Data Wrangling Steps:

Gathering, Assessing and Cleaning

For the completion of

Udacity project two

By

Rashidat Sikiru

Introduction

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog. This report gives a detailed report of the data wrangling steps used for the tweets. These steps are:

1. Gathering
2. Assessing
3. Cleaning

I will explain the above list one after the other.

Gathering Data

The datasets for this project were gathered from the following sources:

1. The WeRateDogs Twitter archive. This a dataset that was provided by Udacity in the csv(comma separated values) data called “The twitter_archive_enhanced.csv”.This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. This data was downloaded and was read in the jupyter notebook using the pd.read_csv(function.
2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was gotten in a TSV(Tab Separated Values) format and it was downloaded using requests library and write it to image_predictions.tsv file
3. Tweet_Json.txt: Twitter API and Python's Tweepy library was used to gather each tweet's retweet count and favorite counts as well as followers counts.

Assessing Data

Quality Issues

Twitter Enhanced Archive data

a. Incorrect data type for some columns (eg.tweet_id,in_reply_to_status_id,in_reply_to_user_id,timestamp)

b. Some rating values with decimals are incorrectly assigned

c. Missing values in columns like expanded urls,in_reply_to_user_id,etc

d. Incorrect names of Dog(e.g a,all,an,bo)

e. The text columns contain a short link at the end of each text

f. The Source Column still in its Html format

g. Extremely large rating numerator and values which are likely to be outliers

h. Retweeted text present and may act as duplicity of data

Image Prediction Data

a. Incorrect data types

b. Missing Images

c. inconsistent alphabet case in p1,p2,p3

Twitter Data

a. Incorrect data types

b. missing texts

Tidiness

Twitter Enhanced Archive data

a. doggo, floofer, pupper, puppo should be merged and renamed to dog stage

Image Prediction Data

a. This table should be together with Twitter Enhanced Archive data

Twitter Data

a. This table should be together with Twitter Enhanced Archive data

Cleaning Data

- Remove all retweeted posts
- Make a new column called dog stage to house dog stages(doggo, floofer, pupper, puppo)
- Delete columns that are not needed any longer and fill all nan in dog stage with none
- fix the wrong data types
- Change the source column from html format to text
- correct numerators decimals extracted wrongly
- Merge the image prediction table and the twitter api table
- dropped rows with no image
- Saved clean data