**CCKS 2021 技术评测任务书**

**通用百科知识图谱实体类型推断**

## 任务描述

本评测任务围绕通用百科知识图谱构建中的实体类型推断展开，评测从实体百科（包括百度百科、搜狗百科等来源）页面出发，从给定的数据中推断相关实体的类型。实体类型推断在知识库中具有非常重要的价值，因此该任务也一直是研究的热点。在CCKS 2020新冠知识图谱构建与问答任务中，其第一个子任务就属于实体类型推断任务，该任务主要围绕着医学领域，构建了7个实体类型的推断任务。而这次，我们对实体类型推断任务做了一定的拓展和改进，主要包括以下几个方面：

1. 任务涉及的领域更加广泛和通用，涉及到包括组织机构、人物、作品、位置等多个领域。
2. 实体类型更加丰富，不同领域下包括多个具体的实体类型，整体任务包括几十个实体类型。
3. 实体类型之间具有层级关系，所有实体类型构成了一棵分类树，并且具有subclassof关系。一个具体的实体需要尽可能的推断到最细粒度的实体类型才算正确。各参赛队伍可以考虑是否把这种层级关系加入到模型当中进行训练。
4. 某些实体可能只属于一个实体类型，某些实体可能属于多个实体类型，因此更加增添了任务的难度。比如“刘德华”既属于“歌手”类型，又属于“演员”类型；“知识图谱概念与技术”就只属于“书籍”类型。

### 实体类型

该任务共包括组织机构、人物、作品、位置等几个领域的实体类型。不同领域包括多个具体的实体类型，具体如下：

* 组织机构：政府机构、公司、教育机构、乐队
* 位置：居住地、基础建筑、自然景观、旅游景点、天体
* 人物：文艺工作者、运动员、商人、老师、医生、政治人物、律师、记者、虚拟人物
	+ - 文艺工作者：作家、编剧、歌手、演员、舞蹈者、导演、摄影师、诗人、画家
		- 运动员：足球运动员、篮球运动员、跳水运动员、田径运动员、游泳运动员、体操运动员
* 作品：影视作品、书面作品、软件程序、音乐作品
	+ - 影视作品：电影、电视剧、动漫、电视节目
		- 书面作品：书籍、小说、诗歌、漫画、歌戏剧、文章
		- 音乐作品：歌曲、专辑
		- 软件程序：游戏

对于某个实体，基于百度百科或者搜狗百科提供的页面数据，根据上面的实体类型定义，需要将实体尽可能预测到最细粒度的实体类型。比如：“林丹”应该预测为“运动员”、“武磊”应该预测为“足球运动员”、一名普通的学生应该预测为“人物”。如果实体属于多个实体类型，预测结果需要包含所有的实体类型。比如“特朗普”，应该预测为“商人”和“政治人物”两个类型。

实体类型的层级关系如下图所示：



### 输入输出

* 输入：包括train.csv、valid.csv、test.csv和type.txt

train.csv：百度百科、搜狗百科等来源的实体页面内容。实体页面文件中包含类型推断可能会用到的名称、简介等信息。选手可以通过分析该文件中的内容，采用各种机器学习方法构建模型，进行实体类型推理。train.csv包含的字段信息如下：



valid.csv：valid.csv里面的数据格式与train.csv文件相同，包含了选手需要进行预测及提交结果的实体数据，A榜的排名取决于这份数据的预测结果。该份数据会随train.csv一起在5月中旬发布。

test.csv：test.csv里面的数据格式与train.csv和valid.csv文件相同，包含了选手最终需要提交的实体推理数据，B榜的排名取决于这份数据的预测结果。该份数据将在7月中旬发布。

type.txt：包含所有实体类型，类型之间的层级关系如上所描述。

* 输出：entity\_type.txt

entity\_type.txt：实体id、实体名及预测的实体类型，每行一个“实体id\t实体名\t类型”对。实体id、实体名、类型之间通过分隔符“\t”进行分割。注意不要包含类型不在给定的实体类型当中的实体。若一个实体属于多个类型，则类型之间用英文逗号“,”分隔，表示为“实体id\t实体名\t类型1,类型2”对。类型1、类型2的书写顺序没有影响。

