2025-06-06周报

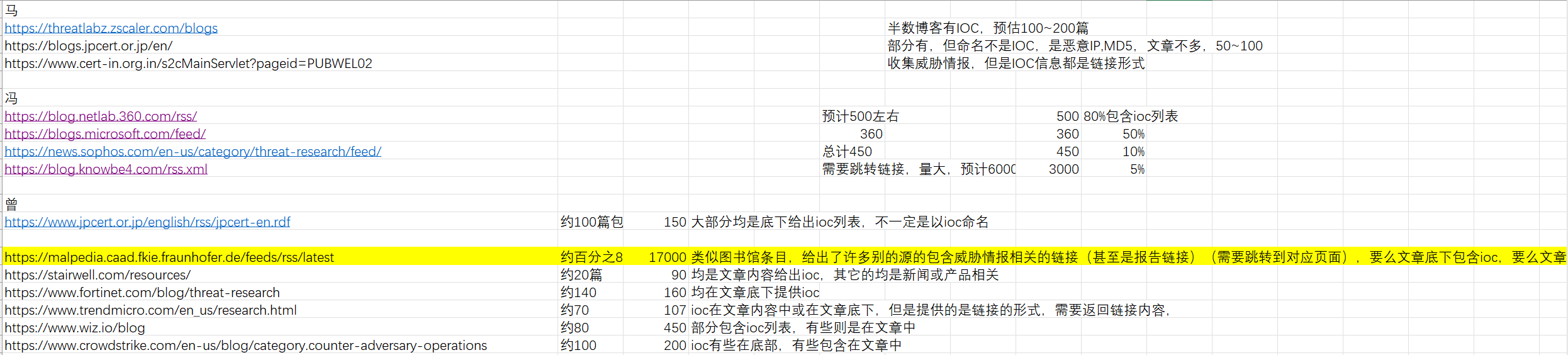
马嘉润

主要工作

1. 筛选数据源，保留文中提到IOC信息的源
2. 对这些数据源进行清洗

问题：

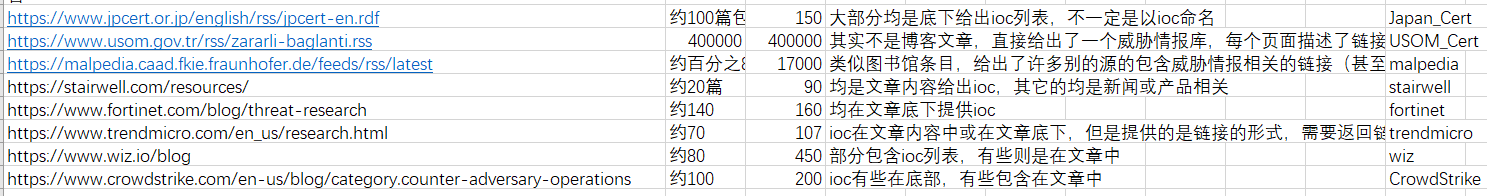
1.非结构化数据源中有效博客很少。



曾家豪-本周周报

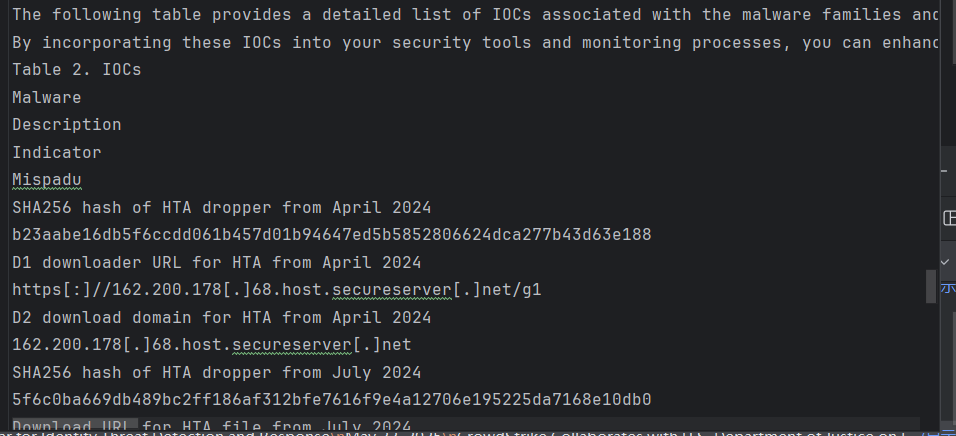
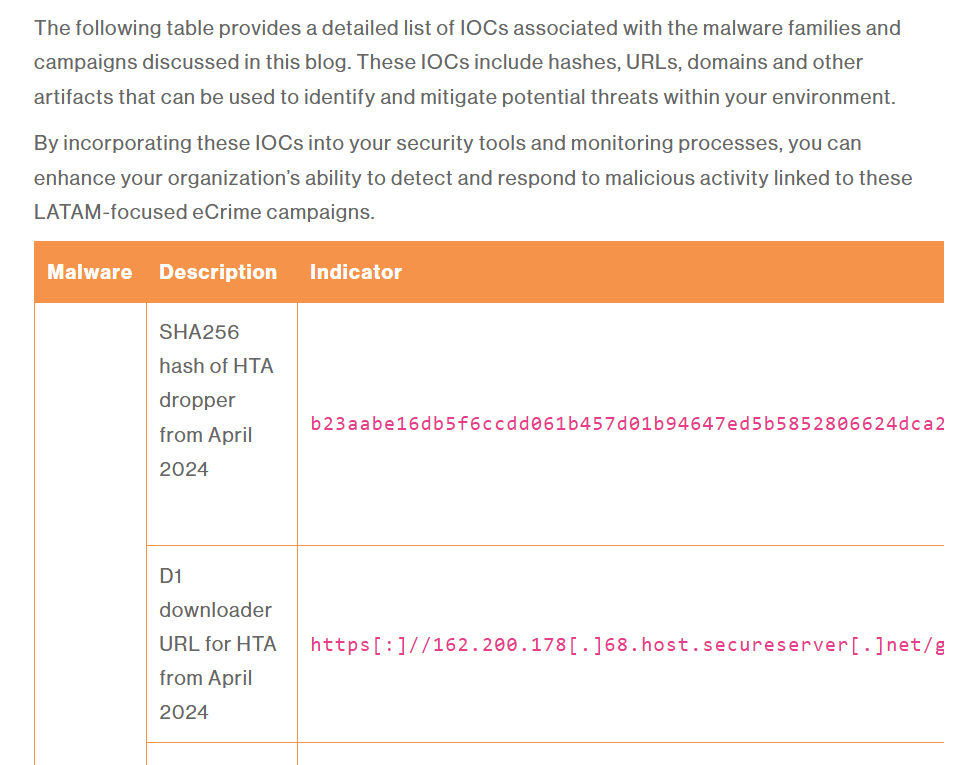
1. 从obs的源筛选大部分有用的源，方便后续从obs中获取对应的原始数据内容

可用源如下：



1. 测试去html标签的代码功能

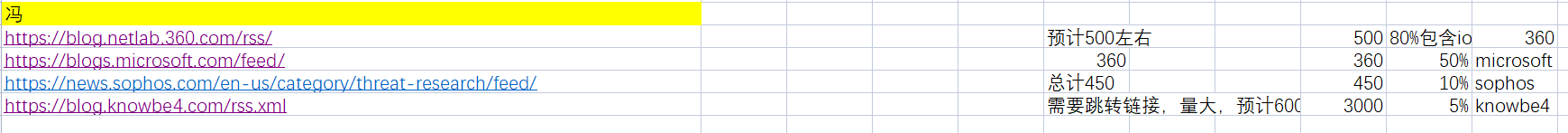
上面是博客底部的ioc数据表，下面是去标签后的结果。



3.配置了一个需要全量爬的源（1w3的数据量）。

冯世玲-本周周报

1. 筛选数据，获得有效源



1. 对数据进行清洗，去除html标签后，判断文本内是否包含ip、url、domain、md5数据，从2031个html文件中筛选出977个有效文本

