**Why?**

作者: **莫烦**编辑: **莫烦 2018-01-03**

学习资料:

* 我的爬虫[教学代码](https://github.com/MorvanZhou/easy-scraping-tutorial)

其实你身边到处都是爬虫的产物, 比如说搜索引擎 ([Google](https://www.google.com/), [百度](http://www.baidu.com/)), 他们能为你提供这么多搜索结果, 也都是因为它们爬了很多信息, 然后展示给你. 再来说一些商业爬虫, 比如爬爬淘宝的同类商品的价格信息, 好为自己的商品挑选合适的价格. 爬虫的用途很多很多, 如果你搞机器学习, 爬虫就是你获取数据的一种途径, 网上的信息成百上千, 只要你懂爬虫, 你都能轻松获取.

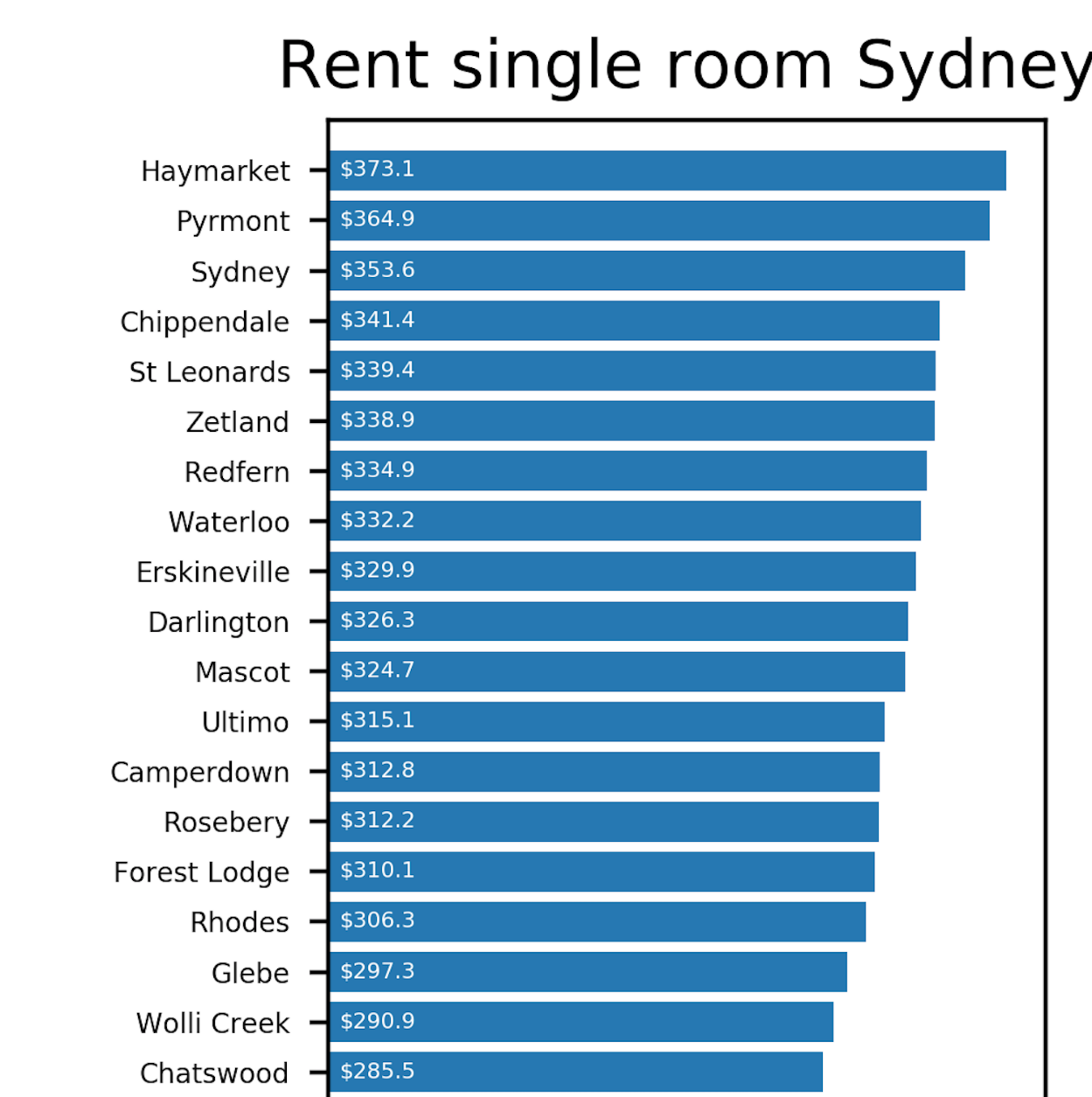
[**我做过的一些有趣的爬虫**](https://morvanzhou.github.io/tutorials/data-manipulation/scraping/1-00-why/#%E6%88%91%E5%81%9A%E8%BF%87%E7%9A%84%E4%B8%80%E4%BA%9B%E6%9C%89%E8%B6%A3%E7%9A%84%E7%88%AC%E8%99%AB)

直接开始教学太没意思了, 不如先来看看我做过的一些爬虫吧. 我有 [Github](https://github.com/) 的代码库, 里面有很多机器学习的[教学代码](https://github.com/MorvanZhou) (爬虫的教学代码也在[这里](https://github.com/MorvanZhou/easy-scraping-tutorial)哦). 为了对比我的代码库质量好坏. 通过[这一节](https://morvanzhou.github.io/tutorials/data-manipulation/scraping/3-01-requests/) 内容的知识, 我爬取了 Github 上分类为 [Machine-learning](https://github.com/search?l=Python&q=machine+learning&type=Repositories&utf8=%E2%9C%93) 的 Python 库, 然后整理了数据, 并可视化了一些潜在关系, 比如 fork (被下载)数 和 star (被点赞)数的关系. 放个视频给大家看看结果.

