实验二 知识图谱部分

一、实验背景

随着人工智能与大数据的发展，**知识图谱**（Knowledge Graph, KG）逐渐成为表示结构化知识的重要形式。它以**三元组**（head, relation, tail）的形式记录实体之间的语义关联，被广泛应用于推荐系统、搜索引擎、问答系统等领域。然而在真实的大型知识图谱中，三元组通常是不完整的。因此，如何根据已有结构预测缺失的三元组成为了一个重要研究方向。

图谱补全是**知识图谱嵌入（Knowledge Graph Embedding, KGE）**最核心的应用场景之一，其目标是在低维向量空间中学习实体与关系的连续表示，从而对知识图谱中缺失的结构进行推断。具体而言，图谱补全通常以链接预测的方式实现，包含**三类补全任务**：在给定（h, r, ?）时预测尾实体（tail prediction），在给定（?, r, t）时预测头实体（head prediction），以及在给定（h, ?, t）时预测关系类型（relation prediction）。通过这些任务，可以有效提升知识图谱的完备性，并为下游智能系统提供更丰富的结构化语义知识支持。

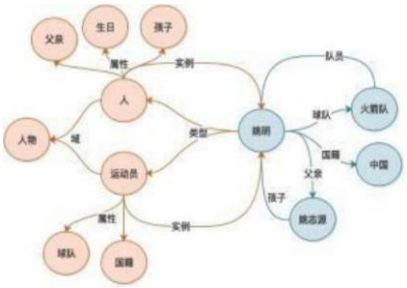
为了进行有效的图谱补全，学术界提出了**基于嵌入表示的算法**，如 TransE、TransR 等，它们能够将实体与关系映射到低维向量空间，通过几何结构表达语义，进而实现对缺失实体的预测与补全。在实验二中，我们希望各位同学从公开图谱中随机**抽取一定规模的知识图谱子图**，并对抽取到的图谱进行处理与**数据集划分**，进而**基于 TransE 模型完成链接预测任务**，体会从“**原始图谱 → 子图构建 → 嵌入学习 → 链接预测评估**”的完整流程。

二、实验背景

（1）Freebase 数据集介绍

Freebase 是一个由元数据组成的**大型合作知识库**，内容主要来自其社区成员

的贡献。它整合了许多网上的资源，致力于打造一个允许全球所有人（和机器）快捷访问的资源库。Freebase 提供数据查询和录入机制。其官网（https://developers.google.com/freebase）提供 N-TripleRDF 格式的数据压缩包的下载，但请注意整个压缩包 30G，解压后 300G+。有关 Freebase 的更多信息可参考相关介绍文章[1]。



本实验中我们采用 freebase 在电影领域的**中等规模图谱 freebase_movie.gz**，以（头实体 URL，关系 URL，尾实体 URL）这种**三元组**的形式进行保存。其中，数据集的头实体与尾实体均采用 Freebase 的官方 MID（Machine Identifier）格式表示，**MID** 是 Freebase 为每个**实体**分配的唯一、稳定且机器可读的标识符，例如实体 URL“<http://rdf.freebase.com/ns/m.03jt67r>”中的 “**m.03jt67r**”；数据集中的**关系**采用 freebase 原生 schema 的关系表示方式，以 /domain/type/property 的 schema 路径形式出现，如下图关系 URL “http://rdf.freebase.com/ns/file.performance.film”中的“**file.performance.film**”。

不过，在本实验中，我们并不关注 freebase 中 MID 或关系路径的具体命名规范。我们仅利用其核心特性，即**所有实体与关系均具有由 freebase 设计的稳定且唯一的标识符**，并基于此进行我们的子图抽取任务。

