学习社群的理念：

02 记录 社群理念

> 我希望经营一个什么样的小型社群？
>
> 我也是初学者；爬虫系列教程80%的脑中关于爬虫的知识都会呈现出来。

1：我和爬虫的故事

> 这是一门小众的技巧，掌握爬虫技术在职业生涯里，可能只能占到微小的分量.
> python在数据处理方面的强大性能和广泛的使用，可以尽早学习数据科学，具体后续会写数据科学专栏。

我学习爬虫知识的想法是因为找实习，一家公司需要网页爬虫实习生，专门负责爬虫。于是我上网找视频，找教程学习，因为当初我已经下定走IT之路，所以自学了python，把python核心编程这本书一路啃下来，一边做笔记，一边练习，一边写代码。半个月的整块时间，学会了爬取豆瓣TOP250电影，并存取在本地。然后就写了简历投向了那家公司。因为是家小公司，同时面试官也是个90后的大神，可能志趣相投，又因为项目紧急，所以顺利招了我。马上入职，一周三天，专门负责爬虫。前两周的学习，带我的头儿，给了我几个网站，要我爬取整个网页的网址，或者整个网页的文字。没有其他的任务。前两周也就还是自己练习爬虫，之后给了我整体的500+信息源，告诉我任务具体是什么，当然，一口吃不出胖子，头儿挑选了比较容易处理的网页要我抓取数据。就一个网页，先是抓取并显示出来，之后要求把数据结构化，还是在同一份代码上进行重构。之前一直使用的是正则表达式匹配文本。再之后又在同一份代码上让我使用BeautifulSoup进行匹配，后写了遍，再之后又让我在同一份代码上使用xpath匹配。白天在公司上班，晚上回来看官方文档，学习模块的常用方法，一般文档都会特别长，但是真正使用的函数并不会特别多。如何甄别？google，好几份教程。看别人使用哪些函数方法，使用的频繁的就应该是你重点需要学习的。光看还学不会，还得写代码。一个月过去，其实还没有正在的涉及任务。当然学习过程中出现各种问题，头儿是个全栈工程师，他让我尽量自己摸索，快下班的那会，把问题集中起来问，这些问题，一般他三下五除二就能给你解决掉。到了晚上，自己在进行总结。一个月后，正式的抓取财经数据，陆续接触MySQL关系型数据库，Mongodb 文档型数据库，elasticsearch分布式搜索数据库. 抓取的大量数据进行存在不同的数据库中为后续的量化分析师进行分析。所以数据需要结构化，字段的名称等都需要统一起来，不然数据分析那块，数据清洗会占据很大的时间。陆续上手后，处理的数据量大了，就会涉及多线程和协程方面的知识。500+信息源，大多是财经数据，所以并没有涉及登入等操作，而且这些新闻网站都没有反爬虫机制，所以可以大规模爬取，大概一个网页会是一个子爬虫，一般是一天一个子爬虫。后来写的快了，一天三个爬虫。包括一系列的抓取，清洗，整理，存取等操作。就这样维持了5个月… 整个爬虫项目都是我负责，头儿把整体的框架搭好后，就让我写子爬虫。整个项目最后放在服务器上实时更新。对了为了实时更新，抓取规则和更新规则的处理也花了好些时间进行布局。协同合作，学习了git分布式版本控制.
最后大概写了100+ 数据表.
20+ 文档
100万+ 数据

学习的过程中，发现公司的那些牛人都有一些共同的习惯：

- 良好的编码风格，甚至公司内部都规定好了一定的编码风格
- 代码的管理，电脑文件夹分门别类的规划好，git上备份代码，不能公开显示的源码，就付费在git上托管。
- 持续的学习，带我的头儿已然是全栈工程师，还不断的学习自然语言，机器学习，图像处理等，听说晚上下班还在编码。全然不看电视。
- 交流与沟通，公司内部每周都有分享会，遇到的问题，解决方式，新知识等都会进行交流

  ….

2：学习的理念

新知识的学习易采用发散式思维，联想和类别是非常好的学习方式。同时新知识的学习，需要非常注意概念的学习，概念是你理解和精进的第一步。一般我会套用自己形成的公式对一个新概念问两个问题：1. 这是什么 2. 它能干什么 . 用自己话复述一遍是知识在现的很好的方法。
新知识易采用重复性学习，大脑对新知识的容纳能力有限，持续学习3小时，装下的知识也不会很多，这个时候应该采用分块化学习，重复性学习。一天2小时，每天持续时间，同时对旧知识的不断回顾。回顾的方法易使用白纸和笔的方法，你写不出来，没记住，那就回去翻旧知识。
新旧知识需要不断的更新，直到新知识变成旧知识，旧知识回顾起来的时间越来越短。再持续接触新知识。

不断循环。

任何领域，想要学会一样技能，离不开重复，实践。事实上这是最显而易见的事实。刻意练习这个词是一万小时理论下的一个概念，了解了这个概念后，我对之前的知识进行了梳理，不会的不管碰的，都尝试学习。这样扩充知识。

知识易先大致框架，在深入的模式。这种叫做集中和发散式学习思维。

总结下：
学习注重概念
学习新旧知识不断循环
学习与实践

3：学习爬虫的建议

专注一个兴趣的网页，不断的使用正则啊，beautifulSoup，xpath等进行代码**重构**，数据库存储等，数据库又可以使用不同的方案。在学习的过程中不断梳理python模块。学有余力在进行特定网页的抓取。

多查看官方文档，不知道重点，那连开几个教程系列，看别人用哪些？
学会了基础的，自己可以回过头再对不熟悉的函数练习使用
摈弃完美主义：8/2法则
多实践

4：推荐读物

- 《把时间当做朋友》 – 李笑来
- 《暗时间》 – 刘未鹏
- 《心理学与生活》
- 《万万没想到》 – 万维钢
- Coursera：视频课程

5：群内规则

- 不鼓励闲聊
- 鼓励分享，最好自己梳理，整理成完整的知识分享，这也是锻炼自己的归纳总结的能力
- 集中时间答疑，一般是晚上
- 理念是：持续不断的精进。对旧知识掌握到一定程度，一定要接触新知识，这样才能精进
- 交流 ： 闭门造车，后果自负