

2017 年第二届全国网络舆情分析技术邀请赛

评测大纲

社交关系预测

一、任务说明

本任务的目标是根据社交网络用户的基本信息、发布消息内容、回复消息内容、关注主题、用户间的社交关系（包括回帖、点赞、评论等），预测用户之间可能建立的关注关系。比赛不允许使用外部数据资源。

二、评测数据集

主办方提供以下数据：用户基本信息集，热门发帖集，热门回帖集，主题标签集，主题关注集，社交关系训练数据调试集，社交关系测试数据集。

1. 用户基本信息集

用户基本信息集包含超过 10 万条用户的基本信息数据，包括用户 id，用户名，位置信息，职业背景，受教育经历，个人简介，收到其他用户的点赞数，收藏帖子数，回帖数，关注数和被关注数。

2. 热门发帖集

包含近 2000 个热门发帖的基本信息，如作者 id，点赞数，评论数，发帖时间、发帖内容、回帖时间、回帖内容所属的所有主题标签。

3. 热门回帖集

包含近 7 万条回帖的基本信息，如作者 id，点赞数，评论数，回帖时间，回帖所属主题，回帖内容。

4. 主题标签集

包含社交网站的 26,000 个主题，包括主题的描述及主题的分类层次。

5. 主题关注集

包含网站用户对热门主题的关注情况。

6. 社交关系训练数据集

训练数据包含 2000 个左右用户之间的关注与被关注关系。

7. 社交关系测试数据集

测试数据集分为 3 份，每一份测试数据的评测分值占最终分值的 10%、30%、60%。

三、参评系统输入输出文件格式

1. 输入文件格式

输入文件包括社交网站用户基本信息集、热门发帖集、热门回帖集、社交关系训练数据调试集、社交用户和主题标签的关系集及主题标签集六组数据文件，文件格式如下：

用户基本信息集

```
<RECORD>
<id>5882418d5cc4dd47d408a8e8</id>
<user_name>5882418d5cc4dd47d418a819</user_name>
<location>广东</location>
<jobs>记者</jobs>
<education></education>
<personal_description>
-BORN THIS WAY-
</personal_description>
<likes_num>32313</likes_num>
<thanks_num>43657</thanks_num>
<collection_num><35435/collection_num>
<answer_num>65</answer_num>
<source_topic_id>1831496722</source_topic_id>
<source></source>
<follow_persons_num>45</follow_persons_num>
<updatetime>2016/12/21 16:57:13</updatetime>
<incId>1</incId>
</RECORD>
```

热门发帖集

```
<RECORD>
<id>435523354523</id>
<author_id>45654366634</author_id>
<topic_1>航天</topic_1>
<topic_2>工业，制造业</topic_2>
<topic_3>IT</topic_3>
<question>Java面试中会不会涉及源码的问题? </question>
<question_description></question_description>
<followCnt>353</followCnt>
<visitCnt>241</visitCnt>
<commentCnt>34</commentCnt>
<questionTime>2015/12/01 11:34:12</questionTime>
</RECORD>
```

热门回帖集

```
<RECORD>
<question_id>5256764332</question_id>
<author_id>6568575464</author_id>
<likes_num>453</likes_num>
<comment_num>123</comment_num>
<answer_time>2013/09/23 03:12:57</answer_time>
<topic>IT</topic>
<answer_content>
新建状态 (New) : 当线程对象对创建后, 即进入了新建状态, 如: Thread t = new MyThread();
就绪状态 (Runnable) : 当调用线程对象的start()方法 (t.start();), 线程即进入就绪状态。
处于就绪状态的线程, 只是说明此线程已经做好了准备, 随时等待CPU调度执行, 并不是说执行了t.start()此线程立即就会执行;
运行状态 (Running) : 当CPU开始调度处于就绪状态的线程时, 此时线程才得以真正执行, 即进入到运行状态。注: 就绪状态是进入到运行状态的唯一入口, 也就是说, 线程要想进入运行状态执行, 首先必须处于就绪状态中;
阻塞状态 (Blocked) : 处于运行状态中的线程由于某种原因, 暂时放弃对CPU的使用权, 停止执行, 此时进入阻塞状态, 直到其进入到就绪状态, 才有机会再次被CPU调用以进入到运行状态。根据阻塞产生的原因不同, 阻塞状态又可以分为三种:
1. 等待阻塞: 运行状态中的线程执行wait()方法, 使本线程进入到等待阻塞状态;
2. 同步阻塞 — 线程在获取synchronized同步锁失败(因为锁被其它线程所占用), 它会进入同步阻塞状态;
3. 其他阻塞 — 通过调用线程的sleep()或join()或发出了I/O请求时, 线程会进入到阻塞状态。
当sleep()状态超时、join()等待线程终止或者超时、或者I/O处理完毕时, 线程重新转入就绪状态。
死亡状态 (Dead) : 线程执行完了或者因异常退出了run()方法, 该线程结束生命周期。
</answer_content>
</RECORD>
```

