13.wrangle_report

2018年6月6日星期三 9:32

一 数据收集

- 拿到手的是三分数据集，一个是含有基本信息的twitter_archive_enhanced.csv，一个是使用提供的URL从网上下载的image-predictions.tsv，最后一个是从Tweepy库查询 API 中每个推特的 JSON 数据，把所有 JSON 数据存储到一个名为 tweet_json.txt文件。
- 当然，由于我没有翻墙，所以用了提供好的这个txt文件，从txt文件中利用.json函数取出需要的id、转发数和点赞数，并生成dataframe。

二 数据评估

1. twitter_archive表

- 首先用.head查看了twitter_archive的大致情况，并结合.info发现有一些数据是有缺失的（很多缺失：in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp；较少缺失：expanded_urls）以及错误的数据类型（tweet_id，in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id，timestamp、retweeted_status_timestamp）
- 因为要使用tweet_id作为后面连接三个表格的key，所以查询一下tweet_id有无重复性数据
- 然后随机输出5个数据，目测有无问题，发现评分这边，分母有不是10的数据，所以对rating_denominator进行.value_counts的操作，发现很多评分都是有问题的；
- 在随机目测中也发现狗狗名字为‘an’的情况，输出这一行代码查看，发现应该是None才对，查看了大部分名字为‘a’的狗狗，text显示其名字应该为None。
- 还有博主自己转发的推特181条以及没有图片的推特100条

2. image_predictions表

- 用.head查看了该表的大致情况，并结合.info发现错误的数据类型（tweet_id、img_num），以及列名p1,p2等可以用更有描述性的列名才对；
- 因为要使用tweet_id作为后面连接三个表格的key，所以随后查询一下tweet_id有无重复性数据

3. df（从 tweet_json.txt提取点赞和转发的表）表

- 用.head查看了该表的大致情况，并结合.info发现错误的数据类型（tweet_id），查看是否读取了txt中的所有数据；

4. 清洁度问题

- twitter_archive表中doggo、floofer、pupper、puppo应该合成一列
- 删除研究中不需要的列
- 三个表应该合成一个表

三 数据清洗

- 先对之前的文件进行.copy操作

1. 数据缺失

- 优先处理缺少数据的问题，结果由于转发等数据拿不到了，所以没有办法进行补充。

2. **清洁度**

- 随后处理清洁度问题，因为整合好整个表格再去处理质量问题，就会避免重复处理。
- 首先根据项目要求twitter_archive表格中删除掉不要的博主自己转发的行和别人转发的列以及image_predictions_clean表中不需要的P2和P3预测的列
- 随后使用.melt函数将twitter_archive 表格中的doggo、doggo、pupper和puppo四列合成nickname列（注意删除合并后出现的重复的None行和重复地位的数据）
- 然后使用merge函数将三个表合成twitter_archive_master一个表，做这一步的时候注意将三个表中的id转化成str

3. **质量问题**

- 质量问题先进行数据类型的转化，由于之前清洁度工作的时候很多类型有问题的列已经删除，所以这里只对还没有处理的列（timestamp、img_num）进行转化处理。
- 处理评分分母有多种的问题，因为分母类型实在太多了，所以进行统一转化，将rating_numerator全部除以原来的分母再乘以10，获得新列，这样转化为以10为分母的评分机制，然后再删除原先的分子和分母列以及超过20的异常值
- 接下来处理狗狗名字不对的问题，就像前面说的，狗狗名字为a和an的大部分是提取的时候出现的问题，查看原来的text，多数是没有名字，所以用.replace替换成None；（这里不采用正则重新取名的原因是：因为大家提出自己狗狗名字的方式各有不同，即使用正则重新读取名字也可能会造成名字很多取错）
- 使用rename将p1,p1_conf和p1_dog列的名称进行修改为prediction、prediction_PR和prediction_dog；

四 探索性数据分析

- 该部分见art_report.pdf

五 结论

- 该部分见art_report.pdf