图片结果可以在这里看到:

* 2016-2017年的结果 [条形图](https://morvanzhou.github.io/static/results/scraping/github_bar2016.png) [点图](https://morvanzhou.github.io/static/results/scraping/github_sca2016.png)
* 2017年的结果 [条形图](https://morvanzhou.github.io/static/results/scraping/github_bar2017.png) [点图](https://morvanzhou.github.io/static/results/scraping/github_sca2017.png)

另外一个有意思的是, 我现在住在悉尼, 悉尼的租房价格可贵了, 为了了解每个区的整体价格, 方便确定自己将要在哪里租房. 通过[这一节](https://morvanzhou.github.io/tutorials/data-manipulation/scraping/3-01-requests/) 和 [这一节](https://morvanzhou.github.io/tutorials/data-manipulation/scraping/4-02-asyncio/)内容的知识, 我做了个爬虫帮我爬取著名华人住房社区-[今日悉尼](http://www.sydneytoday.com/house_rent) 的租房价格. 注意澳洲的租房一般是以周计算, 单位是澳刀, 我给大家展示一下结果的前半部分, 因为图太长了, 后半部分请在[这里查看](https://morvanzhou.github.io/static/results/scraping/sydney-rent-avg_price.png).

[](https://morvanzhou.github.io/static/results/scraping/sydney-rent-head.png)

[**为什么做这个教程**](https://morvanzhou.github.io/tutorials/data-manipulation/scraping/1-00-why/#%E4%B8%BA%E4%BB%80%E4%B9%88%E5%81%9A%E8%BF%99%E4%B8%AA%E6%95%99%E7%A8%8B)

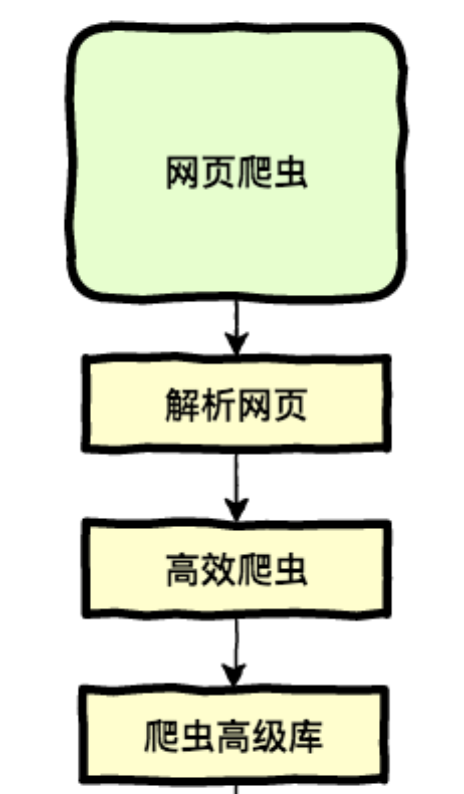
这个教程提供了你一次入门的机会, 当然网上还有很多很好的入门教程, 比如:

* 崔庆才 的[Python爬虫学习系列教程](https://cuiqingcai.com/1052.html)
* [知乎问答](https://www.zhihu.com/question/47883186)中的各种推荐
* 孔淼 的[一看就明白的爬虫入门讲解](https://www.csdn.net/article/2015-11-13/2826205)

这些都是非常好的参考资料, 你的学习请不要只限于一个网站. 因为像机器学习一样, 爬虫也囊括的东西绝非不止一点点. **而你为什么要看看我的爬虫教程呢? 因为我只关注基础, 我认为入门是最重要的, 能帮你成功引上路子, 我想你会轻松很多. 而且搭配视频讲解的形式, 也会更加容易理解. 是为初学者定制的. 如果你已经入过门, 想着如何商业化爬虫, 这个教程应该不能满足你的需求了.**

当我第一次接触爬虫的时候, 其实很陌生, 完全不知道从何开始. 在网上自己搜一些介绍, 但是他们的介绍都太笼统了, 给你丢几个关键词, 让你自己解决. 当时我就懵逼了, 看着那些关键词 (requests, urllib, beautifulsoup, scrapy) 不知道从何下手, 我估计你也会有这种感觉. 当时, 我花了大把的精力, 想弄懂这些东西和爬虫的关系. 而且分清如果只想入门, 我们需要掌握哪些? 毕竟商业化和入门还是有很大不同的. 有些关键词或者模块是为了商业化而用的. 所以我花了很多时间, 整理网上的这些信息. 总结出一条入门爬虫的便捷之路.

[**课程介绍**](https://morvanzhou.github.io/tutorials/data-manipulation/scraping/1-00-why/#%E8%AF%BE%E7%A8%8B%E4%BB%8B%E7%BB%8D)

[](https://morvanzhou.github.io/static/results/scraping/1-1-0.png)

这系列教程按照上面的逻辑来教会你爬虫, 我们会从网页的基本结构开始讲述, 慢慢使用一些简单的工具, 做一些简单的爬虫. 还会有一些小练习, 让你爬爬真正的互联网. 下载美图, 逛逛百度百科, 全网爬取等等. 当你懂得了爬虫的概念, 我们在深入一些, 谈谈如何加速你那和蠕虫(爬的慢)一样的爬虫, 把它升级为一只小飞虫(多进程,异步爬取). 当然这些内容都不会特别深入, 重点是把你带入门. 但是我会在每节内容里加一些链接, 提供给想要深入了解的朋友们.

[下次](https://morvanzhou.github.io/tutorials/data-manipulation/scraping/1-01-understand-website/)我们就正式开始学习爬虫啦.