一、电商黑灰产挖掘（刘倩君）

1.数据情况

      1). 电商黑灰产网站

上周确定了2303个候选网站，本周人工标注了前1000个网站，共142个电商相关的黑产网站。

   2).网站文本数据

对所有候选网站爬取前20个页面的文本，遇到问题：有的网站只爬了几页，有的一页都没爬到，有的爬了几百页都停不下来。

因为我是用了10个爬虫，每个负责爬200多个网站，当一个网站爬到20个页面时，应该停止向这个网站发送请求，但由于一个爬虫要负责200个网站，没法过滤掉特定的请求。试了一下分布式的动态可配置的爬虫，没有成功。

2.算法情况

    黑站识别算法已经实现：基于URL token和网站文本keywords的LR二分类模型。

3.算法效果

    由于数据爬取不全，暂时只有部分结果。

二、微博舆情相关（吕晨阳）

1.数据情况

     目前已有4万个节点，2000万条博文。

希望以阿里当前挖掘到的参与黑产的用户的微博账号作为种子节点，爬取其个人信息、博文信息、评论以及节点的关注人和被关注人，获取150万个节点的微博信息。

2.算法情况

   实验设计：

用BFS挖掘的黑产团体作为baseline，用node2vec结合社交关系和文本信息挖掘黑产团体。

自然语言处理博文，挖掘其中的黑产平台、黑产交流QQ群等信息。

三、遇到的问题

爬取数据困难。后续对微博和对黑站的大规模爬取，也需要专业的爬虫团队，我门把爬虫的需求明确一下，希望能跟相关人员讨论一下。