本周进展

1. 从阿里目前提供的10000条舆情文本数据中提取了约2000条URL。其中有些是黑灰产网站，如

' http://www.zhenshua.com/ '(交流论坛)

'http://t.cn/RbRKaOo'（流量平台）

'http://www.milioo.com'（刷单平台）

'http://www.88dan.com/?uid=393'（空单服务）

'http://www.100estore.com'（店铺转卖）

还有一些是淘宝、考拉海购、京东等的商品链接，还有微博的链接，类似于这样的：

http://t.cn/R6VhICj???

http://t.cn/R6cXgHe??

http://t.cn/8Fv70Jc

http://t.cn/R6cxLlk??

1. 对提取的URL对应的网站进行了内容的爬取，但是只选取了一两个作为start\_urls，还没有进行全量的爬取。

下两周计划：

阅读文本语义分析相关的论文，将舆情文本中的网站命名实体提取出来，为后续的文本处理也做好准备。

优化爬虫，进行全量爬取。