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、 数据集

1. 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。

2. 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。

3. 新闻测试集

新闻测试集共包含 30,000 篇新闻报道供测评阶段使用，文本未经过分词处理。

4. 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。

5. 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。

6. 短文本测试集

短文本测试集共包含 30,000 条短文本供测评阶段使用，文本未经过分词处理。

三、 参评系统输入输出文件格式

1. 输入文件格式

训练数据及测试数据示例

```
<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>
```

训练标注示例

```
<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。

```
3
5
2
11
7
```

输出文件样式

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

新闻结果，准确率___，召回率___,F 值___；

短文本结果，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

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。