爬取数据分为两部分操作，思路如下

1，从几个中心微博（如@中山大学、@中山大学学生会……）中爬取粉丝列表，形成粉丝大名单（相信会有重复项即同时关注了这几个微博的账户），这个大名单估计会在4万个账户以内。然后遍历大名单中每个账户的互粉情况（只获取关注和粉丝的列表，配对互粉，如不互粉可以清除掉），如果一个账户与在大名单中的五个以上账户互粉，则可判断其为社群中账户，这样可以将社群外的账户排除，僵尸粉应该能清理干净。

（最后的账户数估计在3万个以内）

2，在上一步的最终名单中，获取每个微博账户的关注列表、粉丝列表、账户下每条微博的被评论和被转发的情况（如评论/转发者的帐号、评论/转发次数统计）、在微博中“@其他账户”的次数。

除统计@的情况外，不需要获取微博的内容，不对内容进行分析

附：中心微博选取

|  |  |
| --- | --- |
| 微博名 | 微博链接 |
| 中山大学学生会 | http://e.weibo.com/u/1797174742 |
| 中山大学南校区学生会 | http://e.weibo.com/u/2367241050 |
| 中山大学东校区学生会 | http://e.weibo.com/u/2571630322 |
| 中山大学北校区学生会 | http://e.weibo.com/u/2410476754 |
| 中山大学珠海校区学生会 | http://e.weibo.com/u/2376296650 |
| 中山大学团委 | http://e.weibo.com/u/2140532233 |
| 中山大学粤语协会 | http://e.weibo.com/u/1811159122 |
| 中山大学中大青年 | http://e.weibo.com/u/1768001547 |
| 中山大学摄影协会 | http://e.weibo.com/u/2651668322 |
| 中山大学信息技术帮助台 | http://e.weibo.com/u/2355709275 |
| 中山大学-5D空间 | http://e.weibo.com/u/2203552935 |
| 中山大学研究生会 | http://e.weibo.com/u/2095696670 |
| 中山大学逸仙时空BBS | http://e.weibo.com/u/2259379193 |
| 中山大学研究生支教团 | http://e.weibo.com/u/2017780037 |
| 以上是认证微博，以下为普通微博 | |
| 中山大学传设院学生会 | http://weibo.com/u/1896895961 |
| 中大Din | http://weibo.com/u/1994626454 |
| 中山大学逸仙时报 | http://weibo.com/u/1899452321 |
| 中山大学一起跑协会 | http://weibo.com/u/1994389140 |
| 中大GBT | http://weibo.com/u/1848906843 |

补充：

先从几个我知道的确定是中大的人的账户开始  
爬关注，得到关注列表，如关注者1，关注者2，关注者3。。。  
然后查询关注者i是否有我设置的那几个中心微博（有其一就可），则保留此账户，若那几个中心微博一个都没关注的，就删除  
然后查询关注者i的关注中是否有我设置的那几个中心微博

1. 出的账户列表，有用户编号就行，我也不打算对应到微博名的

2. 账户列表中的每个账户的相关数据，这些数据包含“共发布多少条微博”，“关注数粉丝数”，“转发过这个微博账户中微博条目的帐号--次数”，“评论过这个微博账户中微博条目的帐号--次数”（这两个都是一个count的过程），“@其他账户的账户号--对应次数”