

# 知乎—爬虫

2017 年 8 月 4 日

## 1 模拟登录

1. 知乎爬取时必须带上header。
2. `_xref` 可以从登录网页的源码中提取，但是提取和登陆的网页必须是同一个。`urllib2.urlopen` 是打开可能不同的，可以使用requests库建立session。

## 2 爬取follow , 问题等内容

【部分知乎问题已有答案了】

知乎网页是动态加载问题，怎么解决？以爬取问题或关注人为例，通过开发者工具中的Network – XHR / Js 选项，向下拉取页面，可以看出网页在加载时是通过offset 来获取新的内容（不是在类似batch的请求中，注意），复制出URL备用。

编写header, 除了基本header外，必须还要有authorization，authorization貌似在不同的爬取需求中都是相同的，但是header必须包含它。打开网页后的内容好像是json 格式，使用`json.loads(url.content)` 将内容转化为字典形式。

通过正则表达式或是BeautifulSoup 可以提取需要的内容，在提取速度和存储方式上，文件操作等方面有待加强。