Wrangle and Analyze Data with #WeRateADog

SEBASTIEN HANICOTTE

MAY 2020

Introduction

As part of the Data Analyst NanoDegree (D.A.N.D), the project is here to Wrangle datas and Analyze those datas coming from Tweets of the account WeRateDogs. So, will be using Twitter datas, connexion to the API to control datas, and informations coming from Picture Analysis.

Datas

3 Different sources are used in this project :

- Twitter Archive
  An archive of more than 5K tweets is given as entry datas. Those are all original tweets, replys and retweets from the WeRateDogs account for a period of time.

- Twitter Image Prediction
  This is a CSV of Image Prediction depicting Dogs. This CSV is to download from a specific URI `https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv`

- Twitter API
  In order to grab additional datas, I'll be using the Twitter API. This will allow me to complete, validate all datas that were given through the Twitter Archive.

Gathering Data

As requested, I gathered datas from all 3 sources :

- TwitterArchive was loaded from the CSV retrieved manually from the Udacity's servers
- ImagePrediction was loaded from the TSV retrieved from the URI programmatically.
- TwitterAPI was used to retrieve complementary information about all the tweets from WeRateDog account using Tweety library.

Assessing Data & Cleaning

I began the assesment on the TwitterArchive by displaying global informations about the datas (info to retrieve type of columns, describe to retrieve min/max/mean values from numeric datas, shape).

This allowed me to discover several quality and tidiness issues :

Some columns needed a changed of type like `timestamp` to be dealt as timestamp, tweet_id as string (instead of int).

Requirement was to deal only with original tweet, therefore all rows containing non-null values in `retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_status_id, in_reply_to_user_id` were removed either was the columns.

The 4 dogs category (`doggo`, `floofer`, `pupper`, `puppo`) were summarize into one column “stage”.

The `rating_numerator` and `rating_denominator` were analyzed, cleaned from off charts datas and stored in a single column named `rating`.

All dogs “wrongly named” were removed from the dataset.

Few tweets were missing `expanded_urls`, all those tweets were removed from the dataset as was the column.

The Image Prediction datas were also analyzed and dealt with.

For each image came 3 predictions containing respectively “Is it a dog ?”, “Dog race”, “Confidence”. All those datas were summarize into 2 columns “Prediction” and “Confidence”.

The tweet_id column type was changed to string instead of integer as was the column creation_date changed to datetime.

Finally, all datas coming from the 3 sources were merged together using the common column `tweet_id`.

The resulting dataset was saved into a **twitter_archive_master.csv** file.