Data Wrangling: WeRateDogs Twitter Data

Data

1. File on hand: twitter-archive-enhanced.csv

Data : WeRateDogs Twitter archive basic data

2. File from the Internet: image_predictions.tsv

Data: Top 3 predictions based on Twitter ID and associated photo

Location: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Method: requests

3. Twitter API data : tweet_json .txt (created from API)

Data used: Tweet ID, retweet count, favourite count

Method: Use Tweepy to query Twitter's API using Tweet IDs in twitter archive data and save JSON in a text file

Assess and Clean

Visual then programmatic assessment with Python, Pandas and NumPy

Quality issues relate to content: Completeness, Validity, Accuracy and Consistency

Q1. Different number of records:

- Twitter Archive Data: 2356,
- Image Prediction Data: 2075,
- Twitter API Data: 2337

Q2. Mostly missing data in retweeted_status variable and in_reply_to columns

Resolution: Issues Q1 and Q2. Make the datasets consistent. Only include Original Tweets with an image. Exclude records in Twitter Archive Data with *retweeted_status* and *in_reply _to_status*, deleting any related '*retweeted*' and '*in_reply_to*' columns. This in now our Master Dataset

Exclude any records in the Master Dataset that that don't have 1. Corresponding *tweet_id* in Twitter API Data and 2. Corresponding *tweet_id* in Image Prediction Data.

Outcome*:* Master dataset with only original tweets with an image

Q3. Missing data and multiple urls for *extended_urls*

Resolution: Repopulate records with missing or invalid *expanded_urls* using the tweet_json.txt

Q4. *source* column is difficult to read and contains extra/irrelevent information

Resolution: Extract the source, excluding any urls or html

Outcome: Succinct sources of tweets: Twitter for iPhone, Twitter Web Client, TweetDeck

Q5. Erroneous datatypes : *tweet_id* and *timestamp*

Resolution: Convert *tweet_id* data type in Master Data and Image Prediction Data from integer to string. *tweet_id* is a unique identifier and will not be manipulated with maths.

Convert *timestamp* datatype in Master Data from string to datetime in order to perform time related analysis.

Outcome: Datatypes: *tweet_id* - string. *Timestamp* - datetime

Q6. *rating_denominator* > 10

Q7 Extra large *rating_numerator* (> 14)

Resolution: Extract rating denominator and numerator from text column in Master Data using RegEx to ensure correct rating data. Recalibrate rating numerator to its ratio to 10 for denominators greater than 10. Exclude records with extra large numerators after visual inspection. Delete *rating_denominator* column

Outcome: 1 *rating* column. Extra large ratings either recalibrate to ratio to 10 or excluded

Q7. Missing and non- *name* (‘a’, ‘an’,’the). Some names not picked up with RegEx

Resolution: Extract names using RegEx for common introductions, “This is”, “Meet”, “Say hello to”, “Here is”, deleting any non-names. Individuallly search for undetected names in *text* and populate associated *name* column

Outcome: 1378 names

Q8. Capitalized and lower case first letter in *prediction* column of Image prediction data

Resolution: Using dog breed list downloaded with wptools, map predictions that are dogs

Outcome: Standardized list of dog breeds, all starting with capital letter

Tidiness issues relate to structure.

T1. Twitter Archive Data: Stage of dog (4 columns: *doggo, floofer, pupper, puppo*) is one variable so should be one column. 10 cases of single tweets with multiple dog stages

Resolution: Merge 4 columns into 1 *dog_stage* columns, visually inspecting any records with multiple dog stages and either classifying as ‘multiple’ or adjusting individually. Delete the 4 columns

Outcome: only one *dog_stage* column in Master Dataset

T2. *retweet_count* and *favourite_count* in df_tweets table should be part of Master

Resolution: join *retweet_count* and *favourite_count* from Twitter API data to Master Dataset, based on *tweet_id*

Outcome: Only 2 datasets: master and image_predictions

T3. Twitter Archive Data: Along with text, *text* column contains a hyperlink, which is abbreviated *expanded_url*

Resolution: Delete hyperlink in *text* column

Outcome: Only text of Tweet in *text* column

T4. Image Prediction Data: column headers *p1*,*p2* and *p3* are values, not variable names. The associated *_conf* and *_dog* columns should be 2 columns.

Resolution: Combine *p1, p2,* and *p3* and associated *p _conf* and *p _dog* columns into 3 columns, identifying each prediction in the prediction number column (1,2,3 to coincide with p1, p2, p3)

Outcome: Image_predictions dataset. Columns: *tweet_id, jpg_url, img_num, prediction, confidence, is_dog, prediction number*

A bulk of the assessment and cleaning involved reading the Twitter text. Thus being able somehow to assess programatically as opposed to visually the text may be more time efficient and less prone to human error.

Cleaned Datasets:

1. **master.csv.**

   Columns: tweet_id, timestamp, source, text, rating, expanded_url, dog stage, retweet_count, favourite_count, name

2. **image_predictions.csv**

   Columns: tweet_id, jpg_url, img_num, prediction, confidence, is_dog, prediction_number

Supporting data

1. twitter-archive-enhanced.csv
2. image_predictions.tsv
3. tweet_json.txt
4. dog_breeds.csv