以下为数据集中的数据样例，每一行为一个三元组（h，r，t）。

```
<http://rdf.freebase.com/ns/m.03jt67r>  <http://rdf.freebase.com/ns/film.performance.film>  <http://rdf.freebase.com/ns/m.01vg3r>  .
<http://rdf.freebase.com/ns/m.045y4ww>  <http://rdf.freebase.com/ns/common.webpage.topic>  <http://rdf.freebase.com/ns/m.06gjk9>  .
<http://rdf.freebase.com/ns/m.0599qd8>  <http://rdf.freebase.com/ns/base.wfilmbase.siteid.film> <http://rdf.freebase.com/ns/m.033fqh>  .
<http://rdf.freebase.com/ns/m.059x0w>  <http://rdf.freebase.com/ns/film.producer.film> <http://rdf.freebase.com/ns/m.0dllxnf>  .
<http://rdf.freebase.com/ns/m.059x0w>  <http://rdf.freebase.com/ns/film.producer.film> <http://rdf.freebase.com/ns/m.047rkcm>  .
```

（2）TransE 算法简介

TransE（Translating Embeddings for Modeling Multi-relational Data）算法[2]是由 Google 于 2013 年提出的**经典知识图谱嵌入**（Knowledge Graph Embedding, KGE）模型，被广泛视为知识图谱表示学习的奠基性工作之一。其核心思想是：将知识图谱中的实体（entity）和关系（relation）映射为低维向量，通过几何空间中的**“平移操作（translation）”来刻画三元组（head, relation, tail）的语义**

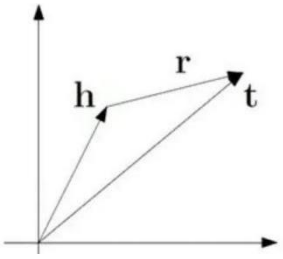
[1] https://developer.aliyun.com/article/717320?spm=a2c6h.14164896.0.0.535f3630po4hs1

[2] Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." Advances in neural information processing systems 26 (2013).

结构。对于一个正确的三元组（h，r，t），其在向量空间里的关系应该为：

$$\vec{h} + \vec{r} \approx \vec{t}$$

其刻画的语义结构类似下图表示：



于是，TransE 算法使用实体与关系的**距离**作为分数来衡量三元组的合理性：

$$f(h,r,t) = \| \boldsymbol{h} + \boldsymbol{r} - \boldsymbol{t} \|_2$$

得分越小，表示三元组越可能为真。常用距离包括 L1 和 L2 范数。

TransE 算法基于简单的**关系平移假设**，不断训练从而调整表征，最终得到了关系和实体在嵌入空间的良好表示从而实现链接预测，即自动推测缺失的关系或实体补全知识图谱。由于其结构简单、计算效率高、易于扩展，TransE 已成为知识图谱补全任务中最经典、最基础的嵌入模型之一，也为后续复杂模型（如 TransH、TransR、RotatE 等）提供了理论基础。

经典的 TransE 算法采用 Hinge loss 作为损失函数。它是一类典型的基于负采样（negative sampling）的 margin-based 损失，**其核心思想是迫使模型将正三元组的得分提升，同时将负三元组的得分压低，并保持二者至少相差一个给定的** `margin`。知识图谱训练集中

$$\mathcal{D} = \{(h_i, r_i, t_i)\}_{i=1}^{N}$$

全部是经过人工或规则系统验证的正确事实，**因此没有自然出现的“负三元组”**。然而知识图谱补全**要求模型具备区分真实关系与虚假关系的能力**，因此需要在训练过程中**主动构造负样本**。

在**本实验中，负样本由正样本随机替换尾实体得到。**具体而言，对于每一个正三元组(h_i, r_i, t_i)，我们从实体集合$\mathcal{E}$中随机采样一个与t_i不同的实体t_i^-。为避免采样到图中真实存在的事实，我们要求：

$$t_i^- \notin \{t \mid (h_i, r_i, t) \in \mathcal{D}\}$$

即**负样本必须不属于当前头实体与关系$(\mathbf{h_i}, \mathbf{r_i})$所对应的所有真实尾实体集合**（否则采到的还是正确的尾实体，即属于正样本）。通过排除潜在正样本，可以确保所构造的三元组在语义上必为错误，从而提高负采样的有效性。

