wrangle_report

December 3, 2020

this project was very interesting, I faced many problems in it and I learnd a lot from the search to solve these.

I used the three resources(twitter-archive-enhanced, data from twitter API, image predictions) in my project, I couldn't make a twitter developer account so I used the json file that exists in the resourses.

I used the url given in the details and using requests package, I got the image_predictions data and stored it an external file as tsv file then read it via `pandas read_csv` function.

after gathering the three datasets using what I learned from the course, I assess them and I found these issues.

quality issues

- in `twitter_archive` dataset, `tweet_id` data type is int instead of str.
- there are many missing values in `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, and `retweeted_status_timestamp`columns in `twitter_archive` dataset, we can remove these columns.
- `name` column in `twitter_archive` has strange names like(a, an, the, o)
- `timestamp` column at `twitter_archive` dataset data type is object instead of datetime.
- `text` column at `twitter_archive` dataset has a link to the tweet.
- `source` column in `twitter_archive` dataset has html tag, we can extract the link from it.
- `expanded_urls` column in `twitter_archive` has null values.
- in `image_predictions` dataset, `tweet_id` data type is int instead of str.
- `rating_numerator` column has outliers, and it's better to be float.
- `rating_denominator` column has values not 10, I want all values equal 10 to drop this column.
- remove the retweeted twitter

tidiness issues

- `doggo`,`floofer`,`pupper`,`puppo` columns should be one column `dog_kind`.
- the three datasets should merged to be only one.
- just 3 columns needed in `tweet_info` dataset (id_str,favorite_count,retweet_count)
- create new colmun (`dog_species`) to store dog species that is the highest probability to be correct.

then, I cleaned these issues one by one using the wonderful python libraries like pandas and finally I saved the clean dataset and wrote my insights.

in this project, Stack Overflow site is my first resource to solve the problems I faced.

In the end I would like to thank you very much for this wonderful scholarship, and for giving me an opportunity to work on this project that helped me practice what I have learned.

thanks.

In [ ]: