数据收集、评估与清洗项目

项目背景

本项目将要整理 (以及分析和可视化) 的数据集是推特用户 @dog_rates 的档案，推特昵称为 WeRateDogs。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。为什么会有这样的评分？因为 "They're good dogs Brent." 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，并通过电子邮件发送给优达学城，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。

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。

项目细节及步骤

你在这个项目中的任务如下：

数据整理，其中包括： 收集数据 评估数据 清洗数据

对清洗过的数据进行储存、分析和可视化 书面报告 1) 你的数据整理工作 和 2) 你的数据分析和可视化

1. 收集以下三份数据，为三种不同的文件类型：

- WeRateDogs 的推特档案。这个数据文件是直接提供的，文件名为 twitter_archive_enhanced.csv
- 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。这个文件你需要使用 Python 的 Requests 库和以下提供的 URL 来进行编程下载。下载用的 URL：https://raw.githubusercontent.com/udacity/new-dand-advanced-china/master/%E6%95%B0%E6%8D%AE%E6%B8%85%E6%B4%97/WeRateDogs%E9%A1%B9%E7%9B%AE/image-predictions.tsv
- 每条推特的额外附加数据，至少要包含转发数（retweet_count）和喜欢数（favorite_count），还可以收集任何你觉得有趣的字段。使用 WeRateDogs 推特档案中的推特 ID，使用 Python 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 的文件中。

2. 对项目数据进行评估

收集上述三个数据集之后，使用目测评估和编程评估的方式，对数据进行质量和清洁度的评估。在你的 wrangle_act.ipynb Jupyter Notebook 中记录评估过程和结果，最终列出至少 8 个质量问题 和 2 个清洁度问题。要符合项目规范，必须对项目动机中的要求进行评估（参见上一页课程的 关键要点 标题）。

3. 对项目数据进行清洗

对你在评估时列出的每个问题进行清洗。在 wrangle_act.ipynb 展示清洗的过程。结果应该为一个优质干净整洁的主数据集（pandas DataFrame 类型） （如果都是以推特 ID 为观察对象的一些特征列，则清理最终只能有一个主数据集，如果有其他观察对象及其对应的特征字段，可以创建其他的数据集，同样需要清理）。同样地，必须符合项目动机的要点要求。

4. 对项目数据进行存储、分析和可视化

将清理后的数据集存储到 CSV 文件中，命名为 twitter_archive_master.csv。

在 wrangle_act.ipynb Jupyter Notebook 中对清洗后的数据进行分析和可视化。必须生成至少 3 个见解和 1 个可视化。

以下为数据清理和分析的具体步骤。

首先进行对三种数据类型的数据集分别进行导入，并同时通过可视化和编程法，对数据集进行查看。

通过查看发现了至少以下 8 种数据质量问题和 2 种数据清洁度问题。

数据质量问题

twitter_archive 数据集：

1. 我们只需要含有图片的原始评级 (不包括转发和回复的)，需要删除转发和回复的数据。同时，由于 twitter_archive 数据集有 2356 行数据，而 image_predictions 数据集只有 2075 行，存在没有图片的数据行，需要同时删除。
2. 多列数据缺失严重，包括有 in_reply_to_status_id，in_reply_to_user_id，retweeted_status_id，retweeted_status_user_id，retweeted_status_timestamp，expanded_urls
3. timestamp 列类型不是时间类型，需转化成时间类型
4. 名字列中，有些数据不准确
5. 打分的分子和分母列中，有些数据不准确
6. 狗的类型列中，有些数据不准确，可能需要手动处理

图片预测数据集

7. 从统计结果来看，大部分都是针对图片 p1 进行的预测，且由于图片 2,3 的可信度较低，直接删除针对图片 2,3 的预测列
8. 删除针对图片 p1 的预测非狗的行

数据清洁度问题

1. 有共同的 tweet_id 列，三个表格应该合并在一起，尤其是 tweet_df 只是在 twitter_archive 基础上的补充
2. doggo, floofer, pupper, puppo 四列表示的实际上是同一个特征，狗狗的地位，按照数据整洁度的标准（Each variable forms a column），应该将其合并为一列来表示，合并时需要注意，存在一只狗狗有多个地位的情况

针对以上出现的问题，都分别运用编程的方式，一一进行了处理。但是，我这次所发现的问题，只是本数据集的部分问题。而且某些问题非常特别和具体，只有通过针对性的编程修改来完成。在数据处理完成后，还进行了检查验证。

最后通过三个表格中共同的 tweet_id 列，将三个表格合并在一起，同时删除了重复列和一些不相关的列，存储在 twitter_archive_master 文件里。