* 样例：

type.txt：组织机构，组织机构>政府机构，组织机构>公司，组织机构>教育机构，组织机构>乐队，位置，位置>居住地，位置>基础建筑，位置>自然景观，位置>旅游景点，位置>天体，人物，人物>文艺工作者，人物>文艺工作者>作家，人物>文艺工作者>编剧，人物>文艺工作者>歌手，人物>文艺工作者>演员，人物>文艺工作者>舞蹈者，人物>文艺工作者>导演，人物>文艺工作者>摄影师，人物>文艺工作者>诗人，人物>文艺工作者>画家，人物>运动员，人物>运动员>足球运动员，人物>运动员>篮球运动员，人物>运动员>跳水运动员，人物>运动员>田径运动员，人物>运动员>游泳运动员，人物>运动员>体操运动员，人物>商人，人物>老师，人物>医生，人物>政治人物，人物>律师，人物>记者，人物>虚拟人物，作品，作品>影视作品，作品>影视作品>电影，作品>影视作品>电视剧，作品>影视作品>动漫，作品>影视作品>电视节目，作品>书面作品，作品>书面作品>书籍，作品>书面作品>小说，作品>书面作品>诗歌，作品>书面作品>漫画，作品>书面作品>歌戏剧，作品>书面作品>文章，作品>音乐作品，作品>音乐作品>歌曲，作品>音乐作品>专辑，作品>软件程序，作品>软件程序>游戏

train.csv表格中一个具体的实体数据例子：



其中，train.csv的contents字段为该实体的完整json描述内容，json解析完如下图所示：



* **输出样例**

 1131993 晶茂电影传媒 组织机构>公司

 22067022 清华大学 组织机构>教育机构

 6072077 细胞的旅程 作品>软件程序>游戏

 2110758 崔志成 人物>商人

 22469665 陈国绪 人物>老师

 20007006 梅尔·吉卜森 人物>文艺工作者>导演,人物>文艺工作者>演员

* **说明**

假设实体“尿TH糖蛋白”在valid.csv中，但由于该实体的类型不是type.txt中所包含的实体类型，因此输出结果中不包含该实体。

### 评价指标

本任务采用精确率（Precision, P）、召回率（Recall, R）、F1值（F1-measure, F1）来评估效果。设$A$为参赛队伍输出文件中的所有实体-类型对的集合，$G$为评测方标注文件中的所有实体-类型对的集合，相关计算公式如下：

$$P=\frac{|A\bigcap\_{}^{}G|}{|A|} R=\frac{|A\bigcap\_{}^{}G|}{|G|} F1=\frac{2PR}{P+R}$$

## 数据描述

我们分别从收集的几千万百科类实体页面中，抽取了部分实体页面作为这次任务的数据来源。任务本身不限定方法类型，可以是无监督、半监督、有监督方法，因此不提供任务相关的有标注结果的训练集，参赛选手如有需要，可以自行对训练数据进行标注。

测试集是通过自动化实体类型推测和人工检验进行标注的。测试集上不允许参赛选手进行任何人工标注。

## 任务提交

本次任务将采取刷榜的方式，将会发布A榜和B榜两个测试集，A榜测试集用于日常刷榜和模型迭代，不计入最终比赛成绩，测试集A发布后，允许参赛队伍每天至多提交一次结果，文件命名为“entity\_type.txt”，格式与任务描述中的示例输出相同。

测试集B为最终评估比赛成绩的有效评测集，待测试集B发布后，允许参赛队伍每天至多提交一次。

**所有获奖团队需额外提交相关材料以供资格审查，要求如下**：

1. 测试集B结果文件，用result.txt命名（UTF-8格式）
2. 相关代码及说明
3. 方法描述文档（非评测论文，评测论文撰写要求见CCKS 2021官网）

将以上三个文件在任务提交截止日期前发送至邮箱lihongyu1@oppo.com。邮件的标题为：“CCKS2021-ETI-参赛队名称”，例如“CCKS2021-ETI-FindX队”。

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用“参赛队名\_code.xxx”命名，要求提交所有的程序代码及相关的配置说明，**程序应当可以运行且所得结果与result.txt相符**。如果方法使用了额外资源，要求说明并提供资源文件或地址。

