项目：分析WeRateDogs的推特数据

> 通过完成本项目来提高数据整理技能。

by **Batu Mengkai**

目录

1. 简介

1.1 背景

WeRateDogs是一个推特主，他以诙谐幽默的方式对人们的宠物狗评分。这些评分通常以 10 作为分母。但是分子则一般大于 10：11/10、12/10、13/10 等等。WeRateDogs拥有四百多万关注者，曾受到国际媒体的报道。WeRateDogs下载了他们的推特档案，专门为本项目使用。这个档案是基本的推特数据（推特ID、时间戳、推特文本等），包含了截止到 2017 年 4 月 1 日的 5000 多条推特。

1.2 项目目标

清洗 WeRateDogs 推特数据，创建有趣且可靠的分析和可视化。这份推特只包含基本的推特信息。在分析和可视化前，还需要收集额外的数据、然后进行评估和清洗。

1.2 探索的问题

- 问题1：推特主一天当中喜欢在哪个时段发推特？一周当中喜欢在哪天发推特？
- 问题2：推特主对狗的评分高低是否与其粉丝中的大多数喜好一致？
- 问题3：哪种地位的狗狗最受人们喜欢？
- 问题4：随着时间的推移，点赞数和转发数是怎么变化的？

2. 整理数据

2.1 收集数据

本次研究收集了三个数据集，分别是`twitter-archive-enhanced.csv`,`image-predictions.tsv`,`tweet_json.txt`,介于本地区无法使用twitter API以及提供的下载链接，都采用使用现成的数据集。`tweet_json.txt`文件内容是json格式，提取了每条tweet信息的tweet_id, favorite_count, retweet_count，并写入名为tweet_json的DataFrame文件中。

2.2 评估数据

通过目视评估和代码评估，总共发现如下问题，按质量问题和结构问题分别开来。
为避免代码评估不准确，先将NaN和None表示空值的方式统一成NaN方式。
目视评估是通过excel查看文件，以及随机打印DataFrame文件，查看发现存在的问题
代码评估是通过info,describe,value_counts,duplicated等方法查看数据

发现的数据集存在的问题：

质量问题：

`twitter_archive_clean` 存在的问题

1. 空值表示方式有NaN, None两种
2. timestamp列,retweeted_status_timestamp列数据类型是object
3. tweet_id列数据类型是int,in_reply_to_status_id和in_reply_to_user_id列数据类型是float,应该是字符串
4. in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp等列数据缺失
5. rating_numerator有一些异常值，
6. rating_numerator有一些值提取错误，例如9.75/10提取了75,没有考虑小数的可能
7. rating_denominator列有一些除了10以外的数据
8. name列含有一些诸如a,an,the等小写的非名字的词

`twitter_image_clean` 存在的问题

9. tweet_id数据类型是int64,应该是str
10. p1,p2,p3可以使用更描述性的词

`tweet_json_clean` 存在的问题

11. tweet_id数据类型是int64,应该是str

结构问题：

1. doggo，floofer,pupper,puppo这几列是属于一个变量
2. 三个数据表格都可以按照tweet_id进行合并。
3. 含有181个retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp的数据，我们不需要转发数据
4. 删除不需要的列（in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp列）

2.3 清理数据

清理数据过程，为清理过程更为有效，首先提高解决了数据缺失问题，然后解决了数据结构问题，最后解决了数据内容问题。

2.3.1 解决数据缺失问题

1. in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp等列数据缺失，一些推特不存在回复及被转发，所以有些记录不存在这些数据是合理，一些数据无法获得。对此数据缺失不做处理。
2. 空值表示方式有NaN, None两种，为了更好用代码评估数据，在进行代码评估开始之前就对已经此错误进行了处理。有NaN, None两种统一为NaN。使用replace函数，将所有None的单元格替换为NaN

2.3.2 解决数据结构问题

1. 含有181个retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp的数据，我们不需要转发数据，使用方法： 筛选出这些列非空值的行，将其删除
2. 删除不需要的列（ in_reply_to_status_id， in_reply_to_user_id， retweeted_status_id，retweeted_status_user_id， retweeted_status_timestamp列），使用drop函数删除这些列
3. doggo，floofer,pupper,puppo这几列是属于一个变量，将其变为一列。使用melt函数将这四列合为一列，介于这四列数据有空值，变成一列会有重复数据，将重复的数据进行删除。
4. 三个数据表格都可以按照tweet_id进行合并。使用merge函数，三个数据集进行两次按照tweet_id对应合并成 `twitter_master` DataFrame

2.3.3 解决数据质量问题

1. 更改数据类型问题：timestamp列数据类型是object，将其使用pd.todatetime函数改为datetime数据类型。将tweet_id列数据类型使用astype函数更改为字符串类型
2. rating_numerator有一些值提取错误，例如9.75/10提取了75,没有考虑小数的可能。重新编写正则表达式，从text提取，考虑小数点的可能性
3. rating_denominator列有一些除了10以外的数据，将小于10的删除，将大于10的同比例缩小到分母为10。(rating_numerator = rating_numerator*10/rating_denominator)
4. rating_numerator有一些异常值,将大于20的异常值删除
5. name列提取了含有一些诸如a,an,the等的非名字的词，对应的text信息里面也没有提供名字，所以删除，具有的特点是都是小写，将小写词筛选出来删除。
6. p1,p2,p3可以使用更描述性的词，p1,p2,p3更改为prediction1,prediction2,prediction3

2.3.4 保存清理后的数据

使用to_csv函数，将清洗后的数据twitter_master存为twitter_archive_master.csv文件。

探索性数据分析及结论请见报告act_report.pdf