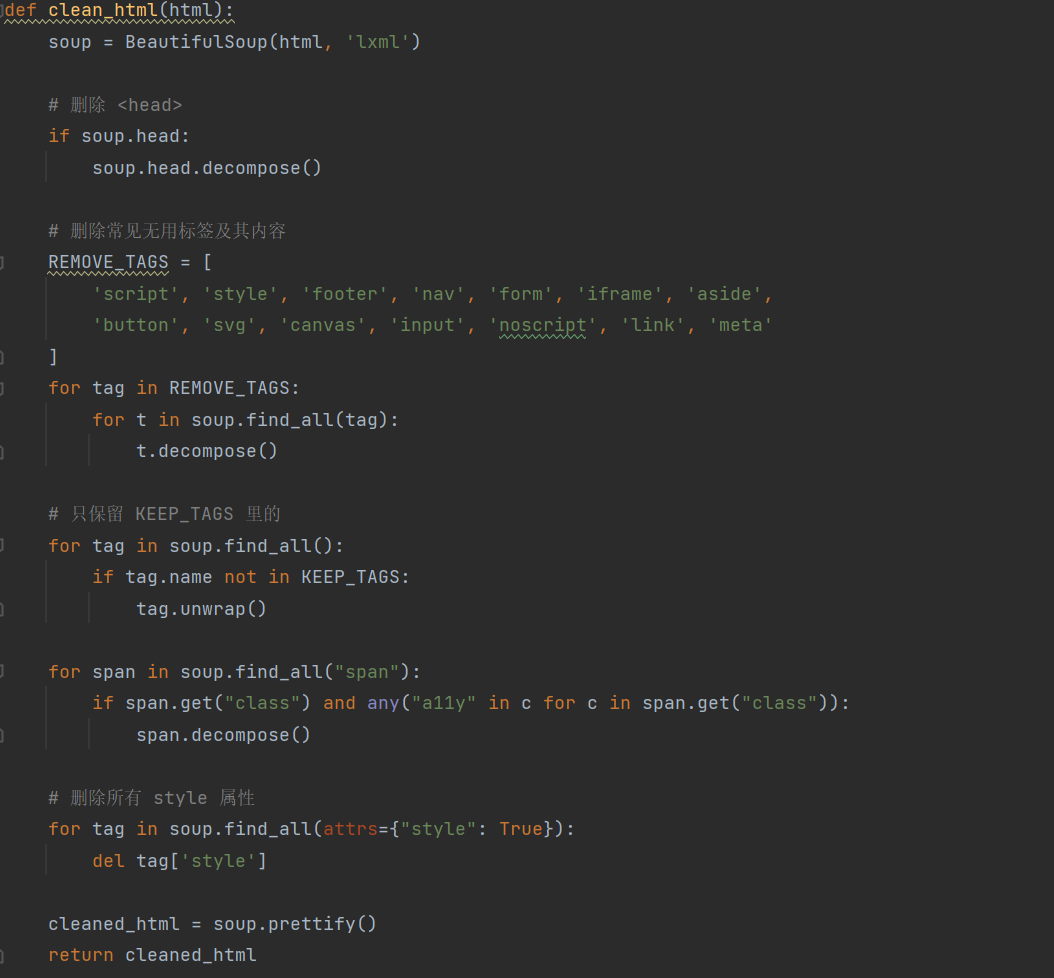
2025-05-29周报

马嘉润——html页面清洗，构建训练集

主要工作

1. 对批量爬取的 HTML 博客页面进行清洗，删除与正文无关的元素，保留正文标签。

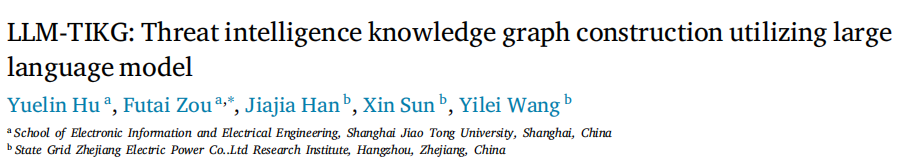


1. 构建训练集，用deepseekR1-671B模型以及prompt来获取input和output。

冯世玲——本周周报

