# 实验二描述文档

## 实验背景

知识图谱是一种特殊的图数据，图中的结点代表实体，边代表关系。一个知识图谱可以由一系列 <头实体，关系，尾实体> 三元组来表示。通常情况下，知识图谱中的三元组并不完整，我们需要根据已知的三元组，来推测出残缺甚至未知的三元组。

## 实验目的

本实验要求在给定的带有文本描述的知识图谱数据集上，设计一种知识图谱补全算法。你需要在给定的测试集上，预测出缺失的三元组的尾实体。

## 任务要求

在本实验中，你需要实现一种利用文本信息的知识图谱的补全方法。此处不限制以何种方法或何种算法使用文本内容，你可以自由探索（如使用 word2vec 计算语义的相似度）。你需要将你的预测结果，提交到指定的网站上，网站会实时计算出相关指标。在规定的时间内，你最多提交 10 次。

你可以自由的设计如何使用数据集所提供的文本信息，但应该尽可能在实验报告中详细的阐述你设计的动机和实现的方式。如果有多种方法，你可以由模型简单到复杂逐步阐述你对模型迭代的动机。并在实验报告中附上每种方法（或不同参数）下的分数。最后，你需要在实验报告中附上你所有提交结果的截图（网站会列出），并在附件中提交你最高分数预测文件。

**可选项：**

1. 除了利用文本信息，你还可以使用知识图谱的图信息/结构信息来预测缺失三元组，如 TransE 方法。如果你感兴趣，可以使用最后提供的相关工具包进行尝试。

**提示：**

1. 因为提交的次数较少，为了不浪费宝贵的提交次数，你可以在线下比较你所设计的方法。因此需要将数据集中的**已知三元组**（训练集和验证集）中的一部分，划分为**线下测试集**。此处你可以自由划分，如按一定比例（70%/15%/15%）将**已知三元组**划分为测试集/验证集/训练集，也可以根据不同关系下的三元组比例划分（分层采样），或者采用交叉验证的方式。但应注意尽量避免划分不均衡导致数据分布产生较大差异，进而影响方法的评估效果。
2. 在线下有提升效果不一定意味着提交的结果有提升，这有可能是已知三元组和待预测的三元组的分布差异导致的，也有可能是线下测试集划分导致的。

## 数据集介绍

实验二提供了一个知识图谱补全数据集，由以下几个文件组成：

1. entity_with_text.txt 包含文本表述的实体信息，每一行是一个实体。其中第一个数字表示实体的 id，后面用\t（制表符）分割，紧接着是一个 token（符号）序列，表示实体的文本描述，每个 token 使用空格分割。
2. relation_with_text.txt 包含文本表述的关系信息，每一行是一个关系。结构和 entity_with_text.txt 类似。
3. train.txt 训练集文件，每一行表示一个三元组，第一个数字表示头实体的 id，第二个数字表示关系的 id，第三个数字表示尾实体的 id，三个数字用\t（制表符）分割。
4. dev.txt 验证集文件，结构和 train.txt 类似。
5. test.txt 测试集文件，结构和 train.txt 类似，其中的尾实体为要求预测的实体，以“？”表示。

下载地址：

链接：https://rec.ustc.edu.cn/share/27cb0d60-3fc8-11ec-b3f7-c7868b6eac26 密码：lab2

**提示：**

1. 测试集中待预测的 <头实体，关系，？> 可能已经出现在训练集中。比如，测试集中第 8114 行要求预测 <12334,7,?>，但在训练集的第 52044 行，66276 行分别出现了 <12334,7,2418>, <12334,7,12287>。因此，在预测时应该排除这些已经出现的尾实体，在剩余的实体中做预测。
2. 所有的文本描述均已经转换为 id，一个 id 可能表示一个单词，一个词组或一个子词（如 playing 可以拆分为 play 和 ing 两个子词。

## 提交要求

### 实验报告及文件提交要求

请以如下文件目录结构组织相关文件结构：

```
exp2/
|----src/
     |----method1
|----submit/
     |----best_result.txt
|----实验报告.pdf
|----README
```

