Data Analysis

Data Wrangling Documentation

By alaa elhariry

Introduction

Real-world data rarely comes clean. Using Python and its libraries, we 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. We will document our wrangling efforts, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as We Rate Dogs. We Rate Dogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The purpose and goal of this project is to create a trustworthy and interesting analyses and Visualization based on the wrangled data

Project Details

This part of the project is divided into three steps, which are as follows:

- Data Gathering.
- Data Assessment.
- Data Cleaning

Data Gathering

This project requires working on three different datasets, which are acquired as follows:

- **Twitter archive**: the twitter_archive_enhanced dataset is a csv file and was downloaded manually, which was provided in Udacity's classroom.

● **Tweet image predictions**: This is a tsv file. It is hosted on Udacity's servers, and was downloaded programmatically using the Requests library.

● **Twitter API and JSON**: Using library, we were able to scrape data from theTwitter account. We looped through all the available tweets and queried Twitter's APIs along with the tweet ids to get each tweet's JSON data. After that, we extracted the data we needed, which were the number of likes, number of retweets, number of followers, number of friends, the date and the source. There were some tweet ids that the code was not able to extract its data and were stored in a list. Finally, a Data Frame with the name Tweet_JSON was created to store all the acquired information. the tweet_ids in the twitter_archive_enhanced and Python's Tweepy

Data Assessment

After gathering the required pieces of data, we assessed them visually and programmatically for quality and tidiness issues. Visually, by printing the datasets in the Jupiter Notebook and by looking at it via Excel. And programmatically, by using Pandas functions and so on. Finally, we filtered the problems into quality problems (which are issues in validity, completeness, accuracy and consistency i.e. issues in the content), and tidiness problems (which are issues in the structure), and pointed them out.

Data Cleaning

After we assessed the gathered data, we made copies of the datasets and fixed the quality and tidiness issues, and this is the third and final step in the data wrangling process. This step consists of three stages that would be applied to each assessment point in order for it to be cleaned. The stages are Define, which states what we are going to do. Code, which is where we fix the issues programmatically. And lastly, Test, which is where we make sure that our cleaning process for the issued point went well. Finally, we merged the three datasets into one with the name twitter_archive_master.

Data Storing

We stored the twitter_archive_master dataset as a csv file. At this point, the dataset was wrangled successfully and ready for analyses.