2017 年第二届全国网络舆情分析技术邀请赛

评测大纲

文本分类

一、 任务说明

给定一组未经分词的文本文档，对每一篇文档指定唯一类别。按文本类型分为两个子任务：新闻报道分类和短文本分类。新闻报道 10 万篇，包含 15 个类别；短文本 10 万条，包含 15 个类别。数据比例严重倾斜且动态变化。参赛者需要基于训练数据实现两个分类系统，实现两类文档的自动分类。

二、 评测数据集

主办方提供：新闻调试集、新闻调试标注集、新闻训练集、新闻训练标注集、新闻测试集，短文本调试集、短文本调试标注集、短文本训练集、短文本训练标注集、短文本测试集。

1. 新闻调试集

新闻调试集包含 150 篇新闻报道，共 15 个类别，每个类别包含 10 篇新闻，用于调试参赛系统。

2. 新闻调试标注集

新闻调试标注集包含新闻调试集中每篇新闻的对应类别，共 15 个类别。

3. 新闻训练集

新闻训练集包含 70,000 篇新闻报道，文本未经过分词处理。

4. 新闻训练标注集

新闻训练标注集包含新闻训练集中每篇新闻的对应类别，共 15 个类别。

5. 新闻测试集

新闻测试集共包含 30,000 篇新闻报道，文本未经过分词处理。新闻测试集分为 3 份，每一份测试数据的评测分值占最终评测分值的 10%、30%、60%。

6. 短文本调试集

短文本调试集包含 150 条短文本，共 15 个类别，每个类别包含 10 条短文本，用于调试参赛系统。

7. 短文本调试标注集

短文本调试标注集包含短文本调试集中每条短文本的对应类别，共 15 个类别。

8. 短文本训练集

短文本训练集包含 70,000 条短文本，文本未经过分词处理。

9. 短文本训练标注集

短文本训练标注集包含短文本训练集中每条短文本的对应类别，共 15 个类别。

10. 短文本测试集

短文本测试集共包含 30,000 条短文本，文本未经过分词处理。短文本测试集分为 3 份，每一份测试数据的评测分值占最终评测分值的 10%、30%、60%。

三、 参评系统输入输出文件格式

1. 输入文件格式

```
<Text>
    <ID>1</ID>
    <Title>深度：萨德抵韩已无情面好讲 中国可出10招打痛韩美</Title>
    <Content>美韩军方3月7日宣布，在韩部署"萨德"的第一批装备于3月6日晚抵韩，
             "萨德"部署进程正式启动，韩国未就该情况向中国做任何通报。
             对此，中国外交部发言人耿爽3月7日回应说，我们坚决反对美韩在韩国部署"萨德"反导系统，
             将坚决采取必要措施维护自身的安全利益。由此产生的一些后果由美韩来承担。
             再次强烈敦促有关方面停止部署进程，不要在错误的道路上越走越远。
    </Content>
</Text>

<Class>
    <ID>1</ID>
    <classname>军事</classname>
</Class>
```

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 text analysis 使用说明文档》**）中注明。结果文件中的每一行对应一篇新闻（或一条短文本）的分类结果，用类别 ID 表示，数值从 1 到 15。

四、 评价指标

文本分类评价采用准确率、召回率以及 F 值作为评价指标。评分综合每个类别的评价结果，每个类别的权重与该类别的样本数量成反比。新闻报道和短文本分类的准确率、召回率和 F 值的计算公式如下：

$$\mathrm{P} = \frac{\sum_{i \in C}(1 - C_i) * P_i}{|C|}$$

$$\mathrm{R} = \frac{\sum_{i \in C}(1 - C_i) * R_i}{|C|}$$

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

其中 C 是类别的集合，Ci 是属于类别 i 的样本数量与样本总数的比值，P_i，R_i分别是类别 i 的准确率、召回率。设类别 i 的分类结果中，正确分为该类的样本数目是 a，错误划归为该类的样本数目是 b，将该类错误划归为它类的样本数目是 c。类别 i 的准确率、召回率的计算公式如下：

$$\mathrm{P_i} = \frac{a}{\mathrm{a} + \mathrm{b}}$$

$$\mathrm{R_i} = \frac{\mathrm{a}}{\mathrm{a} + \mathrm{c}}$$

最终排名以综合评分的 F 值作为依据，新闻报道和短文本分值各占 50%。

五、 测试步骤

1、在调试阶段，参赛方远程登录客户机，进行系统部署与环境搭建，并获取调试数据进行调试。

2、在系统测评阶段，参赛方提供可执行程序。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。参赛方需要填写代理程序的 config 文件，利用代理程序 stokis 提交评测任务。比赛数据和测试步骤请参见**《stokis for text analysis 使用说明文档》**。

3、参赛方允许在规定时间范围内完成分类任务，包括 3 组新闻测试数据集和 3 组短文本测试数据集。每组测试集可以提交任意次测试结果，但只有第一次结果的评测成绩作为最终排名依据。如果参赛方没有提交一组测试集的运行结果，则该组测试集的得分为 0。

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）下载训练数据，3）运行训练任务产生模型文件，4）下载测试数据，5）运行测试任务产生结果文件，6）上传结果文件到服务器，7）提交评测服务获得评测结果，8）将整个运行结果存入数据库，9）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：

新闻第 1 组结果，准确率___，召回率___,F 值___；

新闻第 2 组结果，准确率___，召回率___,F 值___；

新闻第 3 组结果，准确率___，召回率___,F 值___；

短文本第 1 组结果，准确率___，召回率___,F 值___；

短文本第 2 组结果，准确率___，召回率___,F 值___；

短文本第 3 组结果，准确率___，召回率___,F 值___；

综合评测结果，准确率___，召回率___,F 值___（排名由综合 F 值决定）。

六、 注意事项

1、参赛方需要在指定时间内完成全部 6 组测试集的运行测试。每组测试集可以提交多次测试任务，但只有第 1 次提交的测试任务的评测成绩作为排名依据。如果某一组测试集没有对应的输出结果，则该组测试集的评测成绩为 0。该项目最终得分为 6 组评测成绩的加权平均。

2、代理程序开始运行后，如果参赛程序中断，则视为一次失败的提交。参评程序执行过程中，应记录程序中断提示信息。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。

3、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。