WeRateDogs数据整理报告

Seven He

2018年9月30日

I. 收集

本项目需要从三种来源来收集三种不同格式的数据：

- WeRateDogs 的推特档案，直接通过read_csv读取；
- 推特图像的预测数据，通过Requests库从URL编程下载；
- 每条推特的额外附加数据，通过Tweepy库从推特API获取；

II. 评估

对三个数据集采用了可视化评估与编程评估两种评估方式。

评估问题总结如下：

质量问题：

- tweet_id设为索引；
- rate_df中的source可以转化为类型特征；
- doggo，floofer，pupper和puppo空值为'None'，应该改为NaN；
- 数据类型转换：in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id转为int64；
- 数据类型转换：timestamp, retweeted_status_timestamp转为Datetime；
- 可以看出rate_df有2356行，而image_df只有2075行，tweet_df只有2352行，存在行缺失；
- rate_df中name字段存在不正确的名字，如'None'，'a', 'one'，'this', 'not'等等；
- rating_numerator和rating_denominator有不正确的评分；
- rate_df中的特征retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp不为null则说明是twitter转发，应过滤掉；

清洁度问题：

- doggo，floofer，pupper和puppo这四个为狗狗的stage，可以合并成一个类型特征；
- 三个表为同一个观察单位，应合并成一个表格；

III. 清洗

针对评估中总结的问题，具体清洗步骤如下：

1. tweet_id设为索引，再作三表合并，并清除掉image_df和tweet_df中的缺失行；

2. 只需原始评级，过滤掉转发（回复暂时未过滤）；
3. 数据类型转换：
    - source裁剪多余的字符，并转为类型特征；
    - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id转为int64；
    - timestamp转为Datetime；
    - join之后img_num，favorite_count，retweet_count变为float64，也需要转回为int64；
4. doggo，floofer，pupper和puppo空值为'None'，应该改为NaN，并合并为stage类型特征；
5. name移除错误的名字，这里简单的处理方式是把小写开头的都当作是错误名字，改为'None'，然后把所有的'None'设为NaN；
6. 评分不正确，移除掉某些评分明显不正确的数据；