Data Wrangling report

Bo Fan

This report described the details in data wrangling including gathering, assessment and cleaning. For data collection, I downloaded the twitter archive data manually, wrote http request to download the image prediction files from Udacity server programmatically, and queried the tweeter API to save all the JSON data into a txt file.

For data assessment, especially in the twitter archive data set, there were quite a lot of issues. Some data types were incorrect, for example, timestamp, dog stages, tweet id, etc. In the name column, 745 name values were missing (None). Some names especially shown in lower cases were meaningless, for example, “a”, ‘the’, ‘very’, etc. In the expanded urls’s column, 59 elements were missing. The rating numerator column had outliers of which the total counts were less than or equal to 2. Some numerator values were inconsistent with the data from the text column where fractional numerators appeared. The rating denominator were not all 10, and a few scores were not consistent with the information from the text column. Some rows contained retweet information. The “&” string in the text column was quite redundant. Columns such as ‘in_reply_to_status_id’, 'in_reply_to_user_id', 'retweeted_status_id were useless. For the additional data set acquired via tweepy, only ‘id’, ‘favorite count’ and ‘retweet count’ were useful, but ‘id’ was confusing. For data tidiness issues, in the tweet archive data set, four columns: doggo, floofer, pupper and puppo were redundant. Some tweet id did not have corresponding images in the image prediction data set. Tweet id appeared 3 times in 3 different sources.

For data cleaning, I solved all the issues. For the name issues, I cleaned the data in 3 cases. When the names were shown in lower case and the text column contained key words such as ‘name’ or ‘name is’, I used the real name extracted from the text column to replace the lower-case names. Otherwise, those names were set as None. I removed the rows where the expanded url values were missing, the ‘retweeted status id’ values were not none, or the rating numerator values were outliers. For the rows where denominators were not 10, I visually checked the text column and replaced the old denominators and numerators. After removing all the outliers, the numerator value set as 5 (it was 13.5 in the text column) was replaced by 13.5. The “&” in the text column were all replaced by “&”. I also filtered the additional data set by keeping “id”, “favorite count” and ‘retweet count’, where ‘id’ was also changed to ‘tweeter id’ to remove ambiguity. Useless columns such as in_reply_to_status_id', 'in_reply_to_user_id' and 'retweeted_status_id’ were removed. The data types of timestamp, tweet id, dog stage, rating numerator and denominator were all set to the correct ones.
For data quality issues, the four columns (‘doggo’, ‘floofer’, ‘pupper’, and ‘puppo’) were combined to one. Redundant column tweet_id was cleaned and three tables were joined to a master table using inner join.