wrangle_report

January 12, 2019

1 Wrangle report

1.0.1 general

The data wrangling started with manually downloading twitter-archive-enhanced.csv' and downloading the 'image-predictions.tsv' file by using the python module requests. Creating the third dataset was achieved by querying the Twitter-API by using the tweepy module. For this to be done I had to create a developer account and gather tokens.

1.0.2 archive dataset

The archive dataset contained some columns which did not have enough values. I decided to drop them because one would not work with those columns. The source column contained html-tags which I removed in order to only save the source. Furthermore the dataset contained four columns used to categorize the type of dog which I reduced to one column providing the exact same amount of information. The most challenging cleaning step was to find rows where multiple categorization were used (doggo and floofer, or doggo and pupper and so on).

1.0.3 image dataset

The image dataset looked very good in terms of quality. But there were three boolean columns indicating wether or not there was a dog shown in the picture. I combined these three columns into one which only indicated of the picture shows dogs only or at least a dog. I decided to drop rows where there was no dog. I drop one column which seemed to be of no use (img_num).

1.0.4 tweet dataset

The tweet dataset was the widest one in terms of column count. It was also the one where i dropped the most columns. This was due to missing values (3 columns), duplicate information (2 columns), non relevant information (all rows of a column with the same value - 6) or irrelevant information (5 columns).

1.0.5 final words

Most of the cleaning involved dropping columns which were of no use or only had a low number of values. In the end I merged all three dataset together using tweet_id or id as a key. The resulting dataset now contains all relevant information and can be used for analysis. But even at this stage

of the master dataset one could dive deeper and clean further. The question is just how useful further cleaning would be.