Data Wrangling Repor

Introduction

This project 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. The dataset that you will be wrangling is the tweet archive and image predictions.

Wrangling

Gathering

1. twitter_archive_data: The WeRateDogs Twitter archive, which is provided in the course. The data is in CSV format named, ‘twitter-archive-enhanced.csv’. This data is loaded in Pandas data frame.

2. image_data: The tweet image predictions, i.e., what breed of dog (or other objects, animal, etc.) is present in each tweet according to a neural network. The data file 'image_predictions.tsv' is hosted on servers and data is downloaded and written to the file using the request library in the notebook.

3. tweet_data: 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 details like retweet count, favourite count, etc in a JSON data file called 'tweet_json.txt' file.

In the first step, I collect three datasets which called (tweet_json, image-Prediction and Twitter_Archive_Enhanced) each of them has unique information so I need them all.

Assessing

In the second step, I found too many Tidiness issues in the three datasets .

Twitter Archive:

- Combining "doggo", "floofer", "pupper" and "puppo" columns into one column and then clean the text
- Convert timestamp to datetime dtype.
- Replacing wrong(unappropriate) names with NONE
- Extracting the source of tweet from the url's in source column
- Drop columns with most of the NULL values
- rating_numerator and rating_denominator Columns should be merged into one column named "Rating"

Image Prediction:

- In features p1, p2 and p3, there is underscore in between the names. Replacing it with space and then converting text to camel case.
- Extracting the names of Images from the image url.
- Duplicates tweet ID for the url (https://pbs.twimg.com/media/C2kzTGxWEAEOpPL.jpg) with two tweet_id which are (822244816520155136) & (823269594223824897)
- (https://pbs.twimg.com/media/CsrjryzWgAAZY00.jpg) with two tweet_id which are (777684233540206592) & (802624713319034886) which should be dropped.¶
- The last 4 column should be combined into one column named "Category".

tweet_data:

- Rename id column to tweet_id, as it is the primary key in data and will be used later for merging the dataframes.
- Dropping non-useful columns with all NULL values
- - Should merge all the three dataframes into one dataframe.

Cleaning

In the last step, I just took the tidiness issue and fixed it by using jupytr Notebook tool using pandas and python libraries. First I fixed the Twitter archive data frame tidiness issues then the Image prediction And the Twitter API data frame.

Conclusion

In conclusion this report was for Data Wrangling only and how I wrangle the data by using specific tools and programming language.