Twitter WeRateDogs Project

1.[Twitter WeRateDogs Project]

1.1 [Gathering]

I start with gathering data from 3 different resources and 3 different formats.

1.1.1 [Gathering - twitter-archive-enhanced]

The first resource was twitter archived enhanced file which is CSV file and I used its data as df_1 dataset.

1.1.2 [Gathering - image-predictions]

Then I imported requests to download the file from URL given to me and the format was TSV so I used method pd. read_csv('image-predictions.tsv', sep='\t')
And the dataset as df_2 .

1.1.3 [Gathering - tweepy.API]

I read tweet-json.txt file- which I downloaded from resources – line by line and used its dataset as df_3.

1.2 [Assess]

After Visual and programmatic assessment.

1.2.1 [Quality]

'df_1' dataset :
[1] (in_reply_to_status_id ,in_reply_to_user_id) not null values == validity issue.
[2] (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) not null values == validity issue
[3] timestamp column isn't date time.
[4] extract source of tweet from source column and categorize it.
[5] text column contain urls.
[6] expanded_urls have NaN values.
[7] some expanded_urls contain duplicated urls in the same value.
[8] some rating_denominator value == accuracy issue.
[9] some rating_numerator value not == validity issue.
[10] doggo,floofer,pupper,puppo columns has 'none' for missing values.
[11] some name column values == validity issue.

'df_2' dataset :
[12] img_num column == consistency .
[13] df_2 should contain only the highest prediction confident of breed and the breed name instead of 3 predictions.

1.2.2 [Tidiness]

[1] 'df_1' table : doggo,floofer,pupper,puppo are variable not columns.
[2] All should be in one dataset.

1.3 Clean:

Copied all the datasets : 'df_1' to df1_clean , 'df_2' to df2_clean , 'df_3' to df3_clean .

Define	Method
'df_1' dataset :(in_reply_to_status_id ,in_reply_to_user_id) not_null values == validity issue.	null values are the required rows so I used is_null method to drop not_null rows.
df_1' dataset :(retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) not_null values == validity issue.	null values are the required rows so I used is_null method to drop not_null rows.
columns ['retweeted_status_id', 'retweeted_status_user_id' ,'retweeted_status_timestamp','in_reply_to_status_id', 'in_reply_to_user_id'] not required	drop columns ['retweeted_status_id', 'retweeted_status_user_id' ,'retweeted_status_timestamp', 'in_reply_to_status_id', 'in_reply_to_user_id']
'df_1' dataset : timestamp column is string instead of date time.	use to_datetime method
'df_1' dataset : extract source of tweet from source column and categorize it.	extract unique sources extract unique values string between '>' , '<' from column as source.
'df_1' dataset : text column contain urls.	replace url with space.
df_1' dataset : expanded_urls have NaN values.	not_null values are the required rows so I used not_null method to drop null rows.
'df_1' dataset : some expanded_urls contain duplicated urls in the same value.	divid dataset to 2 datasets by expanded_urls depend on contain ',' or not extract one url from expanded_urls column in dataset contain ',' . concat 2 new datasets.
'df_1' dataset : some rating_denominator value == accuracy issue.	rating_denominator value should be 10 acc. to unique classification , so I will drop values not equal 10 .
'df_1' dataset : some rating_numerator value not == validity issue.	rating_numerators less than or equal 5 most of them are not dog ratings , I will drop them . rating_numerators more than 100 has consistency issue , I will drop them . some rating_numerators between 5 and 100 are flots . I will extract the rating_numerator again from text and convert it to float.
'df_1' dataset : doggo,floofer,pupper,puppo columns have 'none' for missing values.	replace none by "" , combine columns doggo,floofer,pupper,puppo as dog_stage.

	drop columns doggo,floofer,pupper,puppo , replace "" by NaN .
'df_1' dataset : some name column values == validity issue.	## replace 'a' , 'an' in name columns by pattern in text .
df_2 dataset : img_num column == consistency .	## drop img_num column
df_2 dataset : df_2 should contain only the highest prediction confident of breed and the breed name instead of 3 predictions.	## choose the maximum p_conf which will be the breed , use conditional method. ## drop p and conf columns not required.
df_1 , df_2 ,df_3 datasets : all should be in one dataset.	merge 3 datasets to twitter_master.CSV

1.4 [Saving master sheet]

Save merged dataset as twitter_master.CSV

1.5 Visualization

1.5.1 Tweet Count / year.

I used tweet_id count and year from timestamp column to get tweet frequency / year and the highest frequent year of tweets was 2016.

1.5.2 Tweet frequency / Source

I used tweet_id count and source column to get tweet frequency / Source the highest frequent source for tweets was twitter for iphone.

1.5.3 Beloved dog_stage from likes(favourite_count)

I used dog_stage column and sum of favorite_count column to get beloved dog_stage from likes(favourite_count) ,The most beloved dog stage is pupper then doggo depend on favourite_count .

1.5.4 Beloved dog_stage from text_rating

I used dog_stage column and mean of rating_numerator column to get beloved dog_stage from text_rating , The most beloved dog stage is doggo-puppo depend on tweet text .

1.5.5 Beloved dog name from likes(favourite_count)

I used name column and sum of favorite_count column to get beloved dog name from likes(favourite_count) ,The most beloved dog name is Aja then Albus .

Many Thanks :)

Hussein Ellabny