《喜羊羊与灰太狼》知识图谱构建简介

本项目以动画《喜羊羊与灰太狼》为例，构建其中角色的人物关系知识图谱。项目任务一共分为四个部分：

1）数据爬取与预处理 2）命名实体识别 3）实体关系抽取 4）可视化与知识问答

1 数据爬取与预处理

在构建知识图谱之前，需要从《喜羊羊与灰太狼》维基百科网站和百度百科梗概文本获得信息，以进行实体（人物、场地）和关系的识别抽取，从而构建初步的关系三元组。首先，检查和爬取维基百科源代码格式以及半结构化数据文本，将其输出为包含两个实体一个关系的三元组格式。其次，二次爬取百度百科角色关系，对关系三元组进行补充。最后，清洗三元组的异常数据，生成初步的关系三元组文件，并爬取《喜羊羊与灰太狼》400 集剧情介绍文本作为后续关系抽取的训练数据。

2 命名实体识别

为了更好把数据中爬取的文本作为关系抽取的输入，通过命名实体识别技术对文本进行处理，提取出包含两个实体以上的句子作为关系抽取的输入。首先，基于 BIOE 编码方法，手动标注部分训练数据。其次，训练并比较 BI-LSTM 模型以及 BI-LSTM+CRF 模型对实体的识别效果，选择识别精确度相对最高的 BI-LSTM+CRF 模型作为本实验场景的最终实体识别模型。最后，基于维特比解码预测全部数据，得到最终实体识别结果。

3 实体关系抽取

为了抽取出文本中的实体关系，通过训练 6 种关系抽取模型，进一步提取文本实体关系，并对关系三元组进行补充。首先，基于开源 OpenNRE 工具包的数据输入格式，对实体识别后的 BIOE 编码结果进行格式化处理。其次，调用 opennre 库，分别训练 wiki80_cnn_softmax、wiki80_bert_softmax、wiki80_bertentity_softmax、tacred_bert_softmax 以及 tacred_bertentity_softmax，5 种关系抽取的预训练模型。然后调用哈工大，BERT-wwm，中文 bert，在 20w 中文人物关系数据，训练 chinese-BERT-wwm 模型。最后，将《喜羊羊与灰太狼》格式化处理后的数据，输入 6 种预训练模型进行推理，输出关系抽取结果并分析模型抽取效果，根据关系置信度对结果进行清洗和筛选，将有效的关系抽取结果补充进关系三元组文件中，形成《喜羊羊与灰太狼》的最终关系抽取结果。

4 可视化与知识问答

通过 Neo4j Deskto，对《喜羊羊与灰太狼》的关系抽取结果进行可视化呈现，并实现 QA。