wrangle_report

January 14, 2018

1 Documentation of Data Wrangling

This is an internal report that documents efforts taken in gathering, assessing and cleaning the data used in this project.

1.1 Gathering

The following datasets were provided at the start of the project: `twitter-archive-enhanced.csv`, `image-predictions.tsv`. Other data needed to be downloaded over the internet from the Twitter API. For this task, tweepy - a third-party wrapper to the twitter api was used. Essentially, the file `twitter-archive-enhanced.csv` was loaded into memory and looped over to build a list of `tweet_id` strings. The list contained lists, each containing up to 100 tweet ids. Each list containing 100 tweet ids was sent as a batch argument in a request to the Twitter API using tweepy to avoid rate-limiting.

1.2 Assessment

The dataset twitter-archive-enhanced.csv was opened in Google Sheets to conduct a visual assessment of the data for quality and tidyness issues. One issue was that there were retweets mixed into the dataset which represented a quality issue and would need to be removed. We wanted the dataset only to contain original tweets with images, not retweets. Another quality issue was incorrect or empty dog names. Another issue involved the image predictions dataset – that of there being datapoints in which no dog was detected at all over the three passes. One unique issue was that of there being two dogs in a single datapoint. There were some variable columns representing the dog types (doggo, floofer, pupper, puppo) however, it was determined that these should be maintained and not melted into a single column, as some datapoints legitimately had multiple dog types. However, the data representation of these variables were in a strange format (ex. the value for the column `doggo` was equal to either doggo or None) were not ideal and would be better represented as a boolean value (true/false).

Question: Interestingly, the rating system often includes a numerator that exceeds the denominator value of 10. This was intended by the data creator and reflects the playful spirit of the dataset. However, from a data analysis standpoint, this denominator is meaningless. In light of this, should the rating denominator be removed entirely?

Next, the dataset image-predictions.tsv was opened in Google Sheets for visual assessment. In some cases, the image recognition software failed to detect dogs that were present (false negative), in others it detected dogs when none were present (false positive). It should be noted that the data

creator intentionally included ratings of images of animals other than dogs in the dataset. For example, one image is of a box turtle, another is of a fish, another is of a human wearing a dog mask).

Question: Since the creator of the data purposefully placed such images, it is unclear whether they should be removed from the dataset in keeping with the spirit of the data. Tidiness Question: should a data point be excluded from the dataset if its image prediction yielded a false negative or false positive result?

1.3 Cleaning

The quality and tidyness issues that were identified in the assessment step were then broken down and addressed one by one – a cleaning task created for each. For each cleaning task, a purpose was defined, code implemented and finally tested to ensure the desired result occurred. The exact definitions and code are laid out in `wrangle_act.ipynb`. After the cleaning tasks were executed, the resulting dataset was stored as a new file `twitter_archive_master.csv`.