

# 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+）

标签 2：用户的性别（共 2 个类别：男/女）

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）

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）

#### 二、评测数据集

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 6k 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

社交网络测试集相应地分为 3 份，分别包含 1k,2k,3k 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
contens	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, contens, 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</verified>
  <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
contens	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, contens, 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测

试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

用户画像评价根据标签预测正确率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。

$$P = \sum_{i=1}^4 \frac{N_i}{N} * P_i$$

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， $P_i$  是第  $i$  类标签的预测准确率，计算如下：

$$P_i = \frac{n_i}{n}$$

其中， $n_i$  是预测正确的账号数量， $n$  是全部账号数量。

最终排名按照准确率从大到小进行排序。

#### 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内完成 3 组用户画像任务。每组测试集可以提交 10 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。该项目的最终得分为 3 组评测成绩的加权平均。

#### 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。

## 六、 注意事项

1、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。

2、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。