wrangle_report.pdf

This document describes the wrangling efforts in order to gather, assess and clean data for werate dogs using 3 different data sources.

1. Gathering data

I collected data from three different sources. The first dataset was downloaded manually using the link provided. I downloaded that file and imported its content into twitter_archive dataframe. After that, I programmatically download the tweet image predictions file from Udacity cloud server and then I imported content of that file into image prediction dataset.

The most challenging part of data gathering was to query the Twitter API and read JSON data using the Tweepy library. This task was challenging it took me a lot of time to figure out that how I can download data. With the help of some articles on stack overflow, finally I was able to query Twitter API and was able to import that data (tweet ID, retweet count, and favorite count) into tweets_data dataframe.

2. Assessing data

After gathering data from above mentioned three sources, I assessed data visually and programmatically for quality and tideness issues.

Following programmatic methods were used to assess the data.

.head (Dataframe and Series)

.duplicated (DataFrame and Series)

.sample (DataFrame and Series)

.info (DataFrame only)

.describe (DataFrame and Series)

.value_counts

Various methods of indexing and selecting data such as iLoc were also used.

I found the following quality and tidiness issues after assessing the data.

[Tidiness] (structure issue):

- In image prediction table, stages of dog (doggo,floofer,pupper and puppo) variables are in 4 separate columns. However, this information should be available in single column.
- The twitter data for analysis should be in single table so twitter archive, image prediction and tweets data table should be combined.

[Quality] (contents issue):

In twitter archive table:

- tweet_id should be string as descriptive analytics cannot be performed on this.
- timestamp column should be date time variable.
- retweeted_status_timestamp column should be date time variable.
- Tweets with no images should be removed i.e. expanded URL is null.
- retweets should be removed.
- There are links in the source column. We should keep the relevant text and remove url.
- Ratings are not extracted properly. Ratings should be extracted from text column and denominator for all records should be 10.
- Name column needs to be cleaned. Names of dogs are misspelled, incorrect.
- Separate columns should be there for tweet date and time (Timestamp Column). It will help us to perform better analysis.

- Columns which are not required for analysis should be removed from dataset

In Image prediction table:

- tweet_id should be string
- Let's keep the dog breed and confidence level for true prediction.

In tweets data table :

- tweet_id should be string

3. Cleaning data

This was the third step in the data wrangling process. I followed the data wrangling principles and therefore defined, coded and tested each of tidiness and quality issues which I identified during data assess step.

The cleaning process was quite cumbersome and it took a lot of effort to clean all the tideness and quality issues. After all the issues were fixed, I exported data into csv file and saved it on disk so that I can use it later for data analyses and visualizations

4. Conclusion

In summary, this was one of the most challenging project I have undertaken while doing data analyst nanodegree. I thoroughly enjoyed this project and learned a lot from this project. Now, I can apply my newly learned data wrangling skills with more confidence.