图书馆数据分析

创建时间： 2019/8/3 14:35 **更新时间：** 2019/8/3 14:39

图书馆数据分析

项目组成

项目应包含以下部分：

- 数据爬取
- 数据储存
- 脚本在服务器上的运维
- 数据分析
- 跟图书馆的交涉（误

数据爬取

- 要求
  - 可以自由更改收集数据的粒度（例如15min扫一遍或者30min扫一遍）
  - 可以自由更改收集数据的范围（比如单独扫信息馆或者全部分馆扫描）
  - 可以自由更换登陆者身份（防止使用单一用户导致被封）
- 输出字段要求
  - id
  - 确切日期
  - 小时时间粒度编号（如8表示8点，20表示20点）
  - 分钟时间粒度编号（0表示0-15min，1表示16-30min等）
  - 分馆名称（或序号）
  - 房间名称（或序号）
  - 座位编号
  - 座位状态编号（0表示空闲，1表示使用中，2表示被预约，3表示暂离，4表示不可用）
  - 是否靠窗（0或1）
  - 是否电源（0或1）
  - 是否电脑（0或1）

> 举例：
> (id)1 (日期)2019-9-1 (小时)8 (分钟)0 (分馆名称)信息分馆 (房间名称)二楼西自然科学区 (座位编号)1 (座位状态编号)0 (靠窗)0 (电源)1 (电脑)0

- 主要流程
  - 从网页上爬取数据
  - 数据清洗和整理

数据存储

- 根据需求建好相应的表
- 配置好服务器数据库访问权限
- 编写将规范化数据存入数据库的代码

脚本服务器运维

- 脚本运行情况的日志代码编写
- 容错脚本编写（防止脚本和服务器宕掉之后数据缺失）
- 被封ip了怎么办
- 图书馆网站崩了怎么办
- 脚本运行的服务器环境配置

数据分析

- 待定，可以从一周的数据，一月的数据，半学期的数据，一学期的数据，整年的数据逐步推进

与图书馆的交涉

- 防止被封ip
- 防止被请喝茶
- 防止自己的用户名被永久封禁
- 可以的话他们能给我们提供一些脱敏的内部数据（如选座学号等）