wrangle_report.pdf

INTRODUCTION

In this project, we gather, review, and clean data related to the WeRateDogs Twitter archive, @dog_rates, also known as WeRateDogs.

GOAL

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The final data will include the @dog_rates twitter archive, image predictions, retweet counts, and favorite counts.

REQUIREMENTS

Assess and clean 8 quality issues and 2 tidiness issues according to the rules of tidy data: https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html Note: You only want original ratings (no retweets) that have images.

DATA GATHERING

We will use the requests library and APIs to gather the data:

1. **Twitter archive file:** twitter_archive_enhanced.csv This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
2. **Image predictions**, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. **Twitter API & JSON:** Each tweet's retweet count and favorite ("like") count at minimum, and any additional data. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library.

DATA CLEANING

1. Tidiness issues addressed:
    a. consolidate all dog stages text into one new dog_stage column
    b. set index on all files to "id" or "tweet_id" as appropriate
    c. image_predictions: change column name
    d. merge three separate data sources on 'id'
2. Quality issues addressed:
    a. archive: change tweet_id to type int64
    b. archive: convert timestamp to datetime object
    c. archive: drop unnecessary columns
    d. image_predictions: drop unnecessary columns such as img_num
    e. image_predictions: change tweet_id to type int64
    f. tweepy: change tweet_id to type int64

g. tweepy: drop unnecessary columns
h. Master data file: remove retweets and remove rows without expanded urls

CONCLUSION

Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets. The three data sets have various issues that need to be addressed depending on the final goals of the project. For this project, it was sufficient to address basic quality issues in order to generally analyze the data and create overview data visualizations.