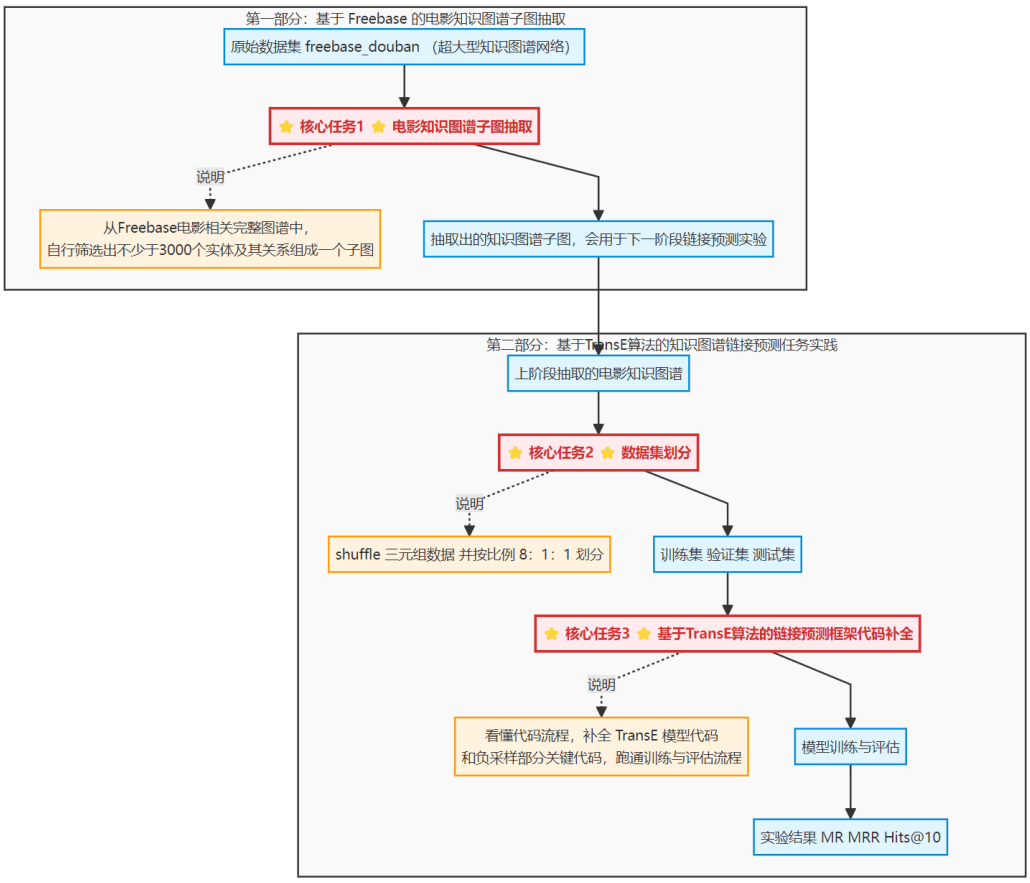
在本实验中，我们改用了同为 pair-wise 损失的 **BRP loss 函数**，它的优化目标是让正样本的得分显著高于负样本，从而提升模型对真实三元组与虚假三元组的区分能力。其形式如下：

$$\mathcal{L}_{BPR} = -\sum_{i=1}^{N} log\ \sigma(f(h_i, r_i, t_i) - f(h_i, r_i, t_i^-))$$

其中，$\sigma(\cdot)$为 sigmoid 函数，(h_i, r_i, t_i^-)为正样本(h_i, r_i, t_i)的负采样三元组。将二者代入 BPR Loss 中进行反向传播，以更新实体与关系的嵌入表示。通过这种方式，TransE 的几何平移假设得以在 BPR 框架下进行优化，使得实体与关系嵌入在训练过程中不断调整，从而更准确地反映知识图谱中的语义关联。

