八、统计知乎大V所有文章标题

实例：知乎张佳玮

https://www.zhihu.com/people/zhang-jia-wei/posts

目标效果图

730	阿伦-艾弗森；2000-01季；逆天
731	怎么，您以为翻译腔只是“见鬼，老伙计，我要踢你的屁股”吗？
732	帕森斯这三年4600万报价：勒布朗们动静的第一块多米诺骨牌
733	6月25日，北京，发布一本曾经的禁书和一本小说
734	希尔、卡特、麦蒂、韦德、科比——乔丹接班人们的“那一步”
735	在罗兰加洛斯的红土亲眼看见纳达尔的上旋，是种什么体验
736	国产航母下水……让我想到李鸿章和北洋舰队
737	湖人勇士跑起来就赢了；勒布朗与希腊怪物到底没对上位
738	趁你还吃得下一切的时候
739	2014年全明星周末碎记：大家一起来找槽点的真人秀
740	道歉信和危机公关
741	说到有趣，立时想起来便会笑的字句
742	父母们老来，如何彼此称呼呢
743	勒布朗-詹姆斯回骑士之理智与情感；以及林书豪、湖人、火箭种种
744	NBA总决赛2014第四场：马刺，那是一整支球队啊
745	统治前场篮板便统治球场的时代，慢慢过去了
746	一段旅行如何开始与结束
747	谢谢你，蒂姆-邓肯
748	爸爸是天下第一高手啊

规律分页

第 1 页：https://www.zhihu.com/people/zhang-jia-wei/posts?page=1
第 2 页：https://www.zhihu.com/people/zhang-jia-wei/posts?page=2
第 3 页：https://www.zhihu.com/people/zhang-jia-wei/posts?page=3
……
第 42 页：https://www.zhihu.com/people/zhang-jia-wei/posts?page=42

规律：
第 n 页：
https://www.zhihu.com/people/zhang-jia-wei/posts?**page=n**

抓取方法

https://www.zhihu.com/people/zhang-jia-wei/posts?**page=n**

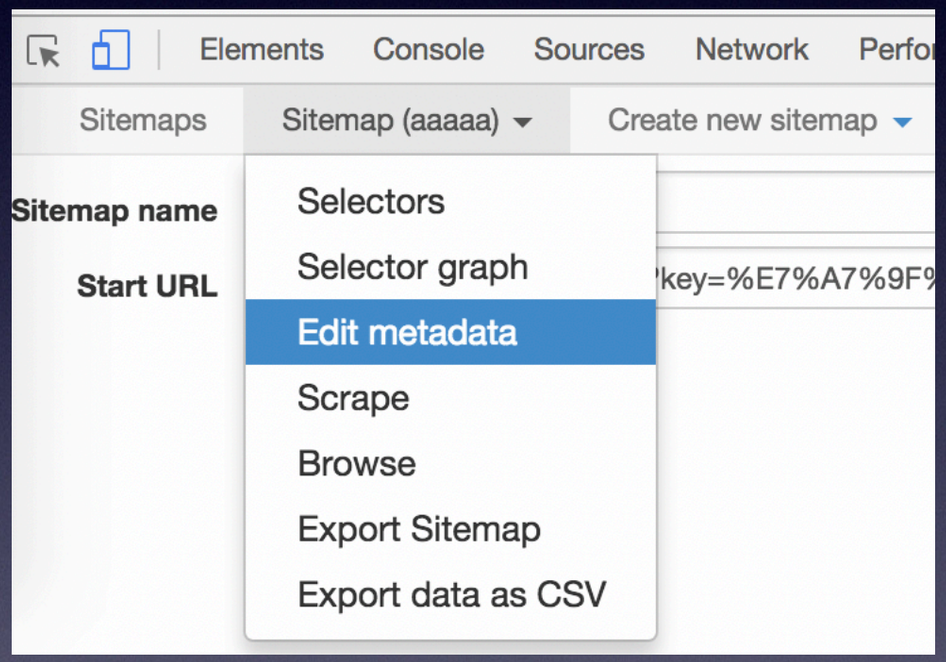
将 start URL 改为：

https://www.zhihu.com/people/zhang-jia-wei/posts?**page=[1-42]**

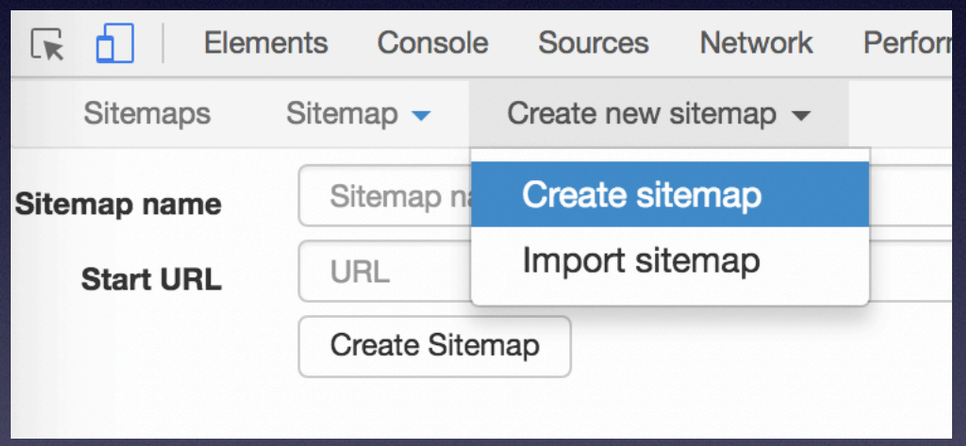
[1-42] 表示从第 1 页到第 42 页
[1-10] 表示从第 1 页到第 10 页
[6-20] 表示从第 6 页到第 20 页

具体操作步骤

修改 start URL：



新建 start URL：



不同网站规律不同

知乎：https://www.zhihu.com/people/zhang-jia-wei/posts?**page=n**

赶集网：http://weinan.ganji.com/fang1/**on**/

链家网：https://bj.lianjia.com/ershoufang/**pgn**/

如何发现这些规律 —— 多点几下不同的页码

警惕第 1 页！！！

其他规律分页

豆瓣

第 1 页：https://book.douban.com/tag/%E5%B0%8F%E8%AF%B4?start=**0**&type=T

第 2 页：https://book.douban.com/tag/%E5%B0%8F%E8%AF%B4?start=**20**&type=T

第 3 页：https://book.douban.com/tag/%E5%B0%8F%E8%AF%B4?start=**40**&type=T

第 4 页：https://book.douban.com/tag/%E5%B0%8F%E8%AF%B4?start=**60**&type=T

……

第 10 页：https://book.douban.com/tag/%E5%B0%8F%E8%AF%B4?start=**180**&type=T

相邻页 的数字差：**20**

其他规律分页

豆瓣

首页对应的数字：0

末页对应的数字：180

相邻页的数字差：20

抓 1-10 页：

https://book.douban.com/tag/%E5%B0%8F%E8%AF%B4?start=**[0-180:20]**&type=T

不同网页的规律不同！！

如果要抓取一个网站的翻页数据
首先，分析 URL 是否有规律
如果有规律，就按照本节的方法抓取
如果没有规律，后面会讲到其他翻页方法。