选题：构建小规模领域知识图谱

组队情况

身份	学号	姓名
队长	221900103	许梁超
队员	221900068	张淳皓
队员	221900084	王诗瑶

研究计划

阶段一：数据预处理与实体识别

目标：从原始文本中提取结构化实体并分类（PER/LOC/ORG/MISC）

任务分解

数据清洗与格式转换：解析CoNLL-2003的BIO标注格式，转换为SpaCy可处理的文本与标签对

基线实体识别：使用SpaCy预训练模型en_core_web_lg进行命名实体识别，并评估性能（计算精确率、召回率、F1值）

进阶：微调自己的NER模型进行命名实体识别

技术点：BIO标签解析、SpaCy训练数据格式转换、模型微调策略

交付成果

conll2003_processed.json（清洗后数据）

data_preprocess.py（完成BIO标签解析、数据清洗、SpaCy格式转换）

ner_model（微调后的模型文件）

ner_eval.txt（包含基线模型与微调模型的F1对比）

阶段二：关系抽取与图谱构建

目标：从实体对中提取关系并生成三元组

任务分解

句法分析与规则设计：通过分析句子的语法结构（如主谓宾、介词短语）和预定义规则（如"[人物] + '在' + [组织]"），抽取出实体间的语义关系

三元组生成与存储：将识别出的实体和关系组合成（头实体，关系，尾实体）形式，并保存为结构化文件（如CSV/JSON）

技术点：句法分析技术、规则设计技术、三元组生成技术

交付成果

relation_rules.py（匹配规则脚本）

kg_triples.csv（三元组数据文件）

阶段三：知识表示学习

目标：训练实体与关系的低维向量表示

任务分解

数据准备与模型选择：将三元组划分为训练集/验证集/测试集，使用PyKEEN或OpenKE等框架下的模型

训练与评估：设置训练轮次和批大小，使用某些指标评估模型性能，并保存最优模型参数

探索不同模型：对比TransE（适合层级关系）、ComplEx（处理对称/逆关系）和RotatE（建模复杂关系）在不同关系类型上的表现，记录各模型训练时间和预测准确率

技术点：采样策略、损失函数选择、训练加速技巧

交付成果

kge_train.py（训练代码）

model_comparison.csv（各模型性能对比表）

阶段四：NE图谱动态更新和增量学习

目标：支持新数据增量更新，避免全量重新训练

任务分解：

图谱更新策略：设计基于规则和嵌入相似度的冲突检测机制，开发实体对齐算法，实现新增三元组的自动化合并与版本化存储，确保知识一致性

增量训练方法：使用continual KGE技术，在动态知识图谱中增量学习新知识同时保留历史知识

图演化表示方法：使用Dynamic KGE技术，通过时间编码机制将时序信息融入实体和关系的向量表示中，建模知识图谱随时间演化的表示学习方法

技术点：图谱更新策略、增量训练方法、图演化表示方法

交付成果

incremental_training.ipynb（增量训练代码）

update_log.txt（动态更新日志）