三、 实验流程

本实验基于真实世界中的大型知识图谱网络 Freebase 数据集，首先从中随机筛选出具有一定规模的知识图谱（子图），进而在该图谱上自行划分数据集，进行电影知识图谱链接预测任务的实践探索。参考的实验流程如下图所示：



具体而言，本次实验的目标可以概括为**两个部分**：

第一部分：基于 Freebase 的电影知识图谱子图抽取

- **任务 1-电影知识图谱子图抽取**：从 Freebase 电影相关完整图谱中，自行筛选出不少于 3000 个实体及其关系组成一个子图。

部分说明如下：

- 为保证质量，最好只保留具有< http://rdf.freebase.com/ns/前缀的实体。
- 一般而言，较为稠密的子图可以保留更多结构关联信息，也可以更有效地支撑图谱补全、图谱推荐等下游任务。如果学有余力，可以比较一下不同稠密程度的子图在下游任务上的性能差异。
- 为保障图谱的质量，也可以根据统计对不常出现的实体或关系进行筛选。例如，可以过滤掉涉及三元组少于 10 个的实体，或只保留至少在 10 个三元组中出现的关系等。
- 为保证提供的源代码顺利运行，请**自行将所有的实体与关系 ID 分别映射为从 0 开始的连续编号**，例如：

	entity
0	m.09gb_4p
1	m.0dgs73j
2	m.09gq0x5
3	m.0211pk
4	m.0bwky98

	relation
0	type.type.instance
1	film.film_crew_gig.film
2	film.film_regional_release_date.film
3	film.performance.film

并将三元组处理成类似下图的格式：

```
578 0 8
579 0 142
579 0 144
579 0 143
580 1 23
581 2 434
```

第二部分：基于 TransE 算法的知识图谱链接预测任务实践

- **任务 2-数据集划分**：将抽取到的三元组划分为训练集、验证集和测试集。

部分说明如下：

- 可采用最基本的“shuffle 数据集+按 8：1：1 划分数据集”的方式。
- 如你选择使用我们提供的代码流程，为使其正常读取数据，请将划分好的三个数据集命名为 train.txt，valid.txt，test.txt（txt 每行三元组内的实体关系以空格隔开）；并保存在 data/freebase/文件夹下。

- **任务 3-基于 TransE 算法的链接预测框架代码补全**：在给定框架代码基础上，补全负采样逻辑与 TransE 模型代码，训练知识图谱嵌入模型；在测试集上

进行链接预测评估，观测并分析结果。

部分说明如下：

- 在本次实验中，我们已经给出了基本的参考代码，但部分代码需要大家自行补全。**具体要求详见 4.2 代码文件说明部分**。注意：提供的代码框架仅作为一种参考方式，可在上面做任何修改。实验验收仅以跑通为准，不要求代码和我们一致。鼓励大家在给定代码基础上自行探索，或者自己搭建框架探索。
- 部分内容为选做实验，仅供学有余力的同学拓展学习使用，不作为计分要求。
- 得到图谱补全（链接预测）结果后，请根据自行划分的测试集评估效果优劣。可以采用 MRR、Hits@K 等指标。由于数据集与评估指标均为自行选择，所以指标优劣不作为给分依据，请不要相互比较，以免造成**不必要的内卷**。

四、相关文件

（1）数据集文件说明

- **freebase_movie.gz：**

该文件为 freebase 在 movie 领域的中等规模图谱。**每行以（头实体，关系，尾实体）三元组的形式进行保存**。此文件为完整的知识图谱数据集，需要在此基础上进行子图的抽取与构建。因原体积（52G）过大，采用压缩形式进行存储，**如无必要请勿解压。可参考如下方式使用：**

```python
import gzip

with gzip.open('./web/freebase_douban.gz', 'rb') as f:
    for line in f:
        line = line.strip()
        triplet  = line.decode().split('\t')
        print(triplet[0:3])
        break
```

