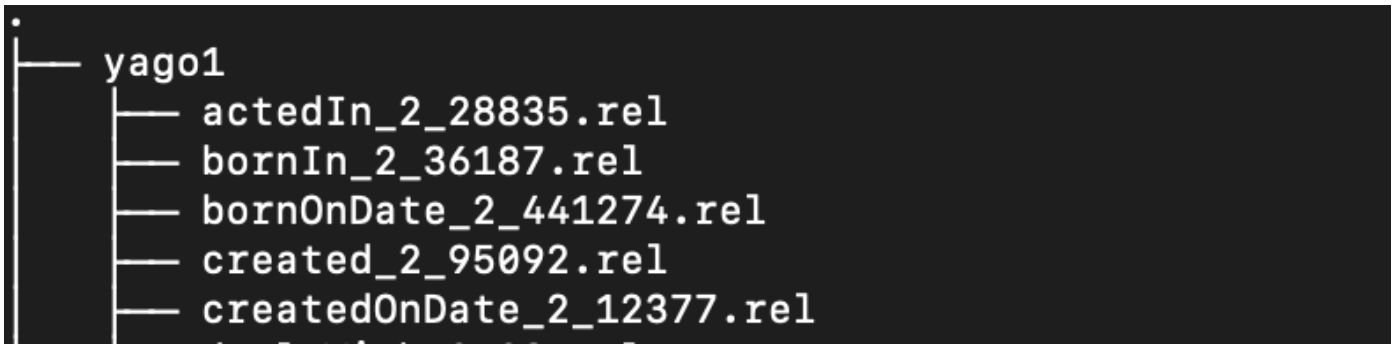
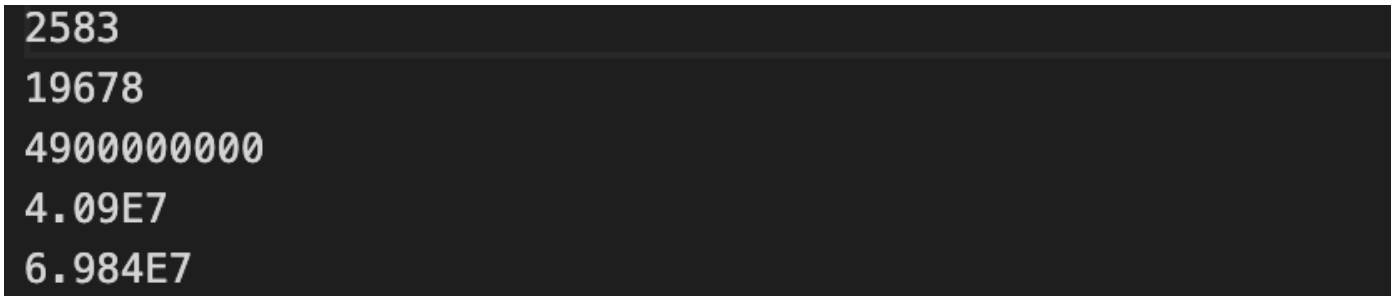
Metadata提取脚本使用说明

数据准备工作

1. 将提取数据集转为`numeratedkb`标准格式，下图为yago1的提取后的文件结构示例(relation和map的文件都存放在yago1(换为其他dataset的名字)的文件夹下)



2. 若对应数据集有`reified`特性，则需先将每个record的index或者除了record之外的所有integer存入一个文件中，格式如下（一行一个数字）



特别注意：判断一个字符串是否是integer的函数应使用定义在`ExtractMetadata.py`中的`isdigit`函数，避免因不同的判断方法导致的数据遗漏的问题。

运行命令与参数说明

要运行脚本，只需到`scripts`文件夹下运行`ExtractMetadata.py`文件即可：`python3 ExtractMetadata.py [Options]`

```
usage: python3 ExtractMetadata.py [-n <name>] [-p <path>] [-i <i>] [-m <indexmode>] [-ipath <indexpath>]
 -n,--name <name>            dataset的名字，应当与存放其数据的文件夹名一致
 -p,--path <path>            dataset所在文件夹的相对路径或绝对路径
 -i,--index <i>              传入1或者0，0代表该数据集不支持reified特性，1代表支持
 -m,--indexmode <indexmode> 传入1或者0，1代表将所有index放入内存，0代表将所有非index的interger放入内存（主要根据不同数据集这两者的集合大小关系来选择），结合数据准备中提到的需要预先提取index信息，若之前提取的是dataset的index，则这里一定要设为1，相反，若是提取的是所有的非index的integer（～index），则这里要设为0
 -ipath --indexpath <ipath> 提取的index信息文件的绝对路径或者相对路径
```

特殊说明

由于脚本只在yago1数据集上进行过正确性测试，再加上不同数据集可能有不同的特性，在提取过程中可能会遇到问题，在`ExtractMetadata.py`文件中，与yago1特性有关的相关处理代码处，我都加了TODO的注释，若遇到报错或者提取结果明显不合理的情况，可以从这些地方下手调整。