

狗狗评分数据清理项目（wraggling adv dog_age@twitter）

[TOC]

Understand Project

/ background

本项目要整理 (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。

WeRateDogs 的推特档案包括 5000 多条推特的基本信息, 但并不包括所有内容。不过档案中有一列包含每个推特的文本, 我用这一列数据提取了评分、狗的名字和“地位”(即 doggo、floofer、pupper 和 puppo) ——这使数据得以“完善”。在这 5000 多条中, 我只筛选出了 2356 条包含评分的推特数据。

text	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU	13	10	Phineas	None	None	None	None
This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the whole bit. 13/10	13	10	Tilly	None	None	None	None
This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know when one may strike. 12/10 https://t.co/...	12	10	Archie	None	None	None	None
This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us https://t.co/tD36da7qLQ	13	10	Darla	None	None	None	None
This is Franklin. He would like you to stop calling him "cute." He is a very fierce shark and should be respected as such. 12/10 #BarkWeek	12	10	Franklin	None	None	None	None
Here we have a majestic great white breaching off South Africa's coast. Absolutely h*ckin breathtaking. 13/10 (IG: tucker_marlo) #BarkWeek	13	10	None	None	None	None	None
Meet Jax. He enjoys ice cream so much he gets nervous around it. 13/10 help Jax enjoy more things by clicking below https://t.co/Zr4hWfAs1H https://t.co/tVJBRMnhxl	13	10	Jax	None	None	None	None
When you watch your owner call another dog a good boy but then they turn back to you and say you're a great boy. 13/10 https://t.co/...	13	10	None	None	None	None	None
This is Zoey. She doesn't want to be one of the scary sharks. Just wants to be a snugly pettable boatpet. 13/10 #BarkWeek https://t.co/...	13	10	Zoey	None	None	None	None
This is Cassie. She is a college pup. Studying international doggo communication and stick theory. 14/10 so elegant much sophisticated	14	10	Cassie	doggo	None	None	None
This is Koda. He is a South Australian deckshark. Deceptively deadly. Frighteningly majestic. 13/10 would risk a petting #BarkWeek hi	13	10	Koda	None	None	None	None
This is Bruno. He is a service shark. Only gets out of the water to assist you. 13/10 terrifyingly good boy https://t.co/u1XPQMI29g	13	10	Bruno	None	None	None	None
Here's a puppo that seems to be on the fence about something haha no but seriously someone help her. 13/10 https://t.co/BxvuXk0Uc	13	10	None	None	None	None	puppo
This is Ted. He does his best. Sometimes that's not enough. But it's ok. 12/10 would assist https://t.co/f8dEDcrKSR	12	10	Ted	None	None	None	None
This is Stuart. He's sporting his favorite fanny pack. Secretly filled with bones only. 13/10 puppered puppo #BarkWeek https://t.co/y70k...	13	10	Stuart	None	None	None	puppo

关于狗地位的评分解释是这样的:

THE DOGTIONARY

doggo
/'dɒɡo/
noun

- A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.
- A pupper that appears to have its life in order. Probably understands taxes and whatnot.

"That's a really good doggo."
"I give my doggo a firm pat every night before bed."

pupper
/'pʌpə/
noun

- A small doggo, usually younger. Can be equally, if not more mature than some doggos.
- A doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo.

"H*ck, that's one pettable pupper."
"How many puppers could I fit on my body at once, if I were lying down?"

puppo
/'pʌpɒ/
noun

- A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
- A dog with a mixed bag of both pupper and doggo tendencies.

"My puppo is still learning what it takes to be a trustworthy doggo."
"I would hug that puppo so passionately."

blep
/'blep/
verb

- An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one's tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

"My doggo did a h*ck of a blep the other day."
"Get a load of this blep I captured."

snoot
/'snoʊ/
noun

- The nose of a dog. Usually found in places the dog may not fit. The location of the snoot may hint at where the dog's interest lies.

"That is a beautiful snoot."
"I've been trying to boop my neighbor's dog's snoot for six years."

floof
/'floʊ, 'floʊ/
noun

- Any dog really. However, this label is commonly given to dogs with seemingly excess fur. Comical amounts of fur on a dog will certainly earn the dog this generic name.
- Dog fur. The term holds true whether the fur is still on the dog, or if it has been shed off.

"Check out that majestic floof!"
"The floof on my dog has gotten out of control but I don't see anybody complaining any time soon."

- 貌似规则是这样的从年龄上分从年轻到老: pupper < puppo < doggo
- 剩下的3类, blep 为听话、sonnt 为嗅觉好、floof 为毛多

/ data files

- twitt_json.txt (读入为 df_api) 后收集来的数据，用于修正原始数据（提供了链接）
- twitter-archive-enhanced.txt (读入为 df_arc)原始数据
- image-predictions.tsv (读入为 df_img)图像预测结果（要求使用 request 库下载，[/LINK/](#)）

/ gole

- 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）。
- 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。
- 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。如果有其他观察对象的数据集存在，需要多个表格，那么要给这些文件合理命名。另外，你也可以把清洗后的数据存储到 SQLite 数据库中（如果这样存储的话，该数据库文件也需要提交）。
- 在 wrangle_act.ipynb Jupyter Notebook 中对清洗后的数据进行分析 and 可视化。必须生成至少 3 个见解和 1 个可视化。
- 项目汇报
创建一个 300-600 字的书面报告，命名为 wrangle_report.pdf，在该报告中简要描述你的数据整理过程。这份报告可以看作是一份内部文档，供你的团队成员查看交流。

