**任务：面向金融领域的事件主体抽取**

# 任务描述

“事件识别”是舆情监控领域和金融领域的重要任务之一，“事件”在金融领域是投资分析，资产管理的重要决策参考。“事件识别”的复杂性在于事件类型和事件主体的判断，比如“公司A产品出现添加剂，其下属子公司B和公司C遭到了调查”，对于“产品出现问题”事件类型，该句中事件主体是“公司A”，而不是“公司B”或“公司C”。我们称发生特定事件类型的主体成为事件主体，本任务中事件主体范围限定为：公司和机构。事件类型范围确定为：产品出现问题、高管减持、违法违规…

本次评测任务的主要目标是从真实的新闻语料中，抽取特定事件类型的主体。即给定一段文本T，和文本所属的事件类型S，从文本T中抽取指定事件类型S的事件主体。

**输入**：一段文本，事件类型S

**输出**：事件主体

**示例**：

*样例1*

*输入：”公司A产品出现添加剂，其下属子公司B和公司C遭到了调查”， “产品出现问题”*

*输出：* “公司A”

*样例2*

*输入：“公司A高管涉嫌违规减持”，“高管减持”*

*输出： “公司A”*

# 数据描述

本次数据主要来自金融领域的真实新闻文本，训练集、验证集及测试集的说明如下：

**训练集&验证集：**

在训练及验证数据发布阶段，我们会发布2w条左右的文本及其所标注事件类型和事件主体和5千条左右的验证文本及其所标注的事件类型和事件主体。训练集每行4列，数据以“\t”分隔，格式为：*文本id\t文本内容\t事件类型\t事件主体，*验证集集每行3列，数据以“\t”分隔，格式为：*文本id\t文本内容\t事件类型。*

**测试集：**

在测试数据发布阶段，我们将会再发布5千条左右的文本数据集，不含标注结果，作为测试。每行3列，数据以“\t”分隔，格式为：*文本id\t文本内容\t事件类型*

# 评价指标

本次任务采用精确率（Precision, P）、召回率（Recall, R）、F1值（F1-measure, F1）来评估事件主体的识别效果。

$$事件主体精确率=识别事件主体与标注相同/识别事件主体总数量$$

$$事件主体召回率=识别事件主体与标注相同/标注事件主体总数量$$

$$事件主体F1值=\frac{事件主体精确率\*事件主体召回率}{事件主体精确率+事件主体召回率}$$

# 任务提交

本次任务将采取刷榜的方式，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（每日最多可上传3次），管理系统会及时更新各队伍的最新排名情况；

测试集发布后，允许参赛队伍多次提交结果文件，结果文件提交格式如下。

**结果文件格式说明：**

选手将结果保存为result.txt，以utf-8编码格式保存，每行2列，以“\t”分隔，第一列为文本id，第二列为事件主体，即“*文本idt事件主体*”。

最终提交文件要求：

每一个参赛队需提交的材料如下：

1. 事件主体运行结果文件
2. 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 2019官网）

以上三个文件需在任务提交截止日期前发送至邮箱：ccks2019task4@126.com。邮件的标题为：“CCKS-评测任务四-最终提交文件-参赛队名称”，例如“CCKS-评测任务四-最终提交文件-火箭队”。

结果文件用result.txt命名，采用utf-8格式存储，文件格式需要与训练数据中的标注结果文件完全一样(每行格式为“*文本id\t事件主体*”)。

代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），用code.xxx命名，要求提交所有的程序代码及相关的配置说明，确保程序能够正确运行，且所得结果与result.txt相符。

方法描述文档用SM.pdf命名，包含算法描述及参数设置，需用pdf格式存储，页数不超过5页。

评测平台：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

• 任务征集截止：3月20日

• 任务准备时间：3月20日—4月1日

• 评测任务发布：4月1日

• 报名时间：4月1日—7月20日

• 训练及验证数据发布：4月30日

•   验证集提交开放：5月1日23:59 - 7月27日23:59 （北京时间）

•     测试数据发布：7月27日23:59（北京时间）

•     测试集提交开放： 7月27日23:59 - 7月30日23:59（北京时间）

•     评测论文提交：8月15日

# •     CCKS会议日期：8月24日—27日规则

1. 所有参赛选手都必须在biendata平台管理系统中注册，本次比赛的参赛对象仅限全日制在校大学生（本科、硕士、博士均可）和企业员工；

2. 参赛选手需确保注册时提交信息准确有效，所有的比赛资格及奖金支付均以提交信息为准；

3. 参赛选手在管理系统中组队，参赛队伍成员数量不得超过5个，报名截止日期之后不允许更改队员名单；

4. 每支队伍需指定一名队长，队伍名称不超过15个字符，队伍名的设定不得违反中国法律法规或公序良俗词汇，否则组织者有可能会解散队伍；

5. 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；

6. 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；

7. 除主办方提供的数据集外，参赛选手允许使用预训练数据（如词向量、字向量等），但是不能用额外的标注数据。

8. 参赛队伍可在参赛期间随时上传验证集的预测结果，一天不能超过3次 ，管理系统会实时更新各队伍的最新排名情况。

# 任务交流平台

任务交流平台采用钉钉群的方式。所有报名参赛的参赛队要求有一名成员加入到该讨论群中，后续所有有关本任务评测及数据的相关说明和通知将只在讨论群中发布和交流，不再另行发邮件或更新官网通知，请所有参赛队务必在报名成功之后加入该讨论群。钉钉群加入二维码如下，在钉钉上扫一扫添加：



加入讨论组时请务必注明所属参赛队队名及所属单位，加入讨论组有问题的参赛队请联系：**ccks2019task4@126.com**。