数据集用于构建社交机器人账号识别模型。数据集数据来源于推特平台，包括要包括账号个人信息、推文信息、邻居信息等数据。

数据集分为训练集和测试集，分别为train.csv和test.csv。另外提供了相关用户信息数据文件和用户关系数据文件，包含数据集中推特用户账号个人信息、推文信息和邻居信息，文件为json格式。

1. 训练集包含一批账号ID及其标签，文件名为train.csv，各字段以\t分隔，格式如下：

|  |  |
| --- | --- |
| **train.csv** | |
| ID | 用户ID |
| Label | 机器人身份标签，“1”表示为机器人账号，“0”表示为人类账号 |

1. 测试集包含一批不含标签的账号ID，文件名为test.csv，格式如下：

|  |  |
| --- | --- |
| **test.csv** | |
| ID | 用户ID |

（3）用户信息数据包含一批账号的详细数据，文件名为user.json，格式如下：

|  |  |
| --- | --- |
| **user.json** | |
| ID | 用户ID |
| tweet | 用户近期发布的200条推文 |
| profile | 用户个人信息，包括用户ID、用户昵称、个人描述、地点等 |
| id | 用户id，用户唯一标识符的整数表示形式 |
| id\_str | 用户id，用户唯一标识符的字符串表示形式 |
| name | 用户定义的用户名 |
| screen\_name | 用户使用的屏幕名称或名别，该值是唯一的 |
| location | 用户定义位置 |
| profile\_location | 用户定义位置的配置信息 |
| description | 用户个人描述 |
| url | 用户配置信息中关联的url |
| protected | 用户推文是否设置保护其推文 |
| followers\_count | 粉丝数 |
| friends\_count | 关注数 |
| listed\_count | 用户所属的公用名单的数量 |
| created\_at | 用户注册时间，UTC格式 |
| favourites\_count | 点赞数 |
| verified | 是否为认证用户 |
| statuses\_count | 发文数 |
| lang | 使用语言类型 |
| profile\_image\_url | 用户头像链接 |
| has\_extended\_profile | 是否存在扩展信息配置项 |
| default\_profile | 是否使用系统默认用户配置 |
| default\_profile\_image | 是否使用系统默认用户头像 |

（4）用户关系数据包含账号之间的关注、粉丝关系数据，文件名为edge.json，格式如下：

|  |  |
| --- | --- |
| **edge.json** | |
| seed\_user\_id | 主节点账号ID |
| relation | 关系类型，包括followers和friends，关系followers表示relation\_user\_id关注seed\_user\_id,关系friends表示seed\_user\_id关注relation\_user\_id |
| relation\_user\_id | 关系账号ID |