WRANGLE REPORT

Step-by-step Wrangle process

Hocepied, Robin
Udacity

Introduction

The main objective of this report is to summarize the efforts deployed to wrangle the data of the WeRateDogs Twitter. This twitter account provides ratings and funny comments on pictures of dogs.

The project goals are:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Storing
5. Analyzing and visualizing the cleaned data (doc Act Report)
6. Reporting via 2 documents (+ doc Act Report)

1. Gather Data :

The first step towards this analysis covered the gathering of the data coming from the WeRateDogs Twitter account and of two other files provided by Udacity:

1. The **WeRateDogs Twitter archive**. The twitter_archive_enhanced.csv file was provided to Udacity practitioners. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

2. The **Tweet image predictions**, this file contains the predictions that were identified from the image url based on the neural network.

3. **Twitter API and Python's Tweepy library** (JSON file) to gather each tweet's retweet count and favorite ("like") count.

2. Assessing Data:

After gathering the data, the next step consists on assessing the data, visually and programmatically. This part covers my own 4 principles, completeness, validity, accuracy, consistency and tidiness to assess the data. You will find below the set of issues found in the data, they are allocated in two sections, Tidiness and Quality.

1. **Tidiness:**

   a. **General:** Three data frames to be analyzed
   b. **Twitter Archive:** The last four columns all relate to the same variable

2. **Quality:**

    a. **Twitter Archive:** A lot of useless columns, such as ('in_reply_to_status_id', 'in_reply_to_user_id' and 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp')
    b. **Twitter Archive:** Wrong datatypes for timestamp and tweet id
    c. **Twitter Archive:** Missing data in the expanded url column
    d. **Twitter Archive:** In the rating denominator column, some values are not in scope as they should respect certain criteria (denominator = 10)
    e. **Twitter Archive:** In the rating numerator, the values should be between 10 and 15
    f. **Twitter Archive:** The column names has 109 mislabeled
    g. **Image Predictions:** The prediction dog breeds involves both uppercase and lowercase issues

3. Cleaning Data:

After the assessment I used the following method to clean the data:

1. Define
2. Code
3. Test

The list of steps were as follow:

- Merge the clean versions of archive, images, and json data frames
- Create one column for the various dog types: doggo, floofer, pupper, puppo
- Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
- Remove columns no longer needed.
- Change tweet_id from an integer to a string.
- Change the timestamp to correct datetime format.
- Correct naming issues and standardize dog ratings.

4. Storing Data:

I stored the data in a csv file to make sure I saved the file before moving to the analysis process