508e461a2d685a28ff4/1800g0882kspk2t2im010501puoj08msi4iuiwif0ind_prv_wlteh_jpg_3?imageView2/format/jpg!1",
16         "http://sns-webpic-qc.xhscdn.com/202406091744/14b585b8d45aa02b19e21b85fd0cd370/1800g0882kspk2t2im010501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
17         "http://sns-webpic-qc.xhscdn.com/202406091744/14b585b8d45aa02b19e21b85fd0cd370/1800g0882kspk2t2im010501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
18         "http://sns-webpic-qc.xhscdn.com/202406091744/14b585b8d45aa02b19e21b85fd0cd370/1800g0882kspk2t2im010501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
19         "http://sns-webpic-qc.xhscdn.com/202406091744/f49a6bce6119d56cdd638cfcfd09bce1/1800g0882kspk2t2im04g501puoj08msi4iuiwif0ind_prv_wlteh_jpg_3?imageView2/format/jpg!1",
20         "http://sns-webpic-qc.xhscdn.com/202406091744/f49a6bce6119d56cdd638cfcfd09bce1/1800g0882kspk2t2im04g501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
21         "http://sns-webpic-qc.xhscdn.com/202406091744/f49a6bce6119d56cdd638cfcfd09bce1/1800g0882kspk2t2im04g501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
22         "http://sns-webpic-qc.xhscdn.com/202406091744/f49a6bce6119d56cdd638cfcfd09bce1/1800g0882kspk2t2im04g501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
23         "http://sns-webpic-qc.xhscdn.com/202406091744/887bd5c6153d21f664b4502a643ac8a/1800g0882kspk2t2im040501puoj08msi4iuiwif0ind_prv_wlteh_jpg_3?imageView2/format/jpg!1",
24         "http://sns-webpic-qc.xhscdn.com/202406091744/887bd5c6153d21f664b4502a643ac8a/1800g0882kspk2t2im040501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
25         "http://sns-webpic-qc.xhscdn.com/202406091744/30459c084e02d8c2f795a2a7971ba82e/1800g0882kspk2t2im040501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
26         "http://sns-webpic-qc.xhscdn.com/202406091744/30459c084e02d8c2f795a2a7971ba82e/1800g0882kspk2t2im040501puoj08msi4iuiwif0ind_dft_wlteh_jpg_3?imageView2/format/jpg!1",
27     ],
28     "author": {
29         "nickname": "爱丽拉健康健康管理",
30         "avatar": "https://sns-avatar-qc.xhscdn.com/avatar/62a7e9579d45ef00018768ce.jpg",
31         "userId": "6039f6260000000001005b92",
32         "city": null
33     },
34     "liked_count": "1k+",
35     "collected_count": "1k+",
36     "comment_count": "10+",
37     "share_count": "100+",
38     "comments": [
39     ]
40 }
```

X:

1. Have official paid X API:
 - 1) Free: No posts access with API
 - 2) Basic: 10,000/month Posts read-limit rate cap; cost: \$100 per month
 - 3) Pro: 1,000,000 posts per month; cost: \$5,000 per month<https://developer.x.com/en/docs/twitter-api>
total posts used in PLIP: 232,067 tweets
2. Does not allow the users to use web scrapers.
 - <https://developer.x.com/en/developer-terms/agreement-and-policy>
3. Scraper: <https://github.com/israel-dryer/Twitter-Scraper/blob/main/scrapper.py>

Zhihu:

1. User policy: <https://www.zhihu.com/term/zhihu-terms>

您应对您使用本平台的行为负责，除非法律允许或者经知乎事先书面许可，

您使用本平台不得具有下列行为：

以任何方式（包括但不限于盗链、冗余盗取、爬取、抓取、模拟下载、深度

链接、假冒、模拟注册等) 直接或间接盗取知乎平台的数据和内容; 恶意注册知乎账号, 包括但不限于频繁、批量注册账号。

2. Official API used for internal applications and partners. There are third-party API but may be not official recognized
3. Scraper: <https://github.com/ZWY24/Web-Scrapping/blob/main/scrapper.py>