其中，各目录/文件具体要求如下：

- src 目录下放置你的源代码文件，其中至少有一个文件夹，method1 表示你设计方法的源代码，如果有多种不同的方法，每种方法单独放在一个文件夹中，每种方法下的文件内容可以自行组织。文件夹的名称也可以自行命名，但应当在 README 或实验报告中说明。
- submit 目录放置你的最终的预测结果。**文件行数应和数据集的 test.txt 行数相同**。
- 实验报告.pdf 应包含对你设计的算法，同时请用**实验数据**说明你的实现或设计的算法在结果上的差异，并适当**分析**造成这种现象的**原因**。其余应当包含的内容可参见任务要求。请在实验报告中注明所有组成员的学号和姓名。
- README 文件中包含你的源代码的运行环境、编译运行方式，以及对关键函数的说明。同时，对于所作要求之外的文件，也请在 README 中注明这些文件的含义。
- 你不需要上传的模型文件，但应当保证最佳结果可以复现。

## 网站提交要求

你需要提交一个文本文件（以 txt 为文件后缀），该文本文件应满足：

1. 每一行有 5 个实体 id，表示对于缺失的三元组，预测的可能的尾实体。id 用英文逗号分割，不包含额外的空格符号。5 个实体 id 的预测的概率应为从高到低依次排列。
2. 文件的行数应该和测试集的行数相同，且末尾不应该包含额外空行。
3. 如果你提交的文本文件不满足上述要求，本次提交将认定为无效，且不消耗提交次数。

所采用的评价指标为 Hit@N，计算方法为

$$\text{Hit@N} = \frac{\#\text{hit}_N}{\#\text{test}}$$

其中$\#\text{hit}_N$表示在考虑预测结果前 N 个 id 的情况下，命中预测结果的数量，#test 表示测试集的数目。此处考虑 N=1 和 5。**主要关注 Hit@5。**

**例如：**

假设有 5 个需要预测的尾实体

| 正确答案 | 预测结果 | 是否在第一位命中 | 是否在前五位命中 |
|---|---|---|---|
| 1 | **1**,2,3,4,5 | **1** | **1** |
| 2 | 1,**2**,3,4,5 | 0 | **1** |
| 3 | 1,2,**3**,4,5 | 0 | **1** |
| 4 | 1,2,3,**4**,5 | 0 | **1** |

| | | | |
|---|---|---|---|
| 5 | 1,2,3,4,6 | 0 | 0 |

$$\text{Hit@1} = \frac{1+0+0+0+0}{5} = 0.2$$

$$\text{Hit@5} = \frac{1+1+1+1+0}{5} = 0.8$$

## 提交说明

以 PDF 或 DOC 格式提交，实验报告提交文件及邮件标题命名格式统一为“ 学号 1_姓名 1_学号 2_姓名 2_web 实验二”。

- 例如：“ PB19111888_法外狂徒张三_PB19010999_懂法狂魔李四_web 实验二”。
- 两人一组，单人进行也可，但无优惠政策，单人请按“ 学号_姓名_web 实验二” 格式提交。
- 标题须写明小组全部成员学号及姓名，也请在文中注明学号及姓名。
- 因未署名造成统计遗漏责任自行承担。
- 实验报告请务必独立完成，如果发现抄袭按零分处理。
- 迟交作业将不再被接收。

请于 **2021 年 12 月 12 日 23:59 之前**打包成 zip 后提交至课程邮箱 ustcweb2021@163.com ，过期不候。

如有未尽事宜，将对本说明进行进一步更新。

## 相关资源

如果你使用 Python 想使用深度学习的方法，并且需要 GPU 计算资源，可以使用线上提供的免费算力，但绝大多数均有周时间限额。

1. AI Studio：深度学习框架仅支持 PaddlePaddle，显卡为 V100。
2. Colab：默认 TensorFlow，但可以安装 Pytorch，显卡为 K80 或 P100。
3. Kaggle：可以自行安装深度学习框架，显卡为 P100。

如果你想探索如何使用知识图谱图上的结构信息来预测缺失的尾实体，可以参考下面的工具包。

1. OpenKE：是一个开源的知识表示学习工具包，有 TensorFlow 和 Pytorch 版本。