Solutions

/ data intergration

根据题目的要求，进行了更新（开始跑偏了，直接自己分析的 df_api）

- 主数据 df_arc
 - 主键 twitter_id，做了一些列的清理
- 附加列 df_img
 - 增加狗狗的第1预测分类,确信度和是否狗狗的分类
 - 增加的列 ['p1','p1_conf','p1_dog']
- 附加列 df_api
 - 增加 retweet_count favorite_count 后又觉得 display_text_range 也有价值也增加了
 - 增加的列 ['p1','p1_conf','p1_dog']
- 最终主数据结构：
 - ['tweet_id', 'rating_numerator', 'rating_denominator', 'name', 'doggo', 'floofer', 'pupper', 'puppo', 'retweet_count', 'favorite_count', 'display_text_range', 'p1', 'p1_conf', 'p1_dog', 'clean_text']

/ data tidying

从以下几个维度进行了数据整理（代码在jupyter notebook中）。

// 转换为时序索引

转换为时序索引方便进行时序分析。

// 提取回复长度

display text range 的格式是嵌套的list，从中提取出最大文字，并转换为数值。

// 提取文本

将 text 中后续评分和链接内容排除，提取为新列，并在后续用这些新列做词云的输入。

/ jupyter index

本分析的所有功能为在 jupyter notebook 中完成，附加利用了 wrangling2.py 作为常用分析函数。并且在词云生成时需要使用 t1, t2 两个图片文件。

jupyter notebook 使用 toc 生成目录的话会比较方便浏览，其中展开的部分为项目涉及部分：

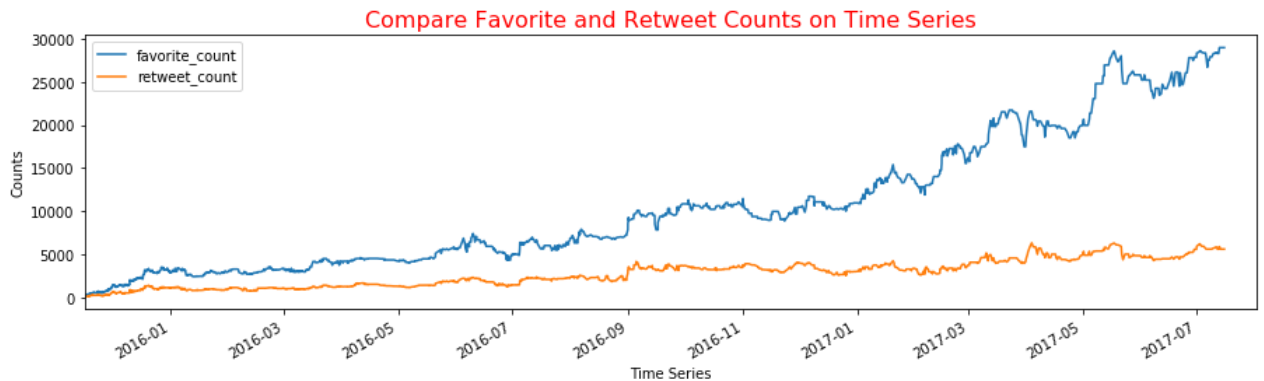
Contents

- ▼ 1 收集
 - 1.1 / import lib
 - 1.2 / display setting
 - 1.3 / load df
- ▶ 2 评估 - twitt_json
- ▶ 3 评估 - twitter-archive-enhanced
- ▶ 4 评估 - image-predictions
- ▼ 5 评估 - 总结
 - 5.1 / load df (inspect)
 - 5.2 / merge df
 - 5.3 / persistence
- ▼ 6 清理 (Tidiness)
 - 6.1 / load df
 - 6.2 / created_at
 - 6.3 / display_text_range
 - 6.4 / full_text
 - 6.5 / data type
 - 6.6 / check - null
 - 6.7 / check - duplicated
 - 6.8 / check outlier
 - 6.9 / persistence
- ▶ 7 探索 (非项目要求)
- ▶ 8 结论 (非项目要求)

Conclution

/ favorite 和 retweet 时序分析

- 2016年上半年之前, favorite 数量大概是 retweet 的两倍
- 但再这之后, favorite 数量大量上涨, retweet 数量上涨十分缓慢(两者之比达到6倍)
- 推测相关因素如下:
 - 可以看出 twitter 增长非常迅速(可惜缺少用户量相关的数据)
 - 但是人们愿意付出更多一点时间 retweet 的时间在减少, 可能原因是当人接触到更多的 twitter 信息后, 能够 retweet 的注意力已经没有什么增长空间了(注意力处于饱和状态)



/ word cloud 分析

- 对评论使用 word cloud 进行分析
- 去掉了 stop words
- 图像为 twitter 英文字符(小鸟图不太美观)



/ text range 分析

- text range 改名为 text range 更为明确
- 数据做了过滤(过滤掉了个别 160 字符的)
- 数据有左偏斜趋势 (不能断定) 因为在140字的限制上有大量出现, 所以明显存在人为调整
- 有些数据超出了140

- 后续可以做异常值分析(按说不应该有超出, 也可能是正则化过滤时留下的问题)

