**爬虫项目实战——项目需求文档**

**项目要求**:

1）豆瓣网数据采集

① 按照关键字搜索图书或者电影

② 采集3-5个字段

③ 采集数据50条左右即可

2）去哪儿网数据采集

① 采集“景点”或者“美食”数据

② 采集3-5个字段，此外需要采集经纬度信息（经纬度不包括在5个字段内）

③ 采集数据50条左右即可

3）blibli弹幕数据采集

① 搜索任意关键字，找到相关视频列表，选取3个视频

② 选取3个视频进行采集，采集其弹幕数据 - 1个视频约1000条弹幕

4）拉勾网数据采集

① 按照关键字搜素岗位信息，或者采集公司信息数据

② 采集3-5个字段

③ 采集数据50条左右即可（岗位信息or公司信息）

**项目结构**:

根目录： ./

Bilibili弹幕采集: ./bilibili弹幕

豆瓣网数据采集: ./豆瓣网

拉勾网数据采集: ./拉勾网

去哪儿网数据采集: ./去哪儿网

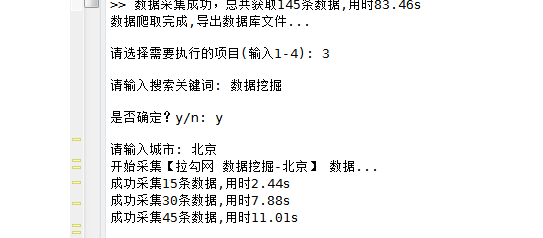
导出数据目录: ./data

共用模块： ./setting.py

多线程模块: ./myThread.py

**调用方法**:

1. 各项目子文件夹中py文件和bat文件均可调用爬虫项目, 不支持交互;
2. 根目录调用 项目执行合集.py文件 和 start.bat 文件可调用4个项目，支持交互，交互方式如下图所示:



**项目特点：**

1. 支持多线程和单线程。
2. 使用Mongodb存储数据；
3. 综合使用selenium和requests/bs，互相补充不足点；
4. 调用方法便捷，爬取速度快且稳定；
5. 可自行在setting中设置代理如阿布云，提高爬虫健壮性，避免多线程任务过程中高并发请求被检测出来；
6. 自定义模块之间相互调用，但亦可独立运行互不影响，且采用了类继承方式进行重写方法等，减少代码冗余。