WeRateDogs 推特数据整理报告

一、项目综述

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。我们的目标是清洗 WeRateDog 的推特数据，为后续的分析和可视化提供可靠的数据。

需要清洗整理的数据有三份：

1、WeRateDogs 的推特档案

本项目中的 WeRateDogs 的推特档案数据是从原始提供的 5000 多条推特中筛选了 2356 条包含评分的推特，并提取了基本信息。原始档案中有一列包含每个推特的文本，项目前期利用这一列数据提取了评分、狗的名字和“地位”（即 doggo、floofer、pupper 和 puppo）。

项目前期采用编程方式提取了这些数据，但是并没有做好。评分并不都是正确的。狗的名字和地位也有不正确的。需要重新评估和清洗这些数据。

2、通过推特 API 获取附加数据

WeRateDogs 的推特档案遗漏了两个关键数据：转发数（retweet count）和喜爱数（favorite count），因此需要从推特 API 中重新收集这部分数据。

因为无法访问 Twitter，所以在本项目中提供替代方案：一份已采集好的 Twitter 数据。

3、图像预测数据

通过一个可以对狗狗种类进行分类的神经网络，运行这份推特档案中的所有图像，获得了一份图像预测结果表格。其中包含了预测结果的前三

名，推特 ID，图像 url 以及最可信的预测结果对应的图像编号（由于推特最多包含 4 个图片，所以编号为 1 到 4）。

从推特中可以看到，对图像的一号预测（p1）是准确最高的。

二、数据整理

2.1 收集数据：

1.WeRateDogs 的推特档案。这个数据文件是项目直接提供的，可以将其当作是“资料来源：手头文件”来收集。

2.通过推特 API 获取附加数据。因无法访问 Twitter，采用项目提供的已采集好的 Twitter 数据，该文件为 txt 格式，每一行为一条独立的 twitter 信息，格式为 JSON。

3. 推特图像的预测数据。这个数据文件可以当作是“资料来源：互联网文件下载”来收集，需要使用 Python 的 Requests 库和项目提供的 URL 来进行编程下载。

2.2 评估数据：

质量问题

1.WeRoteDog 推特档案：

- 狗狗的名字清洗有问题,有大量狗狗的名字为"None",甚至部分狗狗名字为"a"、"an"、"the"等
- 部分带特殊字符的狗狗名字清洗错误，如"O'Malley" 被清洗成"O"
- 缺少转发数（retweet_count）和喜欢数（favorite_count）
- 包含有 181 条转发数据
- 推特发布时间戳“timestamp”的类型不是时间类型

- 狗狗的评分，分子清洗有问题，查看"rating_numerator"维度，发现分子最小 0 分，最大 1776 分
- 狗狗的评分,分母清洗有问题,查看"rating_denominator"维度,发现分母最小 0 分，最大 170 分
- 狗狗的评分，分子是小数，但是只提取了小数点后面的数字的情况。如 11.27/10，错误提取为 27/10
- 狗狗的评分，包含有多个狗狗的总评分：99/90，规律是：分母是 10 的 N 倍，且分子可以被 N 整除
- 狗狗的评分，存在比较特殊的错误，比如"24/7"指的是 7 天 24 小时，并不是一条评分。
- 同一个推特中存在两处分数形式的数字，提取的是第一个，但是可能第二个才是正确
- 包含有 2017 年 8 月 1 日之后的数据（2 条）

2.推特图片预测数据：

- 狗狗的种类名称不规范，首字母有的大写有的小写
- 缺失部分数据（WeRoteDog 推特档案 2356 条，推特图片预测数据 2075 条）

3.Twitter API 的附加数据

- 推特 ID 命名不规范（是"id"而不是"tweet_id"）
- 包含有 177 条转发数据（"retweeted_status"有 177 条非空记录）
- 包含有 2017 年 8 月 1 日之后的数据（2 条）

格式问题

WeRoteDog 推特档案：

- 狗狗的“地位”评级数据分散多列

Twitter API 的附加数据：

- 三个数据集都是以"tweet_id"为观察对象，应该合并为一个数据集

2.3 清洗数据

Twitter API 的附加数据

- 删除 2017 年 8 月 1 日之后的数据
- 删除转发的数据
- 提取感兴趣的字段【推特档案编号（id)、转发数（retweet count)和喜爱数（favorite count)】
- 规范列名

推特图片预测数据

- 提取图片是狗狗的数据
- 提取感兴趣的字段【推特档案编号（tweet_id)、预测的狗狗种类（p1)】
- 统一狗狗种类的命名规范（首字母大写）
- 修改列名

WeRoteDog 推特档案

- 修正错误的数据类型（推特发布时间戳“timestamp”的类型不是时间类型）
- 删除 2017 年 8 月 1 日之后的数据
- 删除转发的数据
- 删除不是狗狗的数据（通过与清洗好的推特图片预测数据集合并数据集的方式）

- 重新提取狗狗的名字
- 重新提取狗狗的评分
- 重新提取狗狗的“地位”评级
- 提取感兴趣的字段【推特档案编号（tweet_id），时间戳（timestamp）,狗狗的种类（dog_type),新提取的狗狗名字(new_name),新提取的狗狗评分(rating_numerator),新提取的狗狗地位（stage)】
- 修改列名

2.4 合并数据集

将清洗干净的"Twitter API 的附加数据"和"WeRoteDog 推特档案"(已经和"推特图片预测数据"合并)进行合并，为后续的分析和可视化提供清洗干净的主数据集