wrangle_report

April 9, 2019

0.1 WeRateDogs Data Wrangling Report

0.2 Introduction:

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. I performed the data wrangling three steps (gathering, assessing and cleaning) on the WeRateDogs data to get clean data set for further analysis.

0.3 1- Gathering

For this project we will gather the data from three resources :

- Enhanced Twitter Archive CSV file : that already prepped and downloaded manually from Udacity.

- The tweet image predictions : hosted on Udacity's servers and downloaded programmatically using the Requests library.

- Twitter API : Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library.

0.4 2- Assessing

After I assess the three data file that I gather I found the below issues:

1.

0.4.1 tweet_arch_df:

-

0.4.2 Quality Issues:

- in_reply_to_user_id has many null.
- in_reply_to_status_id has many null .
- in_reply_to_user_id , in_reply_to_status_id , retweeted_status_id and retweeted_status_user_id are float not int.
- Wrong dog names (a , an , the , such).

- Dogs names with symbol Gšrdşn , Devşn , Ralphľ , Oliviľr , Amľlie , Flvio and Frűnq.
- There are retweet and replays in the tweet_arch_df that may result in misleading and redundant rating for same tweet indirectly.
- have two dog stage for same dog.
- expend url in tweet_arch_df is duplicated for jpg_url in img_prediction_df.
- duplicated expend url in the same recored.
- duplicated expend url in tweet_arch_df.
- source column url can be categorizes like iPhone Twitter, TweetDeck , Twitter Web Client and Vine.
- source has html code with the url.
- Timestamp is string not date time.
- huge number in rating_numerator & rating_denominator, since they are humour tweets some of them are right ratings and other were wrong and captured from other numbers that mentioned in the tweets but not for rating .
- None value insted of NaN.
- less than 621 with dog stage values.
- Undescriptive column names expend url

•

0.4.3 Tidiness Issues:

- Many column for dog stages insted of one .
- Timestamp has date with time

2.

0.4.4 img_prediction_df:

•

0.4.5 Quality Issues:

- duplicated jpg_url.
- Dog breed name with upper first letter and other without.
- Dog breed name with '_' .
- Undescriptive column names .
- Records with False prediction for all algorithms.

•

0.4.6 Tidiness Issues:

- Image algorithms prediction column the dog breed with false results presented.

3.

0.4.7 tweet_df:

-

  #### 0.4.8 Quality Issues:

  - Html code in multiple column like source , entities.
  - Undescriptive column names id .

-

  #### 0.4.9 Tidiness Issues:

  - repeated column like id and id_str.
  - Most columns in tweet_df are duplicate of tweet_arch_df columns.

0.5 3 - Cleaning:

During the assessment I found many issue and in druring cleanning step I cleaned the below issues:

During the assessment I found many issue and in this cleanning section I will clean the below issues:

- ### Quality Issues:
  - Undescriptive column name id in tweet_df .
  - Undescriptive column names expend url in tweet_arch_df.
  - There are retweet and replays in the tweet_arch_df that may result in misleading and redundant rating for same tweet. so, original tweet will be keep it for accurate investigation result.
  - Records with False prediction for all algorithms.
  - Wrong dog names (a , an , the , such).
  - have two dog stage for same dog.
  - source column url has html code and can be categorizes like iPhone Twitter, TweetDeck , Twitter Web Client and Vine.
  - Timestamp is string not date time.
  - None value insted of NaN.
  - huge number in rating_numerator & rating_denominator, since they are humour tweets some of them are right ratings and other were wrong and captured from other numbers that mentioned in the tweets but not for rating .
  - Dog breed name with upper first letter and other without in img_prediction_df .
  - Dog breed name with ' _ ' in img_prediction_df .
- ### Tidiness Issues:
  - Many column for dog stages insted of one .
  - timestamp split timestamp into two column date and time in tweet_arch_df.
  - Image algorithms prediction columns for the dog breed with false results presented.
  - Most columns in tweet_df are duplicate of tweet_arch_df columns.

- columns that repeated and that don't provide useful information for our we rate dog tweet exploration like in_reply_to_user_id , in_reply_to_status_id , retweeted_status_id and retweeted_status_user_id and jpg_url.