Sanatbek Matlatipov
September 13, 2021

Data Wrangling

Introduction

Data wrangling is the process of cleaning, structuring and enriching raw data into planned format for making good decisions in less time. According to my little experience in this course, almost 80% of the Data analysis process related to the data wrangling. I also understand that data wrangling is the most challenging part as real-world data almost never be clean. Fortunately, data can be gathered programmatically using many developed languages, especially, Python has many easy-to-use/handy libraries to gather, assess and clean. So, in this project I am going to use Python to gather data from different sources and assess its quality as well as tidiness, then I clean the data according to my assessment.

I will be wrangling WeRateDogs(@dog_rates) twitter archive dataset provided by Udacity's database for my nanodegree project. According to the Wikipedia [https://en.wikipedia.org/wiki/WeRateDogs],WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account has recognized internationally and gets the media attention both for its popularity. As of today, WeRateDogs[https://twitter.com/dog_rates] has over 9 million followers.

Gathering

The course has provided three types of datasets as following:

1. WeRateDogs Twitter archive file called `twitter_archive_enhanced.csv`[https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/ `[twitter-archive-enhanced.csv]` file. The file has 2356 rows × 17 columns.
2. `image_predictions.tsv` has been programmatically downloaded by using Pandas Requests[https://pypi.org/project/requests/] library. The file downloadable in the following url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
   The file is basically tweet image predictions by using AI/neural network method. So, the table has bunch of predicted breed of dogs by using tweet pictures with its confidence level in several layers.
3. Query real-time Twitter API for gathering by using Python's Tweepy[http://www.tweepy.org/] library and JSON response data had to be saved in a file called `[tweet_json.txt].` library. The reason we query API was getting **tweet ID, retweet count, and favorite count.** Unfortunately, I had to use old tweet_json.txt dataset provided by udacity nanodegree program because restiriction in Uzbekistan(https://tashkenttimes.uz/national/7168-uzbekistan-blocks-twitter-tiktok-and-vkontakte) . Currently, I think this method works best for me. Additionally, I believe I have enough experience of querying API data.

Overall, we have learnt three techniques to gather data in the course.

3.1. Manually downloading data and importing it from local storage

3.2. Programmatically downloading using URL

3.3.Querying API by taking into account of its credentials. Worth to mention that API technology is developing really fast and becoming the core of data gathering.

Assessing

The assessment of the gathered data is the second step in the data-wrangling process in both visually and programmatically (`wrangle-act.ipynb`) file.

Data quality meets six dimensions:

Accuracy – How well does a piece of information reflect reality?

Completeness – Does it fulfill your expectations of what's comprehensive?

Consistency – Does information stored in one place match relevant data stored elsewhere?

Timeliness – Is your information available when you need it?

Validity – Is information in a specific format, does it follow business rules, or is it in an unusable format?

Uniqueness – Is this the only instance in which this information appears in the database?

The requirements for tidiness:

Each variable forms a column

Each observation forms a row

Each type of observational unit forms a table

Assessment of the twitter archive is following:

- In the beginning, I've viewed all three above mentioned dataset using Numbers app of Mac for getting visual understanding. Then I've assessed it using Python features as following programmatically.
- `df.describe(), df.info(), df.value_counts(), df.shape, df.head(), df.sample(n), df.tail(), df.unique(), df.nunique(), df.query().` These are the functions I've mainly used in my assessment

Cleaning

Third step is the cleaning process of data wrangling. Here, we are changing the structure of the dataset to a desired format by following our assessment summaries. Define, Code, Test method has been used for each our assessment point. At the end I've saved the desired clean dataset to local storage for Data visualization process. More about this in the next report.

Conclusion

Overall, the desired dataset has been constructed and its final structure is as following non-null values which is the result of three tables above:

```
['tweet_id', 'timestamp', 'source', 'expanded_urls', 'rating_numerator', 'rating_denominator', 'name', 'dog_stage', 'jpg_url', 'img_num', 'img_predictions', 'img_confidence_level', 'favourites_count', 'retweet_count', 'followers_count', 'created_at'].
```

I have experienced how to gather data using requests and query API(very famous method of calling remote servers), assess data both visually and programmatically and clean it using *define, code, test* method.