

LetPub 爬虫使用说明

- 这组程序用于从Letpub网站上爬取其数据库，能爬取的数据见参考表头.xls文件
- 该爬虫仅作学习交流使用，不得用于任何商业用途
- 语言环境 Python3.7.1
- 需要的第三方库：
 - a. requests (获取页面代码)
 - b. beautifulsoup (解析并获取信息)
 - c. xlrd (用于读execl)
 - d. xlutils (用于写execl)
 - e. lxml (用于解析网页，非必需)
 - f. xlwt (用于创建execl,非必需)
- 程序包内共有两个程序
 - a. once.py (爬取的数据中途不存储，一次读取一次写入，适用于**网络稳定、保证不会中断**的状况)
 - b. save.py (爬取的数据每爬取10个存储一次，适用于网络不稳定的状况，如果对效率并非太过苛求，推荐采用这个)
- 由于本程序工作量不大，并且考虑到目标网站服务器的承受能力，本程序未对爬虫进行任何加速措施。如果确有加速需求，可以酌情考虑异步协程提速等方法，具体方法可参考[爬虫速度太慢用异步协程提速](#)。