wrangle_report

March 16, 2021

1 wrangle_report

1.1 introdaction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

1.2 Gathering Data

- I will gathrer first twitter archive Data with file (twitter_archive_enhanced.csv)
- second I use requst library to get data from this url (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)to get (image_predictions.tsv) image prediction data
- and use Twiiter API with My account on Twitter developer use tweepy library and get ## twitter API my reference on stack over flow : https://stackoverflow.com/questions/47612822/how-to-create-pandas-dataframe-from-twitter-search-api

1.3 Assessing data

1.3.1 Quality

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers not float but i think it useless data because it 78 , 181 non null data in 2356 rows too much missing data
- timestamp should be date not string or object and retweeted_status_timestamp too but it has too much missing values
- source is diffcult to read
- rating_numerator and rating_denominator have invalid data like 0 or numbers less than 10
- tweet id should be object
- p1,p2,p3 are not lower or uppercase
- img_num column does not contain new data i think it useless

- should keep Original tweet
- url has some invalid data like (0 , u , e , y , n , t , elc)

1.3.2 Tidiness

- should merge all data frame togather
- the prediction should be in one column called (dog_breed) and prediction confidence too in one column called (pred_confidence)
- Create one column for dog types: (doggo, floofer, pupper, puppo)

2 Clean Data

- i drop missing values like (n_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,elc)
- i drop the columns like soruse , name , url_expend , img_num becouse i think it useless data
- i convert some data to it data type like (timestamp , id_tweet)
- i grouped dog stage data in one column called dog_age
- i bulid funcation called dogbread to compere with image prediction data that based on highist confidence and if ture or false and collected this data in two column (dog_breed,pred_confidence)
- the rating_numerator and rating_denominator , i bulid data frame from (text) to know the valid data i use refcance in this:(https://docs.python.org/3/library/re.htmlű) use libary re
- url has some invalid data like (0 , u , e , y , n , t , elc) i bulid list from wrong urls and drop it
- i drop retweet because it useless

```
In [ ]:
```