wrangle_report

August 2, 2020

0.1 Wrangle report for WeRateDogs

0.1.1 By Assemgul Kaiyrzhan

In this project, we studied the main section - Data Wrangling. In the project, we examined tweets from the WeRateDogs account, used 3 different sources and methods of data gather. So we use 3 main steps for Data Wrangling:

1) Gather

2) Assess

3) Clean

And after that we Anamyze and make visulization for data wrangling process, please refer to wrangle_act file for steps and codes.

0.2 Gather

1) The WeRateDogs Twitter archive. We have file twitter_archive_enhanced.csv whiche we manually added to our workspace.

2) The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3) Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count

0.3 Assess

Assessing your data is the second step in data wrangling. When assessing, you're like a detective at work, inspecting your dataset for two things: data quality issues (i.e. content issues) and lack of tidiness (i.e. structural issues). So for WeRateDogs we have 8 quality issues and 2 tidiness

0.3.1 Tidiness

1) doggo, floofer, pupper, puppo need make like a one column

2) 3 dataframe join to 1 dataframe

0.3.2 Quality

1) Delete some column which not needed for analysis like in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and etc.

2) Have name "None", "a", "an",which start in Lowercase

3) Change Data Type in Tweet_id, timestamp

4) rating_numerator and rating_denominator have incorrected ratings

5) rating_numerator and rating_denominator change data type

6) In Source column make one format link without html code

7) expanded_urls 2297 non-null

8) dog_rating should be integers, not floats

0.4 Clean

In this step i clean all quality and Tidiness issues which we identify in Assess process
As a result, it was analyzed, separate analytics, which you can see in act_report

```
In [ ]:
```