以下为数据的样例，每一行为一个三元组（h，r，t）。

```
<http://rdf.freebase.com/ns/m.03jt67r>  <http://rdf.freebase.com/ns/film.performance.film>  <http://rdf.freebase.com/ns/m.01vg3r>  .
<http://rdf.freebase.com/ns/m.045y4ww>  <http://rdf.freebase.com/ns/common.webpage.topic>  <http://rdf.freebase.com/ns/m.06gjk9>  .
<http://rdf.freebase.com/ns/m.0599qd8>  <http://rdf.freebase.com/ns/base.wfilmbase.siteid.film> <http://rdf.freebase.com/ns/m.033fqh>  .
<http://rdf.freebase.com/ns/m.059x0w>   <http://rdf.freebase.com/ns/film.producer.film> <http://rdf.freebase.com/ns/m.0dllxnf>  .
<http://rdf.freebase.com/ns/m.059x0w>   <http://rdf.freebase.com/ns/film.producer.film> <http://rdf.freebase.com/ns/m.047rkcm>  .
```

（2）代码文件夹说明

- **KG_Link_Prediction_TransE ：**

该文件夹包含基于 TransE 模型的知识图谱链接预测任务的完整框架流程（含数据加载、模型定义、模型训练、评估等完整流程），各代码文件具体情况详述如下，其中内部有些地方需要同学们进行代码补全。（需要补全的模块在代码中有一些注释提示，你可以找所有报错的地方或搜索“TODO”。**按要求补全代码即可跑通**。）

代码简介：

- main_kg.py：负责整个实验流程，包括加载数据、构建并训练模型、在验证和测试集上评估模型性能，并保存训练结果。
- parser_kg.py：负责定义和解析实验运行时使用的所有参数，例如学习率、批大小、训练轮数、设备设置和文件路径等。（可在此处探索不同超参数对模型效果的影响。）
- 【**包含 TODO 部分**】 loader_kg.py：负责读取知识图谱数据文件，添加反向三元组补充图谱结构，并在训练过程中为模型生成正样本和负样本。
  - 在本实验中，负样本由正样本随机替换尾实体得到。
  - 【选做】有余力的同学也可探索由正样本随机替换头实体进行负采样，或通过正样本随机替换关系类别，进行“关系判别”的实验。
- 【**包含 TODO 部分**】 KG_embedding_model.py：负责实现知识图谱嵌入模型，包括实体和关系的向量化表示、TransE 或 TransR 的打分函数以及对应的 BPR 损失函数。
  - 【选做】学有余力的同学，可考虑改换其他的知识图谱表征模型，如 TransR、TransH、DistMult 等。其中，如想实现 TransR 模型，可直接补全 KG_embedding_model.py 的 TransR 部分。
- metrics_kg.py：负责计算链接预测任务中的评价指标，例如 MR、MRR 和 Hits@K 等，用于衡量模型的预测性能。
- log_helper.py：负责日志系统，包括创建日志文件和输出训练过程中的信息，

便于监控训练进度和调试。

- model_helper.py：负责模型的保存与加载，包括保存模型参数和从已有的检查点恢复模型。

※ 切记：实验指标高低不作为评分的必然要求。同时，如加了某些优化却发现指标下降也是正常现象。但即使失败或效果不升反降，也鼓励同学们在实验报告里体现自己做的尝试与探索。

※ 本次实验的数据及代码文件可从以下链接下载。

链接：https://rec.ustc.edu.cn/share/b4620980-cc25-11f0-bff0-8763cb145930

密码：web2025

附：提交说明

本次实验要求分组完成，每组最多 3 人（可以少于 3 人，但无优惠政策）。

请于截止日期（2025 年 12 月 31 日晚 23:59）前提交到课程邮箱 ustcweb2025@163.com，具体要求如下：

1. 【重要】邮件标题以及压缩包命名为"组长学号-组长姓名-实验 2"格式。请在**邮件正文、实验报告中**都列出小组所有成员的姓名、学号。
2. 因未署名造成统计遗漏责任自行承担，你可以将邮件抄送你的队友。
3. 实验报告请务必独立完成，如果发现抄袭按 0 分处理。
4. 迟交实验将不被接收。