Luiz Iurk

Iurk.lf@gmail.com

Report

On Data Wrangling Steps of WeRateDogs Twitter Data

# Wrangle Report

The dataset wrangled in the project is the tweet archive of Twitter user @dog\_rates, also known as WeRateDogs. This is a Twitter account that rates people's dogs with humorous comments about their dogs.

The goal of this project is :

* Wrangling the twitter data through the following processes:
  + Gathering data
  + Assessing data
  + Cleaning data
* Storing, analyzing, and visualizing the wrangled data
* Reporting on the data wrangling efforts and data analyses and visualizations

## Gathering Data

Three sets of data were provided to start this report. Each set of data was provided from a different sources as following:

* **WeRateDogs Twitter Archive**: This set of data was previously provided by Udacity in a .csv file under the name twitter\_archive\_enhanced.csv. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) of 2356 tweets as they stood on August 1, 2017.
* **Image Predictions**:This set was stored as a .tsv file and was stored at Udacity’s online storage folder under the name image-predictions.tsv. The file contains the result of each tweet of the machine learning algorithm that returns the dog breed based on one of the 4 possible picture of each tweet.
* **Additional Tweet Information**:This set of data was supposed to be fetched through the Twitter API Tweepy and stored in a .json format. However, due to a limitation of access it was not possible to fetch the data this way, so it was used the already available file in the Udacity storage library called tweet-json.txt. From this extraction, it was gathered each tweet's retweet count and favorite ("like") count at minimum, and additional data that could be used in statistics.

## Assessing Data

After the data was gathered, the data was assessed in two ways : Visual and Programmatically.

In both cases the objective was to evaluate quality and tidiness issues.

Quality Issues

* Completeness:
  + Some records had the field **expanded\_urls** empty in the WeRateDogs extraction;
* Validity:
  + dog names: Typo in dog names;
    - Remark: In this field, it was identified that when the word started with lower case letters, it was not a proper name, but a preposition, an article or even an adverb. This was proven programmatically in the python report.
  + Data types: Fields like **in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id** and **tweet\_id** should be as string, not as int;
  + Discard retweets: The columns **retweeted\_status\_id, retweeted\_status\_user\_id** and **retweeted\_status\_timestamp** should be empty;
  + Invalid Predictions: Since what is being evaluated are dogs, predictions of animals/things other than dogs should be discarded;
* Accuracy:
  + Rating Numerators with big values: Some numerators are decimals and were wrongly extracted
  + Timestamp with incorrect data type: switch it to date and time type
  + Issues on the programmatic rating extraction: Due to the similarity of the rating with dates or other strings that take forward slash (‘/’) , the free text extraction interpretated this mention as rating. In these cases, they must be manually corrected when identified.
* Consistency:
  + rating\_denominator should be a standard 10, but there are a multitude of other values
  + **p1, p2** and **p3** not following a case logic: Sometimes are lower, sometimes upper case. They should be standardized as Title case and replacing ‘\_’ by space

Tidiness Issues

* the column source showing in html tags shows 3 distinct values in the same field. It should be divided by url\_ref, rel and tweet\_source;
* Columns **doggo, floofer, pupper, puppo** should be turn into a single column called dog\_stage;
* Identify and discard unnecessary columns;
* Merge dataframes for analytics purposes

## Cleaning Data

The cleaning process happened, to each of the mentioned issues, following these three steps:

Define: Describe with action verbs exactly what will be done in the section;

Code: Shows all the python code lines used to achieve the definition;

Test: Shows that the defined action works.

## Storing Data

After the data was cleaned, the unnecessary fields were discarded and the dataframes were merged into one single dataframe.

After that, this dataframe was stored into a .csv file called “twitter\_archive\_master.csv”

## Analyzing and Visualizing Data

The following insights were taken with the processed data:

1. Total Number of tweets over time
2. Correlation between tweets and average of followers
3. Top 10 most predicted breeds
4. Top ten 10 most common dog names