2017 年第二届全国网络舆情分析技术邀请赛

评测大纲

用户画像

一、 任务说明

给定一批社交网络用户数据，包括用户个人信息、用户行为信息、用户社交网络文本以及用户粉丝列表等，预测社交网络用户的四类属性标签。

标签 1：用户的年龄（共 3 个类别：-1979/1980-1989/1990+，单类别）

标签 2：用户的性别（共 2 个类别：男/女，单类别）

标签 3：用户的地域（共 8 个类别：东北/华北/华中/华东/西北/西南/华南/境外，单类别）

标签 4：用户的兴趣（共 13 个类别：体育/健康/军事/女性/娱乐/教育/旅游/汽车/社会/科技/航空/读书/财经，多类别）

二、 数据集

主办方提供：社交网络训练集，社交网络训练标注集，社交网络测试集。

1. 社交网络训练集

社交网络训练集包含 6000 条社交网络用户信息，以及每一个社交网络用户的关注用户和粉丝用户数据。

2. 社交网络训练标注集

社交网络训练标注集为社交网络训练集中每一个社交网络用户的属性标签。

3. 社交网络测试集

社交网络测试集共包含 6000 条的待分类用户信息，以及每一个待分类用户的关注用户和粉丝用户的用户数据。

三、 参评系统输入输出文件格式

1. 输入文件格式

测试集包含两类数据：社交网络用户信息数据 users 以及部分的相关（关注与粉丝）用户数据 relations。其中社交网络用户信息数据 users 为所分类的主用户。

其中单条用户数据 user 格式如下：

id	用户 id
name	社交网络用户名
verified	表示是否为认证用户，0 为未认证，1 为认证用户
fan_list	社交网络用户的粉丝列表
follow_list	社交网络用户的关注列表
contents	社交网络用户所发的社交网络内容
retweets	社交网络用户的转发内容
supports	社交网络用户的点赞内容

其中 fan_list，follow_list，contents，retweets，supports 为列表，包含多个子节点。具体文件格式范例如下：

```
<user>
    <id>65435234</id>
    <verified>0</userified>
    <name>932f05493246552er</name>
    <content>
        <childnode>当警钟为他人响起，你不要问警钟为谁而鸣，警钟是为你而鸣。
        </childnode>
    </content>
</user>
```

其中单条相关用户数据 relation_info 格式如下：

id	用户 id

name	社交网络用户名
verified	表示是否为认证用户，0 为未认证，1 为认证用户
location	用户的位置信息
sex	用户性别。由于数据标记问题，男性用户可能取值有”男”或者”m”；女性用户可能取值有”女”或者”f”
birth	社交网络用户的生日信息
fan_list	社交网络用户的粉丝列表
follow_list	社交网络用户的关注列表
contents	社交网络用户所发的社交网络内容
retweets	社交网络用户的转发内容
supports	社交网络用户的点赞内容

其中 fan_list，follow_list，contents，retweets，supports 为列表，包含多个子节点。具体文件格式范例如下：

```
<relation_info>
    <id>353545</id>
    <info>个人</info>
    <name>3543634652531</name>
    <location>山西 运城</location>
    <fan_list>
        <childnode>643636</childnode>
        <childnode>654656</childnode>
        <childnode>775443</childnode>
        <childnode>356432</childnode>
        <childnode>878654</childnode>
        <childnode>342565</childnode>
        <childnode>664322</childnode>
    </fan_list>
    <sex>f</sex>
</relation_info>
```

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，输出数据的结果文件参赛选手可以自行命名，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，例如 yhhx_predict_label_1.txt。结果文件中的每一行对应数

据集中单条用户数据 user 对应的类别。其中包含 17 个数字，用英文逗号字符’,’隔开。数字分别为：

用户 id，年龄类别，性别，区域，以及 13 个用户兴趣标识（顺序为：体育/健康/军事/女性/娱乐/教育/旅游/汽车/社会/科技/航空/读书/财经，某标识位为 0 时，表示对应的用户兴趣为否定；反之则为肯定）。输出文件格式范例如下：

```
110001,3,1,3,0,1,0,0 ,0,0,0,0 ,0,0,0,0, 0,0,0,0,0
35546,1,1,3,0,1,0,0 ,0,0,0,0 ,0,0,0,0, 0,0,0,0, 0
117075,3,1,0,0,1,0,0 ,0,0,0,0 ,0,0,0,0, 0,0,0,0, 0
70386,2,1,4,0,1,0,0 ,0,0,0,0 ,0,0,0,0, 0,0,0,0, 0
81402,3,0,3,0,1,0,0 ,0,0,0,0 ,0,0,0,0, 0,0,0,0, 0
28991,2,1,0,0,1,0,0 ,0,0,0,0 ,0,0,0,0, 0,0,0,0, 0
```

四、 评价指标

用户画像根据 F 值以及相应的标签权重对比赛结果进行评分。评分综合四类标签的评价结果，每个类别的权重与该类别的标签数量成正比。按照得分 S 从大到小进行排序，得到各参赛队排名。

$$S = \sum_{i=1}^{4} \frac{N_i}{N} * F_i$$

其中 N 是全部四类标签的数量，N_i 是第 i 类标签的标签数量， F_i是第 i 类标签的预测得分，其中 F 值的计算公式如下：

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

式中准确率为 P、召回率为 R，准确率、召回率的计算公式如下：

1）$P = \frac{a}{a+b}$，$R = \frac{a}{a+c}$，适用于标签 1，2，3 的评测，其中 a 为正确预测为该标签的样本数目，b 为错误预测为该标签的样本数目，c 为将该标签错误划归为其他标签的样本数目。

2） $\mathrm{P}=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|A_i\cap B_i|}{|B_i|}$，$\mathrm{R}=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|A_i\cap B_i|}{|A_i|}$，适用于标签 4 的评测，其中 A_i 为样本 i 的正确标签集合，B_i 为样本 i 的预测标签集合，$|D|$ 是测试集中的样本个数。

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。代理程序自动运行用户画像评测工具，给出相关性能指标，测试结果形式如下：

评测结果，S___；

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。