Untitled

December 8, 2020

1 1-Introduction

This Wrangle Report is a part of a Data Science Course Project offered by Udacity. The project aims to gather data from Twitter and combine it with a three data frame to create analysis about the tweets and the predicted dog's breed.

2 2- Gathering Data

1- we have three files for data gathering 2- there ara different methods for each file to deal with this files

3 Three data sources

1- Csv file---Download this file manually 2- link data -----using the Requests library to get the data and import it in .Tsv file 3- API twitter---Scrape data from an API and storing it to the file tweet_json.txt

4 3- Assessing data

After gathering the required data, and investigating it visually then programmatically I came up with the following issues with it:-

4.0.1 twitter_archive

```
1- Quality  -Validity-Visual    name    Invalid names or non-standard names.
2- Tidiness -Visual-source  HTML tags, URL, and content in a single column.
3-Tidiness-Programmatic-doggo, floofer, pupper, and puppo   This is a categorical variable, and
4-Tidiness-Programmatic-text--There is two information in a single column. Split the text from t
5-Quality-Validity-Programmatic-timestamp----Convert to date.
6-Quality-Accuracy-Programmatic-retweeted_status_id The same dog could be recorded twice or more
7-Quality-Accuracy-Programmatic in_reply_to_status_id   The same dog could be recorded twice or
--------------------------------------------------------------------------------
```

4.0.2 image-predictions

```
1-Quality-Consistency-Visual p1, p2, and p3 Dogs breed has no standard. Capital letter or  lower
  2-Quality-Validity-Programmatic-jpg_url--It has duplicated images and consequently double entr
```

###twitter_archive_master 1-Tidiness-Programmatic -Merging these two tables (twitter_archive and image-predictions) into one. 2-Quality Validity Programmatic "many columns" Remove in_reply_to_status_id, in_reply_to_user_id, retweeted_status_timestamp, retweeted_status_id, and retweeted_status_user_id.

5 4-Data Cleaning

solve previous issues 1- I used my knowledge of python 2- searching over the internet for references and possible guidance to the best of my knowledge. Most of the issues involving non-usual values to rating_numerator and rating_denominator were solved using check the image that take unusual rating and drop the record that contain data not related to gogs. Overall, I learned a lot about how to use python effectively and efficiently to clean data and store it Finally, I have solved the tidiness issues combining the tables twitter_archive_enhanced.csv and image_predictions.tsv in one called twitter_archive_master.csv. I have also merged 4 columns (doggo, pupper, puppo, and floofer) into one column. once the data was ready I analyzed it using visualizations as document in act_report.htm

6 5-Conclusions

I have documented issues and the twitter_archive_master.csv file is the final file version with a minored number of issues, and ready for a Data Analysis. This file has 1993 observations and 22 features.

In [ ]: