20/12/2018
GIULIO CESARE MASTROCINQUE SANTO

WRANGLING EFFORTS

1. INTRODUCTION

This report will briefly discuss the wrangling efforts made during the second project (WeReteDogs) of Udacity's Data Science Foundations II course. The project is about the Twitter page "WeRateDogs". Project motivation is the ability of join multiple sources of information and being able to have meaningful insights about the posts that circulate in this web page.

2. DISCUSSIONS

The data wrangling process used in this project followed three main steps

- Gathering Data
- Assessing Data
- Cleaning Data

Each one of these steps will be discussed as well as the methodologies adopted during the project.

A. Gathering Data

In order to visualized data correctly and draw conclusions on top of them, one needs to have enough amount of data. That means data must be gathered from different sources and combined to make a solid data frame with rich information that can be used in further analyses.

In this particular project, data was gathered from three main different sources:

- A coma-separated values (CSV) file with 2356 tweets information's, such as tweets texts, tweets rating values and so on so forth.

- A flat file hosted in Udacity's webservers that can be accessed through the URL https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. This file has tweet image predictions that estimate the dog's breeds.

- Twitter API. Using Python's Tweepy library, data such tweets ID, retweets count and likes count was gathered.

The gathering process was done using Python and a Jupyter Notebook. To read flat files, pandas was used through its "pd.read_csv" method. Moreover, data was extracted from the internet using "requests" library.

Data was opened and visually validated and three initial data frames were created for each gathering source: twitter_archive, image_predictions and tweet_json respectively.

B. ASSESSING

Assessing is required to find data issues. That can be done visually and programmatically. Both approaches were used in this project using the Jupyter Notebook.

For visual assessment, functions such as df.head(), df.sample() and df.iloc() as well as columns deployments were used. For programmatic assessment, functions such as value_counts(), df.info(), df.shape, isnull(), sum() and others were used.

As a result of the assessing efforts, thirteen quality issues were documented, being eight of them related to the "twitter_archive" table, four related to the "image_predictions" table and one related to the "tweet_json" table. Moreover, five tidiness problems were pointed out.

C. CLEANING

Cleaning data is the final step of the wrangling process. Both quality and tidiness issues were addresses during this process.

Tidiness issues are cleaned first. However, in this project I first addressed some quality issues such as missing values and excessive unnecessary values (tweets that doesn't exist anymore) to make cleaning easier. Then tidiness was targeted.

To make the data tidy, cleaning was made following the three main following concepts:

- Each variable (attribute) is a column of the data set
- Each observation is a row of the data set
- Each type of observational unity is a table

The result was two main data frames: twitter_archive_clean and dogs_data. The first one containing variables specifically related to tweets information (such as tweet text and number of likes and retweets) and the second one containing only information about the dogs.

Finally, quality issues were addressed to correct problems such as wrong data type, inaccurate data values, unappropriated columns labels and so on so forth.

3. DATA STORE AND VISUALIZATION

The final cleaned data frames were stored in CSV files as well as in an SQLite database. Insights and visualizations were documented producing a report.