数据整理总结

收集

1 WeRateDogs 的推特档案

2 推特图像的预测数据，即根据神经网络，对出现在每个推特中狗的品种（或其他物体、动物等）进行预测的结果

3 使用 Python Tweepy 库查询 API 中每个推特的 JSON 数据要包含转发数（retweet_count）和喜欢数（favorite_count）

评估

质量

archive

·错误的数据类型 : 'tweet_id'，'timestamp'，需要更改

· Rating （评分） 包含不准确数据，需要对可疑数据进行搜索并更新

·包含了 retweets ，在分析中不需要 : 'retweeted_status_id','retweeted_status_user_i','retweeted_status_timestamp' ， 'in_reply_to_status_id','in_reply_to_user_id' 所以删除

· Source 列包含不需要的内容，简化更新

· Name 列包含不正确的数据 (eg. 'a','the','an'，None...)，字母大小写不统一，搜索丢失的名字然后统一格式更新

tweets

·错误的数据类型 :tweet_id，更改数据类型

image

·错误的数据类型 :id， 更改数据类型

· p1 & p2 & p3 的数据首字母大小写不统一，p2 和 p3 的内容在分析中似乎不必要，统一格式并更新

整洁度

archive

·狗的地位分为 4 列，要整合在一列

·把 tweets 和 image 合并到 archive

评估类型

目测评估&编程评估

清洗

编程方法清洗