wrangle_report

January 9, 2018

1 Data Wrangling Efforts

During this project many data wrangling processes and efforts were made. Firstly, it was given a file with close to 2500 Tweets entries with several different information about each tweet. Opening this csv file and exporting it to a Pandas Dataframe was rather easy and straightforward.

The most demanding exercise was: - Using the Twitter API to query information from each tweet present on the CSV file given. - Establishing the connection and querying the information was a process that took some time, but once established and set-up, worked exceptionlly well.

Once the information was gathered, it was saved on a json .txt file. - This file contained a json entry with many different fields from each individual tweet. - Close to 2500 entries. This file resulted to be extremely long and extensive in information. - The Tweet API return a json file from each tweet on which we performed a query. - This Json information contains many fields, most of which were not used during this project.

The major problem on data gathering was: - The data of the json file created was organized into a comprehensive dataframe - Information from each tweet ID present was gathered. - Json returned from Tweet API was not very clear - Determining which Tweets were a retweet, a reply, original, or self retwet was rather complicated. - This information was not readily available, or the format would change from tweet to tweet - Json format seems to have changed with time, which complicated the gathering process - This meant some extra work investigating programmatically each tweet - Each Tweet's text was analysed individually, looking for specific entries that determined the type of Tweet

Once all this information was gathered, a dataframe was created. - Dataframe was complemented with extra information - From the CSV file it was merged the Dog Stage information - From the TSV file the Breed prediction was gathered

Cleaning the data is always the most demanding process. Efforts wre made in the following areas: - Cleaning entries that do not represent a tweet about a dog - Cleaning wrong ratings or missing ratings - All tweets without a link were corrected - Missing Dog's names were obtained, as much as possible. - NaN values were removed or replaced by the correct ones - Information coming from 3 different sources were merged in one comprehensive dataset - From that unique dataset, derivatives datasets were created with important information

The author attempted to clean the data as much as possible, all efforts made are extensively reproted on the project code.