KDD 竞赛任务

1 目标

给定作者 ID 和论文 ID，判断该作者是否写了这篇论文。

2 数据集描述

1. 作者数据集: **Author.csv**。包含作者的编号（Id），名字（Name），单位（affiliation）信息。相同的作者可能在 Author.csv 数据集中出现多次，因为作者在不同会议 / 期刊上发表论文的名字可能有多个版本。例如：J. Doe, Jane Doe, 和 J. A. Doe 指的均是同一个人。

字段名称	数据类型	注释
Id	int	作者编号
Name	string	作者名称
Affiliation	string	隶属单位

2. 论文数据集: **Paper.csv**。包含论文的标题(title), 会议 / 期刊信息, 关键字(keywords)。同一论文可能会通过不同的数据来源获取，因此在 Paper.csv 中会存在多个副本。

字段名称	数据类型	注释
Id	int	论文编号
Title	string	论文标题
Year	int	论文年份
ConferrenceId	int	论文发表的会议 Id
JournalId	int	论文发表的期刊 Id
Keywords	string	关键字

3. (论文-作者)数据集: **Paper-Author.csv**。包含 (论文 Id-作者 Id)对 的信息。该数据集是包含噪声的(noisy)，存在不正确的(论文 Id-作者 Id)对。即，在 Paper-Author.csv 中的(论文 Id-作者 Id)，该作者 Id 并不一定写了该论文 Id。因为，作者名字存在歧义（存在同名的人），和作者名字存在多个版本（如上面的例子：J. Doe, Jane Doe, 和 J. A. Doe 指的均是同一个人）。

字段名称	数据类型	注释
PaperId	int	论文编号

字段名称	数据类型	注释
AuthorId	int	作者编号
Name	string	作者名称
Affiliation	string	隶属单位

4. 会议和期刊数据集：**Conference.csv**, **Journal.csv**。每篇论文发表在会议或者期刊上。

字段名称	数据类型	注释
Id	int	会议 / 期刊 编号
ShortName	string	简称
Fullname	string	全称
Homepage	string	主页

5. 共同作者的信息：**coauthor.json**。目前，coauthor.json 文件给出每个作者合作频率最高的 10 个共同作者，该文件的格式为 json。coauthor.json 文件的内容格式形如：

```
{"A 作者 ID": {"B1 作者 ID": 合作次数, "B2 作者 ID": 合作次数}}
```

   第一层的 key 为作者的 ID，对应的 value 为共同作者信息（同样为 key-value 形式，key 为共同作者的 ID，value 为合作次数）。

6. 论文&作者 pair 字符串信息：**paperIdAuthorId_to_name_and_affiliation.json**。文件内容是从 Paper-Author.csv 提取的, 将 Paper-Author.csv 中相同的论文 ID 和作者 ID 对的 name 和 affiliation 合并，存储为 key-value 形式，key 为论文 ID 和作者 ID 对：'paperid|authorid', value 为 {"name": "name1##name2##name3", "affiliation": "aff1##aff2##aff3"}。

7. 训练集：**Train.csv**。其中 ComfirmedPaperIds 列对应的论文，表示该作者写了这些论文。DeletedPaperIds 列对应的论文，表示该作者没有写这些论文。

字段名称	数据类型	注释
AuthorId	int	作者 ID
ComfirmedPaperIds	string	以空格分割的论文(PaperId) 列表
DeletedPaperIds	string	以空格分割的论文(PaperId) 列表

8. 验证集：验证集 **Valid.csv** 文件的格式如下：

字段名称	数据类型	注释
AuthorId	int	作者 ID
PaperIds	string	以空格分割的论文(PaperId) 列表，待测的论文列表

9. 验证集答案：**Valid.gold.csv** 是验证集的标准答案，文件格式与训练集 Train.csv 格式相同。

10. 测试集：**Test.csv**。测试集 Test.csv 文件的格式与验证集 Valid.csv 格式相同，将在之后发布。测试文件命名为 Test.##.csv，其中##为各个小组的编号，如 Test.01.csv 表示第一个小组的测试集。

11. 因此，各个小组最终需要提交的是**测试集预测结果**，提交文件的格式与 Valid.gold.csv 相同。文件命名为 Test.P##.csv，其中##为各个小组的编号，如 Test.P01.csv 表示第一个小组提交的测试集预测结果。

12. **数据集的统计**

数据集	(作者-论文)对 个数
训练集（Train.csv）	11,263
验证集（Valid.csv）	2,347
测试集（Test.csv）	每个队伍的测试集不同，约 1,300;

3 数据目录介绍

data
dataset: 数据目录
train_set: 训练集文件夹
- Train.authorIds.txt: 训练集。的所有作者列表
- Train.csv：训练集

valid_set: 验证集文件夹
- Valid.authorIds.txt: 验证集的所有作者列表
- Valid.csv：验证集
- Valid.gold.csv：验证集的标准答案

test_set: 测试集文件夹（各个小组不同的测试集）
- Test.authorIds.txt: 测试集的所有作者列表，如 Test.01.authorIds.txt 是第一小组
- Test.csv：测试集，如 Test.01.csv 是第一小组的测试集

Author.csv: 作者数据集
coauthor.json: 共作者数据

Conference.csv: 会议数据集
Journal.csv: 期刊数据集
Paper.csv：论文数据集
PaperAuthor.csv: 论文-作者 数据集
paperIdAuthorId_to_name_and_affiliation.json：包含论文-作者对(paperId，AuthorId)到名字-单位(name1##name2; aff1##aff2)的映射关系

4 提交格式

最终提交的文件是对“测试集”的预测结果。该预测结果文件的格式与训练集 **Train.csv** 的格式相同，包含 AuthorId、ComfirmedPaperIds、DeletedPaperIds 字段。该预测结果文件的命名为 Test.P##.csv，其中##为各个小组的编号，如 Test.P01.csv 表示第一个小组提交的测试集预测结果。

5 评估标准

使用在“测试集”上的结果的准确率（Accuracy）作为评估标准。

评估脚本位于 model_trainer 文件夹下，名为 evalution.py，通过运行该脚本可以获得评估结果。

```
python  evalution.py  gold_file_path  pred_file_path
```

其中，gold_file_path 为标准答案所在的路径，pred_file_path 为预测文件所在的路径。

例如，第一小组在验证集合上的预测结果与标准答案的评估：

```
python  evalution.py  valid_set/Valid.gold.csv  valid_set/predict.csv
```