主题标签集

```
<RECORD>
<incId>2</incId>
<topic>航天</topic>
<id>453532425</id>
<super_topic_id>3543466354</super_topic_id>
<topic_level>3</topic_level>
<topic_description>航天（Spaceflight），又称空间飞行、太空飞行、
宇宙航行或航天飞行，是指进入、探索、开发和利用太空（即地球大气层以外的宇宙空间，
又称外层空间）以及地球以外天体各种活动的总称。
</topic_description>
</RECORD>
```

主题关注集

```
<RECORD>
  <author_id>54365635645373</author_id>
  <topic_id>4535435324</topic_id>
<RECORD>
<RECORD>
  <author_id>4543654673657</author_id>
  <topic_id>4535435324</topic_id>
</RECORD>
<RECORD>
  <author_id>5647566765765</author_id>
  <topic_id>4535435324</topic_id>
</RECORD>
```


社交关系训练数据集

```
<RECOR>
  <author_id>543565654345</author_id>
  <followee_id>55646645t</followee_id>
</RECORD>
<RECOR>
  <author_id>765767346746</author_id>
  <followee_id>654654634</followee_id>
</RECORD>
<RECOR>
  <author_id>245465654543</author_id>
  <followee_id>654546665</followee_id>
</RECORD>
<RECOR>
  <author_id>234235466565</author_id>
  <followee_id>656576474</followee_id>
</RECORD>
```

社交关系测试数据集

```
<RECOR>
  <author_id>7675564353243</author_id>
  <followee_id>434536436</followee_id>
</RECORD>
<RECOR>
  <author_id>2134343255454</author_id>
  <followee_id>676654634</followee_id>
</RECORD>
<RECOR>
  <author_id>9768645634534</author_id>
  <followee_id>765756768</followee_id>
</RECORD>
```

2. 输出文件格式

输入文件左侧到右侧是用户 ID，用“----”符号隔开，表示左侧用户和右侧用户具有关注关系，具体格式如下图所示。最终输出数据的结果文件，参赛选手可以自行对其命名，

只需在 config 文件（config 配置文件请参见《stokis for text analysis 使用说明文档》）里注明。

```
454654635735----654643552235
876875464565----576739024064
767643546767----768564245354
878564634543----767545532564
876323565743----875623578534
763257684552----788745425457
576684325546----787435785322
124354665735----989745667896
768797353454----657425324655
878325787943----878452547686
```

四、评价指标

社交关系预测评价采用准确率、召回率以及 F 值作为评价指标。F 值的计算公式如下：

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

其中准确率 P、召回率 R。在分类结果中，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。准确率、召回率的计算公式如下：

$$P = \frac{a}{a + b}$$

$$R = \frac{a}{a + c}$$

最终排名以综合评分的 F 值作为依据。

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社交关系预测任务，共包括 3 组社交关系测试数据集。每组测试集可以提交任意次测试结果（上限 10 次），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：

社交关系预测第 1 组结果，准确率___，召回率___, F 值___；

社交关系预测第 2 组结果，准确率___，召回率___, F 值___；

社交关系预测第 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 值按照 1:3:6 的分值比例决定）。

六、注意事项

1、参赛方需要在指定时间内完成全部 3 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 3 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。