用爬虫采集人人好友信息

## 运行

### 环境配置

1. Python2.7.6（32位）
2. 非系统自带的库：requests（http://docs.python-requests.org/en/latest/）

### cookie文件

1. 需要文件名为：cookie.cookie
2. 内容格式如下图：  
   [](cookie.jpg)
3. 如何获取该项值：  
    [](cookie_find.jpg)

### 运行方式

1. 完成以上设置。
2. 运行run\_spider.py文件

### 结果文件

1. 结果将生成两个文件夹。
2. Original\_data文件夹保存的是原始ID号的文件
3. Mapped\_data文件夹保存的是映射到新ID的文件
4. 各个文件的意义：
5. Center\_node.txt：只有一行，保存的是中心点的ID值
6. Edges.txt：保存的是经过清洗后剩下的边。每一行表示一条边，每条边只出现一次。每条边的表示方法是：id值较小的点在前，id值较大的点在后，中间用空格分隔。
7. Edges\_unselect.txt：保存最原始的、还未清洗的边。
8. Index.txt：保存id及对应的姓名、标签，分别用冒号分隔。
9. Mapping\_table.txt：保存原始id与映射后id的对应关系，用空格分隔。

## 系统构成

### 网页爬取

对应于Page.py。功能是读入cookie.cookie文件，爬取指定url对应的网页并将其返回。

### 网页解析

对应于FriendlistPageParser.py。功能是解析爬取到的网页。分别提取中心点id、好友列表的总页数和获取当前页面上的好友列表。

### 数据处理

* + 1. 数据采集、保存逻辑
       1. 对应于get\_edges.py。
       2. 功能是获取并保存所有中心点的好友及好友的好友，并将好友关系转化为边的关系。
       3. 将得到的数据写入到文件中。
    2. 数据清洗
       1. 对应于select\_nodes.py。
       2. 功能是从所有爬取到的好友关系中，筛选出一部分点及其组成的网络作为被分析的社交网络。
       3. 清洗逻辑。设置阈值。如果某个点和其它点的连接数小于阈值，那么该点及该点所在的边都被去除。
    3. ID映射
       1. 对应于map\_id.py。
       2. 功能是将原始ID映射到一个新的ID。新ID从0开始连续的整数。
       3. 原因。很多算法要求输入的点的id号是连续的，且不能过大。