wrangle_report

October 5, 2020

For this project, I was tasked to gather, assess, and clean data from multiple sources into a final, clean dataset used for analysis. The three sources were as follows: Twitter archive data provided by the project assigner (“archive data”), Twitter Tweet data as parsed via the Tweepy package using the Twitter API (“api data”), and dog images as gathered by a neural network provided by the project assigner (‘image data”). This report will summarize the wrangling efforts completed as part of this project.

Gather

The gathering process differed for each data source. For the archive data, the data was loaded using Pandas via the ‘read_csv’ function. For the api data, a Twitter developer account was created. Using the Tweepy package, the Twitter API was initialized. Then, using the listing of Tweet IDs provided in the archive data, the ‘get_status’ function was used to gather all api data for each Tweet ID. These were then written into a txt file via the ‘json.dump’ function and was loaded using Pandas via the ‘read_json’ function. The image data was loaded using Pandas via the ‘read_csv’ method and a tab ‘’⁀as the specified delimiter.

After the data was gathered from the individual sources, I decided to merge the data into a dataset prior to assessing and cleaning. This was done via Pandas using the ‘merge’ function. I performed an outer merge on the ‘tweet_id’ variable as it was the key that linked the three datasets. This dataset was named “df” and will be referred to as such within this report.

Assess

I assessed the data both manually/visually and programmatically. To assess it manually/visually, I used ‘df.sample(5)’ to view a random sample of my dataset. This allowed me to see areas that need cleaning. To assess the data programmatically, I used ‘df.info()’ to view datatypes and counts of non-nulls within the dataset.

My efforts resulted in the following list of findings, and thus, cleaning tasks:

Quality

1. Timestamp information isn’t in datetime format and contains unnecessary information.

2. Population contains retweets and replies.

3. Retweet count is in float format.

4. Favorite count is in float format and has invalid counts.

5. Various uncessary columns containing many NaN/Null/None data.
6. Not all tweets have an associated image.
7. Various invalid dog names such as "a" or NaN/Null/None.
8. Invalid dog types in the p1/p2/p3 columns.
9. Some denominator values aren't 10.

Tidiness

1. Multiple data store in a columns: entities/extended_entities.
2. Data that should be stored in a single column is stored in multiple: doggo/floofer/pupper/puppo dog types.

Clean

1. Dropped the timestamp column.
2. Filtered for no Retweets/Replies.
3. Set the retweet_count column to an integer type and dropped invalid (<0 values).
4. Set the favorite_count column to an integer type and dropped invalid (<0 values).
5. Dropped unnecessary columns and those with large amounts of N/A, Null, NaN, None, etc.
6. Dropped rows without an image URL.
7. Dropped invalid names and dogs with no names.
8. Dropped p1, p2, p3 data with counts less than 1 (as these were usually not actually valid dog breeds) and manually dropped rows with invalid dog breeds.
9. Dropped rows with invalid (<0) dog denominators.
10. Dropped columns with multiple data in one column.
11. Dropped the dog-tionary columns (see notebook for rationale).

---

jedraynes