Wrangle report

The wrangling processes in this project include data gathering, data assessing, data cleaning and data storing.

Data gathering

There are 3 main files needed in this project:

1) Twitter archive provided by Udacity
   This file is download from Udacity hosted storage manually and move to project folder. This file is in .csv format.
2) Image prediction result provided by Udacity
   This file is download programmatically using python request library from Udacity hosted storage. This file is in .tsv format.
3) Twitter post data retrieved using Twitter API
   This file is requested using Twitter API and dump it into a .txt file

Data Assessing

Data were assessed using different method include dataframe in jupyter notebook, Ms Excel, and Visual code.

Data Cleaning

[twitter-archive-enhanced.csv]

Quality

- timestamp should be in datetime type
  Converting datatype using pandas.astype function
- retweeted_status_timestamp should be in datetime type
  Converting datatype using pandas.astype function
- retweeted_status_id should be in int type
  Converting datatype using pandas.astype function
- retweeted_status_user_id should be in int type
  Converting datatype using pandas.astype function
- remove the retweet post
  Remove all rows with retweet_status_id exist
- expended_urls should be in list type
  Split the url text by using comma

- some inconsistent values for rating numerator and rating denominator
  Retrieve the rating from text and check for the different manually

Tidiness

- the link in text column should be separated
  Extract the post url from text and put in a new column
- the stage should be melted into a single column
  Melt 4 of the stage columns into a single column

[images-prediction.tsv]

Quality

- breed prediction should be all in lower case
  Using pandas.str.lower() function to change all text into lower case
- p1, p2 and p3 should be in categorical type
  Convert datatype using pandas.astype function

Tidiness

- aggregate classifier models result
  Pick the prediction with highest confidence and is predicted as dog
- remove unuse columns
- merge it with twitter-archive-enhanced
  Using inner join

[twitter-json.txt]

Tidiness

- remove unuse columns
- merge it with twitter-archive-enhanced
  Using inner join

Data Storing

The clean data in store in a csv file named twitter-archive-master.csv,