**CCKS 2021 技术评测任务书**

**网页文件中学者画像任务**

# 背景与任务描述

随着大数据和人工智能的发展，刻画用户的基本画像、行为画像、心理画像等已广泛应用于搜索、推荐系统等。高端人才画像旨在对专家学者的精准事实知识进行抽取、标签化和统计分析，进而用于学术搜索、科研服务、人才挖掘等方面，推动学术人工智能和科技情报分析系统的发展。

本次大赛主要挖掘专家的基本画像，但现有的挖掘方法受限于标注数据集或者基于规则统计的方法。希望选手能够结合当下最先进的AI和NLP技术，深入挖掘web文本中的专家标签信息，构建模型实现精准的高端人才画像。

**比赛任务：**

本次比赛将为每个学者提供一组搜索引擎返回的网页，其中包括学者本人相关的网页、同名者的网页和其他不相关的网页。选手从这组网页中提取关于这名学者的结构化的信息，如学者主页（不同网站）、性别、所在国家等。

此次任务的输入输出定义如下。

**输入：**

学者姓名和学者机构名，以及通过这两个关键词在谷歌和百度检索返回的20个页面（谷歌和百度各10个）。

**输出：**

输出学者的关键画像信息，包括人主页域名、性别、使用语言、Google Scholar主页域名、职称、所在国家。

# 数据集介绍

本次评测数据集来自国内领先的科技情报大数据挖掘与服务系统平台AMiner。数据集中共包括1万名学者的数据，其中6000人用于训练集，2000人用于验证集，2000人用于测试集。每个人的数据都包括：1）姓名和机构名，2）用姓名和机构名在百度和谷歌搜索后返回的20个页面，3）学者画像信息，包括个人主页域名、性别、使用语言、Google Scholar主页域名、职称、所在国家。

**示例输入：**

**{“1”, “Geoffrey Hinton”, “University of Toronto”, 1.html, 2.html,…,20.html}。**

**其中，每个html文件的格式如下：**

<html>

<body>

<td valign="TOP" class="main">

<h2> Geoffrey E. Hinton</h2>

</td>

…

<td>

<b>

<a href="http://www.toronto.edu/">University of Toronto</a>

</b>

</td>

…

…

</body>

</html>

**示例输出：**

{

"scholar\_id":"1",

"personal\_url":" <https://www.cs.toronto.edu/~hinton/>",

"Google\_url":" https://scholar.google.co.uk/citations?user=JicYPdAAAAAJ&hl=en ",

"position":[

{

"org":" University of Toronto ",

"position":" University Professor Emeritu",

},

{

"position":" Chief Scientific Advisor ",

"org":"Vector Institute",

},

,

"gender":"male",

"location":"Canada",

"language":"English",

]

}

# 评价指标

每个学者共有k（在该项任务中 k = 6）项个人画像数据需要预测。这 k 个中除了职称/职位外都通过完全匹配的方式进行评测，完全匹配上得1分，否则得0分；而职称/职位由于是一个集合，所以通过Jaccard index （即两个集合交集的元素个数除以两个集合并集的元素个数）来进行计算提取出来的集合跟标注答案给出的集合的相似度 （介于0~1之间）。一个学者的画像数据的预测值的最终得分为这 k 个预测分数的平均分。一个参赛选手在该任务上的最终得分为对各个学者的画像数据的预测得分的平均分。即：

# 任务提交方式

本次任务将采取刷榜的方式，将会发布验证集和测试集。验证集用于日常刷榜和模型迭代，不计入最终比赛成绩，验证集发布后，允许参赛队伍每天至多提交3次结果，文件命名为“\*.json”，格式与任务描述中的示例输出相同，排名在提交后实时更新。

测试集为最终评估比赛成绩的有效评测集，待测试集发布后，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. 测试集结果文件，用\*.json命名（UTF-8格式）；
2. 训练代码及说明，训练好的、可以复现最优结果的模型；
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）

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），作为邮件附件传送，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\*.json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。

# 奖励设置（待定）

总奖金2.5万元

1、第一名（1支队伍）：1万元

2、第二名（2支队伍）：各2000元

3、第三名（4支队伍）：各1500元

技术创新奖：5000 元

# 时间安排

比赛页面上线：4月20日

报名时间：4月20日-7月10日 23:59

样例数据发布：4月20日

全部训练数据和不带标签的验证集数据发布：5月10日

测试集数据发布：7月15日

测试集截止提交：7月18日 23:59

提交最终测试结果：7月21日

评测排名及结果复审：7月25日

评测论文提交截止日期：8月5日

CCKS会议日期(评测报告及颁奖)：8月18日-21日

# 组织者信息

王路路 llwangxju@163.com

姜鹏 [peng.jiang@aminer.cn](mailto:peng.jiang@aminer.cn)

刘佳 jia.liu@aminer.cn