数据整理过程

一、 数据收集

主要采取三种方法收集:

1.收集手头文件

定义 WeRateDogs 的推特档案转换后的 dataframe 文件为 df_t.

2.从互联网下载文件

定义神经网络预测狗品种的文件为 response. 新建 tsv 文件 prediction.tsv,并在其中写入 response 内容。定义神经网络预测狗品种的数据转换后的 dataframe 文件为 df_p.

3.API 下载文件

因为手机账号问题，无法使用推特 API，使用项目提供的文件，并定义 API 下载转换后的 dataframe 文件为 df_a.

二、数据评估

1.方法一：目测评估

将 df_p 转换为 csv 文件以便目测评估，命名为 prediction.csv，将 df_a 转换为 csv 文件以便目测评估，命名为 tweet_json.csv。

2.方法二：编程评估

采用 info，.describe ，.value_counts，.sort_values 等方法进行。并通过循环，定义函数等方法检查狗的‘地位’和评分是否与推文中的数据一致。

3.结果一：质量问题

（1）WeRateDogs 的推特档案 df_t

前两行数据是 2017 年 8 月 1 号以后的，因为图像预测权限，无法使用。

有 181 行转推内容。

有 78 行回复内容。

rating_numerator 列有异常值,rating_denominator 列有异常值。

部分狗的“地位”与 tweet 内容里的不一致。

timestamp 列数据类型为 str。

rating_denominator 列最小值为 0.（可以与这两列异常值问题一并处理）

（2）图像预测数据 df_p

p1 列值有大写有小写。

（3）Tweepy 下载数据 df_a.

前两行数据是 2017 年 8 月 1 号以后的，因为图像预测权限，无法使用。

id 列数据类型为浮点，与其他两个数据集 id 列类型为整型不一致。

created_at 列与 df_t 中 timestamp 列重复。

favorite_count 和 retweet_count 列数据类型为浮点

4.结果二：整洁度问题

tweet_id, in_reply_to_status_id, in_reply_to_user_id, source 这四列同时出现在 df_t 和 df_a 中，一个观察单元同时出现在两个表格中。

WeRateDogs 的推特档案 df_t 中，狗“地位”（即 doggo、floofer、pupper 和 puppo）4 列的列名是值，不是变量名。

三、数据清理

1 .删除 df_t_clean 和 df_a_clean 的前两行数据。

2 .删除 df_t 中 181 行转推内容及数据集的 retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp 列。

3. 删除 df_t 中 78 行回复内容及数据集的 in_reply_to_status_id, in_reply_to_user_id 列。

4 .删除 rating_numerator 值大于等于 20 和等于 0 的行，删除 rating_denominator 值不等于 10 的行。

5. 在 df_t_clean 新建 stage 列，值是推文里提取的狗狗地位数据，然后修改该列数据为 category 类型并删除 doggo, floofer, pupper, puppo 列。

6 .将 df_t_clean 中 timestamp 列数据类型改为 datetime.

7. 将 df_a_clean 中 created_at 列删除。

8 .将 df_a_clean 的 id 列数据类型改为整型。

9. 将 df_a_clean 中的 favorite_count 和 retweet_count 列数据类型改为整型。

10. 清除 df_a_clean 的 source 列，保留三个数据集中的 tweet_id 和 id 列，用以合并数据集。

11 .将 df_p_clean 的 p1 列全部改为小写。

12 .合并三个数据集并分别导出至 csv 文件。

13 .根据不同分析目标合并相关的两个数据集。