UDACITY

NANODEGREE DATA SCIENCE II

DATA WRANGLING PROJECT

This document briefly describes the data wrangling steps: gather, assess, and clean. All of them were executed during the completion of the Project: “Data Wrangling”, from the Nano Degree Data Science II course.

Phase 1 - Gathering Data

- Import the udacity provided file (‘twitter-archive-enhanced.csv’) into a dataframe to begin analysis
- Create ‘downloads’ directory to programmatically download image predictions .tsv file
- Download and import image predictions programmatically to a dataframe
- Authenticate to tweeter using the APP and Keys using the ‘tweepy’ library.
- Return number of retweets and favorites based on the tweet id's read from the udacity provided file(‘twitter-archive-enhanced.csv’) and save it
- Save tweet_id, favorite_count and retweet_count to a text file named tweet_json.txt
- Save tweet_id and exception message to a text file named tweet_json_errors.txt

Phase 2 – Assess Data

Identified 9 Quality issues with the data using visual analysis, working with excel and the jupyter notebook.

- Several incorrect recognized dog names,set all to null.
- Convert timestamp column to datetime type.
- Convert columns (in_reply_to_status_id, in_reply_to_user_id) to integer.
- Remove any rows that contain data in columns (retweeted_status_id, retweeted_status_user_id e retweeted_status_timestamp), since we do not want retweeted info.
- Remove 'expanded_urls' rows with missing records, '2297'.
- Rextract ratings that have numerator with a dot('.').
- Convert rating columns to numeric
- Remove double dog stage classifications
- Extract and keep only relevant information from source column

Identified **4** Tidiness issues

- Several columns (doggo, floofer, pupper, puppo) could be converted to one categorical "dog_stage".
- Columns (rating_numerator and rating_denominator) could be one column 'rating', calculated from these 2.
- Merge dataframe df_tweet with clean dataframe df_wrdclean
- Merge dataframe df_imgpred with clean dataframe df_wrdclean

Document Category: Internal

Phase 3 – Clean Data

Clean data based on the observations made during the assessment phase

1. Quality issues cleaning process

    - Copy original dataframe to avoid the need of rework in case of mistakes during the cleaning process.
    - Removed any dog name starting with a lower-case letter using regular expression selection method
    - Removed rows with retweets
    - Removed columns referring to retweets
    - Removed tweets without images
    - Converted timestamp column to datetime datatype
    - Extracted ‘rating_numerator’ again taking dots before the '/' into consideration
        - https://regex101.com/r/7UENUr/1
    - Converted numerator and denominator columns to numeric
    - Removed double "dog stage" classifications
    - Extracted and kept only relevant information from source column

2. Tidiness issues cleaning process

    - Several columns (doggo, floofer, pupper, puppo) were converted to one categorical "dog_stage" column.
    - Columns (rating_numerator and rating_denominator) were deleted and one column named 'rating’ was calculated from these 2.
    - Merged DF_TWEET to DF_WRDCLEAN dataframe.
    - Merged DF_IMGPRED to DF_WRDCLEAN dataframe.

Warm regards,

Rodrigo Neves de Barros
Data Scientist

Document Category: **Internal**