**任务： 公众公司公告信息抽取**

# 任务描述

随着金融科技的发展和全球资本市场的不断扩大，在金融领域， 每一天都有海量的数据产生， 而与之形成强烈对比的是有限的人力以及人脑所能处理信息的极限能力。因此，依靠传统的人工方式已经无法应对投研分析、风险控制、金融监管和事件关联等需求，而亟需引入新的技术来提高信息处理效率，包括大数据分析、自然语言处理、知识图谱等技术，都已经开始被积极用于金融分析和金融监管领域。在监管方面，每一家公众公司都具有相关信息披露义务，由此也产生了大量的公告阅读和信息抽取需求。 据不完全统计，以沪深股市为例，2017年共披露公告44万余篇，2018年共27万余篇，并且随着上市公司数量的增加这一数字也在逐年增加。每年3月底、4月底、8月底、10月底为定期报告披露高峰期，最多的一天所发布公告达 10297 篇。

本次评测的主要目标是针对公告文件（均以PDF方式发布）中的信息抽取。作为知识图谱构建的基础，结构化数据是必不可少的。由此，如何通过自动化的技术来从各类公告中抽取信息，将非结构化数据转化为结构化数据是知识图谱领域所面临的一大挑战。

此次评测将包括两个任务：公众公司定期报告表格中的信息点提取和文本段落中的信息点提取。 具体内容如下：  
**任务一：表格中的信息点提取。**公众公司定期报告中财务报表信息点提取（包括合并资产负债表，母公司资产负债表，合并利润表，母公司利润表，合并现金流量表和母公司现金流量表），除表头部分外，对其中每一行采取原样提取原则。  
  
**输入：**公众公司年报pdf文件  
**输出：**该表格所对应的结构化数据（JSON格式）  
**示例：**  
**输入：**以信宜股份（证券代码：872176）2018年年度报报告为例（ <http://static.cninfo.com.cn/finalpage/2019-02-01/1205830955.PDF> ），其中定位至：第十一节 财务报告 → 二、 财务报表 → (二) 母公司资产负债表   
**输出：**

{

"872176-信宜股份-2018年年度报告": {

"证券代码": "872176",

"证券简称": "信谊股份",

"资产负债表（母公司）": {

"单位": "元",

"项目": [

{

"名称": "流动资产：",

"附注": "",

"年初至报告期末": null,

"上年年初至报告期末": null

},

**…… ,**

{

"名称": "负债和所有者权益合计",

"附注": "－",

"年初至报告期末": "47,333,600.42",

"上年年初至报告期末": "37,153,800.58"

}

**…… ,**

]

}

}

}  
**注：**1)其中"872176-信宜股份-2018年年度报告"为给定PDF文件名；2)给定输出示例有省略，仅包含表格中部分内容，实际处理需尽可能提取完整的报表信息；3）同一份报告中的六张财务报表须在同一JSON对象中返回，具体格式请参见给定训练数据；4）给定输出示例对各表表头进行了语义转换（如给定示例中，"项目"🡪"名称"，"期末余额"🡪"年初至报告期末", "期初余额"🡪"上年年初至报告期末"），具体完整格式说明参见给定训练数据。   
  
**任务二：文本段落中的信息点提取**  
此次提取公告为“人事变动”类型公告，参赛者需从中提取出离职高管信息及继任者信息。  
**输入：**“人事变动”类型公告pdf文件  
**输出：**包含离职高管信息和继任者信息的结构化数据（JSON格式）  
**示例**：  
**输入：**以闽灿坤B（证券代码：200512）关于董事辞职的公告为例（<http://static.cninfo.com.cn/finalpage/2018-08-16/1205289300.PDF>）  
**输出：**

{

"200512-闽灿坤B-人事变动": {

"证券代码": "200512",

"证券简称": "闽灿坤B",

"人事变动": [

{

"离职高管姓名": "陈彦君",

"离职高管性别": "先生",

"离职高管职务": "第九届董事会董事、董事会战略委员会委员",

"离职原因": "个人原因",

"继任者姓名": null,

"继任者性别": null,

"继任者职务": null

},

{

"离职高管姓名": "庄兴",

"离职高管性别": "先生",

"离职高管职务": "第九届董事会董事",

"离职原因": "个人原因",

"继任者姓名": null,

"继任者性别": null,

"继任者职务": null

}

]

}

}

注：其中"200512-闽灿坤B-人事变动"为给定PDF文件名。人事变动可能有多个对象。若无继任者，则继任者信息为空。若高管为增补高管，则离职高管信息为空.

# 数据描述

本次测评训练数据为公开发布的公众公司定期报告文件。组织者将提供训练数据集和验证集（包括公告PDF原文和对应的结构化数据），供参赛选手训练算法模型和参与验证排名。评测数据为训练数据集和验证集同一类型公告，为防止作弊和人工介入，评测数据将包含非公开发布的年报和公告PDF文件（人为制造），同时，评测数据不提前进行发布。参赛者须提供可调用的API，组织者将基于评测数据在指定评测时间统一调用API来给出最终分数。

# 评价指标

本次任务的评价采用正确率、召回率和F1值作为评价指标，内容如下：

**正确率** = 提取出的正确信息点数 / 提取出的总信息点数

**召回率** = 提取出的正确信息点数 / 样本中的总信息点数

**F1值** = 2 \* 正确率 \* 召回率 / (正确率 + 召回率)

注：（评测过程中，将对参赛者的API调用处理时间和结果返回时间设置上限，若一条评测数据未能在限定时间内返回结果，该评测数据及对应的信息点将按抽取失败计入评价指标计算中，具体限定时间及最终评价指标将同训练数据一同发布）。

# 任务提交

本次任务在训练阶段将采取刷榜的方式，验证集发布后，允许参赛队伍多次向平台提交结果，文件命名为“参赛队名称\_valid\_result.txt”，格式与结果文件result.txt相同，排名随时更新。参赛队伍可在评测集发布之前随时上传验证集的计算结果，管理系统会及时更新各队伍的最新排名情况。但最终排名由测试集评测决定。

**提交文件说明：**  
每一个参赛队需提交的材料如下：

* 方法描述文档
* 可调用的Web API 及说明（具体的API格式规定将同训练数据一同发布）

本次评测将依托biendata平台（https://biendata.com/）展开，请有意向的参赛队伍关注平台上的竞赛列表。

# 时间安排

* 评测任务发布：4月1日
* 报名时间：4月1日—4月20日
* 训练及验证数据发布：4月20日
* 评测文件提交：7月20日
* 评测时间：7月25日
* 评测论文提交：8月15日

# 规则

* 所有参赛选手都必须在管理系统中注册；
* 参赛选手在管理系统中组队，每支队伍最多不超过5名队员，报名截止日期之后不允许更改队员名单，只允许进行队伍合并操作；
* 每支队伍需指定一名队长，队伍名称不超过15个字符；
* 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，将取消相关队伍的参赛资格；

组织者： 东南大学 认知智能研究所 漆桂林

东南大学 认知智能研究所 王萌