Wrangle Project Report

Name : Sidhika Mahadik

Introduction

The Aim of this Project is to put into analysis several things which are necessary for twitter analysis during completion of data analytics nanodegree Program. The dataset which is used during this project is the twitter archive of Twitter user @dog_rates. WeRateDogs is a Twitter account that rates people's dogs with humorous comments about dogs.

This report describes the project details:

Project details
The tasks of this project are as follows:
• Gathering data
• Assessingdata
• Cleaning data

Gathering data

The data for this project consist on three different dataset that were obtained as following:

• Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
• The tweet image predictions, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's ervers and was downloaded programmatically using the Requests library and URL information
• Twitter API & JSON: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python'

Tweepy library and stored each tweet's entity set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing Data

We have to assess the data for any inconsistency. Once the three tables were obtained I assessed the data as following:

- Visually, I used two tools. One was by printing the three entire data frames separate in Jupyter Notebook and two by checking the csv files in Excel.
- Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).

Then I separated the issues encountered in quality issues and tidiness issues. Key points to keep in mind for this process was that original ratings with images were wanted.

Cleaning data

This part consists of the defined code and test of the data to be assessed and checked for cleaning purpose. Each time we have to create a copy of the data and next to that we have to assess and clean the data.

There were a couple of cleaning steps that were very challenging. One of them was in the image prediction table. I had to create a 'nested if' inside a function order to capture the first true prediction of the type of dog. The original table had three prediction and confidence levels. I filtered this into one column for dog type and one column for confidence level.

The html strings in the source column were replaced with the display portion of itself.

The rating_numerator and rating_denominator columns were checked for value ranges; I decided to keep only tweets with single ratings. Several tweets' ratings were manually corrected with values from the text. Tweets with large numerators were dropped, as the text didn't contain a valid rating (# out of 10). After the ratings were fixed, I dropped the rating_denominator column (it contained only '10's) and renamed the rating_numerator column to rating .

Conclusion

For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases.

It is Strong in dealing with big data with the help of data wrangling process. This process is also useful when dealing with a variety of data.