**基于Stanza模型的病毒知识图谱构建方法**

1. **在互联网上爬取构建所需的知识信息**

数据源网址为：[https://web.archive.org/web/20000815055127/http://vil.nai.com/villib/alpha.asp](https://web.archive.org/web/20000815055127/http:/vil.nai.com/villib/alpha.asp)

在爬取的同时将其储存在数据库和文件( .json )中

1. **通过Stanza提取命名实体与关系并生成三元组**

使用Stanza 英文模型，并配置一个 NLP 处理流水线来进行分词、词性标注、词干化、依存解析和命名实体识别（NER）。

具体步骤如下：

初始化需要处理的文本，并定义实体、关系和属性的空列表。定义的三元组 triples 列表将用于存储最终提取的三元组数据。

使用 stanza 分词和实体识别。实体类型为 ORG、PRODUCT、QUANTITY、CARDINAL 或 DISEASE。过滤出的实体会作为二元组 (实体文本, 实体类型) 添加到 entities 列表关系和属性提取。

扫描句子中的每个词，若匹配到指定的词汇（如 infected、available、free、active、malicious 等），则认为是状态属性；若词汇为 bytes，则识别其前面的词作为大小值。

利用词语的依存关系（obj、nmod、obl 等）提取灵活的主-谓-宾结构，识别谓语动词（如“感染”）并形成 (subject, relation, object) 三元组关系。

将提取的实体、属性和关系整合为三元组。每个实体被格式化为 (实体, "is\_a", 实体类型)；每个属性被格式化为 (属性名, "has", 属性值)，并将已提取的关系三元组直接加入 triples 列表。

总处理方法：process 方法依次调用 extract\_entities、extract\_relations\_and\_attributes 和 create\_triples 三个方法，实现实体、属性和关系的全面提取。

1. **构建知识图谱**

对每条病毒数据进行如下处理：

创建初始病毒节点：每条病毒数据形成的知识图谱三元组列表 knowledge\_graph\_triples 初始化为 [(virus\_name, "is\_a", "Virus")]。

提取和添加属性：从数据中提取以下字段：aliases、discovery\_date、length、origin、risk\_assessment、minimum\_dat、dat\_release\_date、symptoms、method\_of\_infection、removal\_instructions。对于非空值的字段，将 (病毒名, "has\_属性", 属性值) 添加到 knowledge\_graph\_triples 中。

提取病毒特征信息：调用 extract\_virus\_characteristics 函数获取症状和文件长度增加信息，将其作为三元组追加到知识图谱列表中。

进行深入提取：创建 TextProcessor 实例，传入 virus\_characteristics 文本，并调用 process 方法进行实体、关系和属性的提取。生成的三元组直接追加到 knowledge\_graph\_triples 列表中。

实时写入 JSON 文件：对于每个三元组，按行写入 output\_file 文件中，每行表示一个三元组的 JSON 对象，以便后续查询和分析。

实时写入 Neo4j 数据库：通过调用 create\_knowledge\_graph\_in\_neo4j 方法，将三元组实时写入 Neo4j 数据库，以实现数据的可视化展示。

更多详细信息请访问

https://github.com/HoshinoIsumi/KnowledgeGraph-viruses