Data Wrangling Report

Jie Gu
03/02/2019

This file documents the data wrangling process of the WeRateDogs dataset.

Gathering

Three pieces of data were gathered.

1. The WeRateDogs Twitter archive. It was downloaded manually from the following link: `twitter_archive_enhanced.csv`
2. The tweet image predictions. It was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Each tweet's retweet count and favorite ("like") count. Python's Tweepy library was used to query the Twitter API and the data was stored in a file called `tweet_json.txt` file.

Assessing

A few quality and tidyness issues have been detected.

Quality

twitter-archive-enhanced.csv table

- Missing values in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls.
- Presence of retweeted_status_* actually means the tweet is a retweet not the orginal one.
- Several id columns are int64 or float not string.
- timestamp is string not datetime.
- source is in html tag format not the url itself.

image-predictions.tsv

- Less rows in image-predictions.tsv than the twitter dataset. Therefore, tweets in twitter-archive-enhanced.csv either have no images or are predicted.
- tweet_id is int not string.
- In p1, p2 and p3 column, upper case and lower case are mixed. Dashes and underscores are

mixed too.

retweet count table

- Much less tweets compared to the twitter dataset.
- tweet_id is int not string.

Tidyness

- doggo, floofer, pupper and puppo are just different stages of a dog.
- There are three tables not one table containing all the information.
- Index are not tweet_id.

Cleaning

A copy of the dataset was made for cleaning. Each issue has been defined, coded and tested.

The cleaned data has been stored in twitter_archive_master.csv.