

2017 年第二届全国网络舆情分析技术邀请赛

评测大纲

事件样本发现

一、任务说明

本任务的目标是从给定语料中查找与指定事件对应的新闻语料。任务的输入为 N 个专题事件及每个事件对应的若干篇种子文档, 要求从给定新闻语料中找出与这 N 个事件对应的全部新闻文档。

二、评测数据集

主办方提供: 训练语料集、测试语料集。

1. 训练语料集

训练语料集共包含 6000 篇新闻报道, 新闻报道均来自互联网, 涵盖国际、国内、军事、财经、社会五类。新闻内容未经过分词处理。

2. 测试语料集

测试语料集包含 4000 篇新闻报道, 新闻报道均来自互联网, 涵盖国际、国内、军事、财经、社会五类。新闻内容未经过分词处理。

三、参评系统输入输出文件格式

1. 输入文件格式

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<Samples>
  <Sample>
    <EventID>123123</EventID><!--事件编号-->
    <EventTitle>一带一路国际合作高峰论坛</EventTitle><!--事件名称-->
    <RelSampleID><!--与事件相关的3篇报道-->
      <SampleID>1231</SampleID><!--样本编号-->
      <SampleID>899</SampleID>
      <SampleID>89798</SampleID>
    </RelSampleID>
  </Sample>
  <Sample><!--以下含若干组样例，同上说明--> ...
</Samples>
```

图一：测试事件样例文档

```
<?xml version='1.0' encoding='UTF-8'?>
<Samples>
  <Sample>
    <SampleID>1743</SampleID><!--样本编号-->
    <SampleTitle>全会审议通过《监督执纪工作规则》</SampleTitle><!--样本标题-->
    <publishDate>2017-01-09</publishDate><!--样本发布日期-->
    <SampleContent>《中国共产党纪律检查机关监督执纪工作规则（试行）》<!--样本内容-->
    十八届中央纪委七次全会8日在京闭幕，全会审议通过了《中国共产党纪律检查机关监督执纪工作规则（试行）》。
    全会一致认为，制定监督执纪工作规则，是纪检机关贯彻党的十八届六中全会精神，落实全面从严治党部署，带头强化自我约束，把监督执纪权力关进制度笼子，
    做到正人先正己的实际行动，充分表明了严格自律的担当和决心。
    全会审议通过的《中国共产党纪律检查机关监督执纪工作规则（试行）》，紧扣监督执纪工作流程，明确请示报告、线索处置、初步核实、立案审查、案件审理、
    涉案款物管理等工作规程；规定谈话函询的工作程序，执纪审查的审批权限，调查谈话和证据收集的具体要求；提炼有效管用实招，上升为制度规范，
    把纪委的自我监督同接受党内监督、社会监督等有机结合，确保党和人民赋予的权力不被滥用。
    全会指出，中央纪委要带头执行规则，加强对监督执纪工作的领导、管理和监督，各级纪委要切实履行自身建设主体责任，严肃处理执纪违纪、失职失责行为。
  </SampleContent>
  </Sample>
  <Sample><!--以下含若干新闻样本，同上说明--> ...
</Samples>
```

图二：测试语料文档

2. 输出文件格式

每组输出结果为一个扩展名为 txt 的文本文件，结果文件名称由参赛者自定，只需在 config 配置文件（**config 配置文件请参见 stokis for event sample find 使用说明文档**）中注明。结果文件格式说明见下：

```
EventID:209
SampleID:1123,1822,33842
EventID:547
SampleID:8989,2222,101
EventID:741
SampleID:12323,22,13113

/*说明：
(1) 每组数据分为两行输出；
(2) 第一行为EventID，对应测试集文件中每个事件样本的EventID；
(3) 第二行为SampleID，对应程序从新闻语料集中找到的相关事件样本的SampleID。
(4) 提交此文件时将说明部分删除。
*/
```

图三：输出文件格式说明

四、评价指标

事件样本发现评价采用准确率、召回率、F 值作为评价指标。计算公式如下：

$$P = \frac{\sum_{i \in C} a_i}{\sum_{i \in C} b_i}$$

$$R = \frac{\sum_{i \in C} a_i}{\sum_{i \in C} c_i}$$

$$F = \frac{2 * P * R}{P + R}$$

其中 C 是专题事件的集合， a_i 是关于事件 i 的预测结果中正确的新闻数量， b_i 是关于事件 i 的预测结果中全部的新闻数量， c_i 是测试语料中关于事件 i 的新闻数量。

最终排名以综合评分的 F 值作为依据。

五、测试步骤

1、在训练阶段，参赛方远程登录客户机，进行系统部署与环境搭建，并获取训练数据进行调试训练。参赛方可调用大赛提供的代理程序 **stokis** 对模型进行测试验证，测试数据由参赛方从训练数据中自行选取，代理程序可以反馈测试结果。调用代理程序需要参赛方提供可执行程序并填写代理程序的 **config** 文件。其中，使用 **windows** 虚拟机的队伍训练程序和测试程序分别提供 **xxx.bat** 的执行文件，使用 **linux** 虚拟机的队伍训练程序和测试程序分别提供 **xxx.sh** 的执行文件。具体调用步骤请参见《**stokis for text analysis** 使用说明文档》。

2、评测阶段开始前，参赛方需在 **config** 文件中配置最终版程序的路径，由大赛平台自动进行评测，并将评测结果反馈给参赛队伍。

说明：

1、代理程序的功能。代理程序会自动完成以下操作：1) 切断客户机的访问连接，2) 下载测试数据，3) 运行测试任务产生结果文件，4) 上传结果文件到服务器，5) 提交评测服务获得评测结果，6) 将整个运行结果存入数据库，7) 恢复客户机访问连接。

2、技术指标评分。代理程序自动运行文本分类评测工具，给出相关性能指标，测试结果形式如下：

新闻评测结果，MRR_____；

六、 注意事项

代理程序开始运行后，参赛方与客户机断开连接，无法通过调试程序来优化结果。代理程序运行完毕后，参赛方可重新登录客户机，通过调试程序来优化结果。如果代理程序出现问题，请与主办方工作人员及时沟通解决。