本次评测将依托Biendata平台，请有意向的参赛队伍关注平台上的比赛列表。

## 评测规则

1. 参赛选手需要提交“参赛队名，队长信息（姓名，邮箱，联系电话），参赛单位名称”等信息，报名方式稍后在评测网站发布。
2. 报名截止到最终测试数据集B发布，在测试数据集发布之后，未报名的选手/队伍不能再报名或提交。
3. 每支队伍需指定一名队长，队伍名称不超过15个字符，队伍成员不超过4人。
4. 每名选手只能参加一支队伍，一旦发现某选手以注册多个账号的方式参加多支队伍，将取消所有相关队伍的参赛资格。
5. 允许使用开源代码或工具，但不允许使用任何未公开发布或需要授权的代码或工具。允许使用外部数据，但该数据必须是公开的，并在提交最终结果时一并提交（如数据过大，需提供下载地址）。
6. 参赛选手最终需要提交可运行的代码和方法描述文档，**若在排行榜上的结果无法复现，将取消参赛资格**。
7. 欢迎国内外在校生及社会在职人士参加。比赛组织方成员不可参赛。
8. 关于比赛的所有解释权归比赛组织方所有。

## 评测奖励

第一名：10000

第二名：8000

第三名：6000

技术创新奖：6000

## 组织者

**任务组织者：**

李洪宇（OPPO）

荚济民（OPPO）

**任务联系人：**

李洪宇：lihongyu1@oppo.com

荚济民：jiajimin@oppo.com

附录：实体类型简要说明表

|  |  |  |  |
| --- | --- | --- | --- |
| 一级标签 | 二级标签 | 三级标签 | 简要说明 |
| 人物 | 文艺工作者 | 作家 |  |
| 编剧 |  |
| 歌手 |  |
| 演员 |  |
| 舞蹈者 |  |
| 导演 |  |
| 摄影师 |  |
| 诗人 |  |
| 画家 |  |
| 运动员 | 足球运动员 |  |
| 篮球运动员 |  |
| 跳水运动员 |  |
| 田径运动员 |  |
| 游泳运动员 |  |
| 体操运动员 |  |
| 商人 |  |  |
| 老师 |  | 学校教师、大学老师、教授、副教授等 |
| 医生 |  |  |
| 政治人物 |  | 历史上真实存在的官员、国家政府重要官员、政治家、政协委员、人大代表等 |
| 律师 |  |  |
| 记者 |  |  |
| 虚拟人物 |  | 游戏、动漫、小说等等非现实中的人物或者角色等 |
| 作品 | 影视作品 | 电影 |  |
| 电视剧 |  |
| 动漫 | 动漫影视 |
| 电视节目 | 电视节目、网络节目、综艺节目等 |
| 书面作品 | 书籍 |  |
| 小说 |  |
| 诗歌 | 诗词、诗歌 |
| 漫画 |  |
| 歌戏剧 | 歌剧、戏剧等 |
| 文章 | 杂志、报纸或其他媒体发表的散文、杂文、评论、报告等等书面作品 |
| 软件程序 | 游戏 |  |
| 音乐作品 | 歌曲 |  |
| 专辑 |  |
| 组织机构 | 政府机构 |  | 政府机关，行政部门、居委会、村委会、各类银行等 |
| 公司 |  |  |
| 教育机构 |  | 各类学校、高校、研究机构、教育相关的机构组织等 |
| 乐队 |  | 乐队、歌手组合等 |
| 位置 | 居住地 |  | 包括国家、大陆、岛屿、省份、城市、县、镇、乡、区、村等人口居住的地方等 |
| 基础建筑 |  | 包括车站、地铁站、机场、电影院、超市、酒店、饭店、一般展馆、医院、一般公园、一般寺院、小区楼房、大楼大厦等 |
| 自然景观 |  | 包括各种森林、沙漠、火山、河流、沙滩、山川、湖泊、峡谷、地质公园、森林公园、自然保护区等 |
| 旅游景点 |  | 包括有游客游玩旅游参观的地方等 |
| 天体 |  | 包括宇宙中星系、行星、彗星、小行星、卫星、恒星、星云等等 |