项目目的：获取twitter社交网络中，某话题整个传播过程的数据，包括网络拓扑信息，用户信息和推文信息。

限制：

（1）推特以账号为单位限制访问服务器的频率，引入账号资源池和线程池的概念。

假设我们能按照某种规则把任务切分成小任务，我们就可以抽象出一个master模块，从资源池申请可用账号，把账号和任务绑定，按照我们自定义的调度规则交给线程池里的工作线程去执行数据获取工作。

（2）获取整个传播过程的数据属于一个超大任务量的工作，单线程顺序慢，如何引入并发计算。换句话，也就是如对大任务进行切分成小任务，交给工作线程去执行抓取任务，如何根据各工作线程的结果生成最终结果（类比mapreduce）。

主要从两个角度分析如何切分任务（map）？

a)在社交网络的好友网络中，按照广度优先遍历规则去遍历网络时，同一层节点的数据获取工作是可以并发执行的，依赖仅存在与直系祖先节点和直系孩子节点之间，他们不能并发执行，其余节点间是可以完全并发执行的。

b)对于任务种类来说，我们主要存在三种类型的任务：获取网络结构，也就是抓取粉丝（谁关注了你）和关注者（谁关注了你）；用户的个人资料；用户的历史推文。这三类任务间不存在依赖关系可以并发。

\*）具体实现：有三个任务队列：抓取网络拓扑任务队列（任务间存在部分依赖关系），获取用户个人资料的任务队列，获取历史推文的任务队列。Master负责把任务分发到不同队列中，并按照调度规则进行任务调度。

如何整合各个工作线程的计算结果（reduce）？

Master的工作：

a）对于网络结构类的计算结果，只需要汇总就可以了，对于推文信息和个人资料，按照用户id进行分类存储就好了。

B）解析出新的任务，由于我们是按照广度优先遍历的思想进行网络遍历，所以我们需要筛选出需要继续往下遍历节点，没有参与话题传播的节点直接进行剪枝，然后将抽取出的任务放入消息队列中。

（3）千万级别的数据如何去重？

使用bloomfilter进行去重