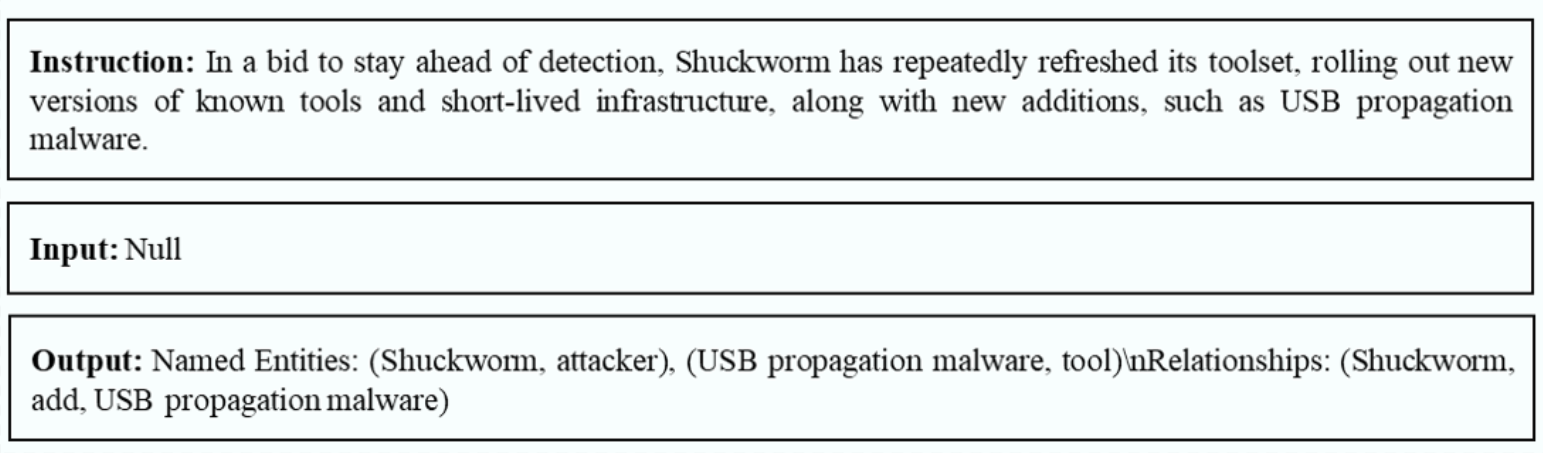
本周调研了具体如何提取实体和关系构建知识图谱

1

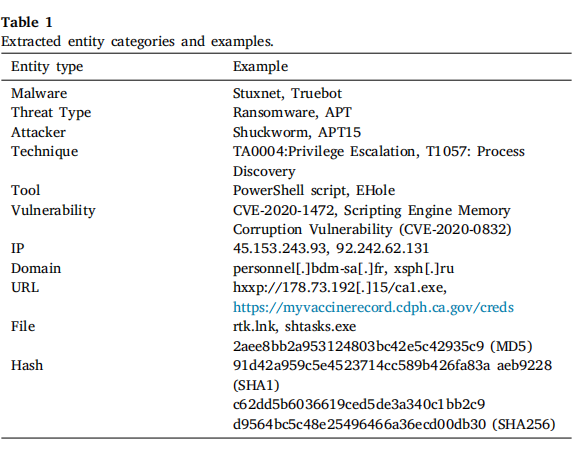


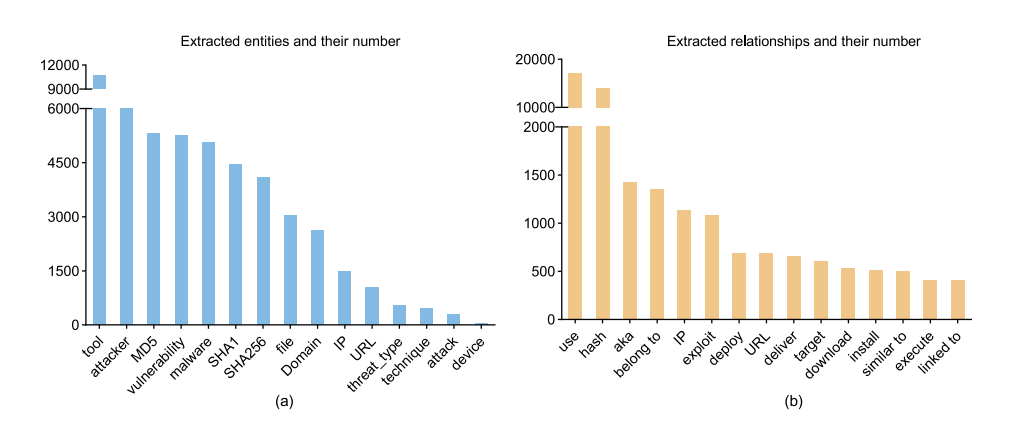
依靠微调的大模型提取实体和关系

用于微调的数据集公开可查

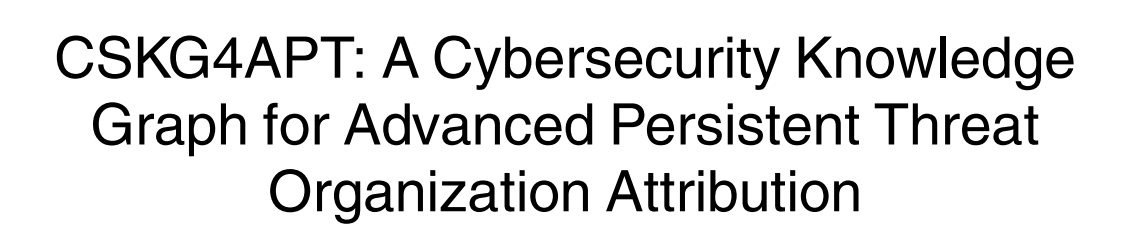


设计的实体和关系如下





2



正则提取和深度学习提取实体

