2018年7月27日 星期五

Tweeter Data Wrangle Report

收集

twitter-archive-enhanced 为现有文件通过pandas中的read_csv直接读取。

tweet_json 是通过 json load 从 tweet_json.txt中读取。

image_predictions 是通过 request 请求库从网站上下载所得。

评估

通过在wrangle_act.ipynb中的分析，得出如下数据问题。

质量问题：11条

twitter_archive 表格

- 'tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id', 'retweeted_status_user_id' 是浮点，不是字符串
- 'rating_denominator' 有23条记录不是10为分母
- 'rating_numerator' 分子记录值有 Outliner，分子大于50的记录有18条
- twitter_archive 表格中有2356条记录，image_predictions 表格有2075条记录；部分记录没有图片
- 需要清理掉推特喜爱数为 0 的记录
- 列名'name' 改成 ‘dog_name’
- 'dog_name' 列中的‘None，应该使用np.nan

`tweet_json` 表格

- 列名'lang' 改成 ‘language’
- ‘lang’ 列中的语言都是缩写，修改成完整语言名称

`image_predictions` 表格

- 表格中列名过于简要，不易理解。p1 改成 p1_name; p1_dog 改成 p1_dog_outcome，依此类推更改 p2，p3
- 统一p1，p2，p3中狗品种的名字格式，用空格替代原本的‘-’，并让单词首字母大写
- p1_conf 中数据Format应改成百分比

清洁度问题：3条

- `twitter_archive` 表格中 doggo,floofer,pupper,puppo 四列应该合并为 dog_stage
- `tweet_json` 表格中的 favorite_count 和 retweet_count, lang 是 `twitter_archive_master` 表格的一部分
- image_predictions表格是 `twitter_archive_master` 表格的一部分

清理

根据之前评估出的问题，逐条做数据清理。

清洁度问题：3条

- `twitter_archive` 表格中 doggo,floofer,pupper,puppo 四列应该合并为 dog_stage

  使用melt方法，将原来的4列合并，针对有的纪录存在2个dog_stage，将这两个dog_stage合并在一条纪录下。

- `tweet_json` 表格中的 favorite_count 和 retweet_count, lang 是 `twitter_archive_master` 表格的一部分

  使用pd.merge方法将其合并

- image_predictions表格是 `twitter_archive_master` 表格的一部分

  使用pd.merge方法将其合并

质量问题：9条

`twitter_archive` 表格

- 'tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id',

'retweeted_status_user_id' 是浮点，不是字符串

  使用 astype 方法转化成 str 格式

- 'rating_denominator' 有23条记录不是10为分母，'rating_numerator' 分子记录值有 Outliner，分子大于50的记录有18条

  使用正则表达式再次从Text中提取 rating_numerator 和 rating_denominator，提取带有小数的分子，Text中如果有2条 rating，到单独处理。

  对 rating_denominator 做10的Normalize

- `twitter_archive` 表格中有2356条记录，`image_predictions` 表格有2075条记录；部分记录没有图片

  merge 表格时使用 inner merge，并且使用 dropna(subset=[‘jpg_url']) 方法清理。

- 需要清理掉推特喜爱数为 0 的记录

  使用条件筛选去掉推特喜爱数为 0 的记录

- 列名'name' 改成 ‘dog_name’

  使用 rename 方法清理

- 'dog_name' 列中的‘None，应该使用np.nan

  定义 repalce_none 函数，使用 apply 方法调用

`tweet_json` 表格

- 列名'lang' 改成 ‘language’

  使用 rename 方法清理

- ‘lang’ 列中的语言都是缩写，修改成完整语言名称

  定义函数 abbreviate_language，使用apply方法清理

`image_predictions` 表格

- 表格中列名过于简要，不易理解。p1 改成 p1_name; p1_dog 改成 p1_dog_outcome，依此类推更改 p2，p3

  使用 rename 方法清理

- 统一p1，p2，p3中狗品种的名字格式，用空格替代原本的’-’，并让单词首字母大写

  使用apply方法，配合 str 方法清理