wrangle_report

July 14, 2020

1 WeRateDogs - Data Wrangling Summary

The following document summarizes the data wrangling procedure undertaken for the WeRateDogs Twitter archive.

Data Gathering:

Data for this analysis was obtained from three different sources, and stored in three separate Python DataFrames: - WeRateDogs Twitter enhanced data archive. Downloaded directly from Udacity servers. - Data containing neural network generated image predictions. Accessed from Udacity servers and programmatically downloaded. - Data regarding favourite count and retweet count for all Tweets from the WeRateDogs account. Extracted after accessing the entire JSON data downloaded by querying Twitter's API using Python's tweepy library.

Data Assessment and Cleaning:

Data was assessed visually and programmatically, identifying quality and tidiness issues over all DataFrames. Once identified, these were subsequently cleaned appropriately. These two are detailed together for ease of clarity.

Regarding the `twitter_archive` table: - According to the Key Points in the project requirements, all retweet and reply tweet data is irrelevant. Hence, rows containing non NaN values in the **retweeted_status_id**, **retweeted_status_user_id** and **retweeted_status_timestamp** columns are dropped regarding retweets, and all rows containing non NaN values in the **in_reply_to_status_id** and **in_reply_to_user_id** columns are dropped regarding reply tweets. These columns are subsequently redundant and so dropped as well. - The **timestamp** column is modified from string to datetime datatype. - Numerous cases of denominator ratings not equal to 10 for original tweets were identified. Since these were a very small proportion of total entries in the dataset, these were dropped. Subsequently, the denominator rating column was dropped since all values were equal to 10, making it redundant. - Some cases remained with numerator ratings well in excess of 15. Since these were excessively high, they were considered to be inaccurate, and hence dropped from the dataset. - 3 tweets were identified to lack **expanded_urls** data. These were dropped. - Incorrect names for 109 cases of data were identified (lower case, connector words rather than names). These were replaced by 'none'.

Regarding the `api_data` table: - This DataFrame was not cleaned. Alternatively, data for **favourite_count** and **retweet_count** was merged into the `twitter_archive` DataFrame, and its datatype changed from string to integer.

Regarding the `image_predictions` table: - This DataFrame was not cleaned. Alternatively, the most likely predictions for breed based on the highest confidence levels were extracted and merged

into the `twitter_archive` DataFrame by firstly converting **twitter_id** from string to integer. - Rows with incomplete data following this were dropped.

Finally, the combined master DataFrame was reorganised to contain the numerical properties to the left of the DataFrame for easy visualization. This was stored as "twitter_archive_master.csv", whilst the remaining DataFrames were left unsaved as they were not cleaned and not needed for analysis.

[ ]: