**多线程爬虫（简易版）脚本说明文档**

**made by 0x11111111**

**blog: https://blog.0x11111111.top/zh https://blog.0x11111111.top/en**

1. 脚本介绍
   * 该脚本由python3编写，实现功能为从https://www.shiyanlou.com/爬取各门课程相关信息并且保存在xls文件中。
   * 为保证脚本正常执行，请认真阅读此文档后再使用该脚本。并且确保包中文件完整、处于同一个目录下。
   * 未经授权允许，请不要随意改变该脚本用途，造成一切后果由使用者承担、
2. 脚本使用说明
   * 该脚本为简易版本，完整文件由爬虫主体（multiThreadSpider.py）第三方多线程包（threadpool.py）以及第三方xls读写包（xlwt）构成
   * 运行该脚本之前，请修改第21行的end数目，该数字为终止爬取id数（不含），请务必调整大小已保证抓取完整。数字设置过大会导致太多失败抓取，造成时间与资源浪费。
3. 脚本算法与原理说明
   * 该脚本通过网址的遍历id数，使用urllib.request抓取网页源代码，再利用正则表达式提取有关信息，最后按行号写入文件，每次写入完毕将row下移一个。
   * 采用的反反爬虫机制包括：添加用户代理伪装成普通浏览器用户，出现404错误后减缓抓取速度，以及二次抓取初步抓取失败的网页，以防由于反爬虫机制造成的漏抓。
   * 抓取完毕后，会在该脚本所在目录下生成一个以当天日期为文件名的xls文件，其中课程号由于多线程机制以及二次抓取，不是严格递增排列，建议排序以提高可读性。
   * pool = threadpool.ThreadPool(2) 中数字为线程池线程数，设置过大会导致抓取过快，触发网站反爬虫机制，表现为连续数十个网页抓取失败。为防止遗漏，本脚本采取二次抓取。
4. 主要函数介绍
   * get\_info(id, captured):从captured中的网页源文件提取信息，返回一个包含信息的元组。
   * write\_in(towrite, row=0):将表头headings与get\_info获取的信息按行号row写入。
   * capture(id):抓取id的网页源代码，调用get\_info()函数提取信息，再用write\_in()函数写入对应行号。