Report on Wrangling of the datasets for the analysis of Dog Ratings

TABLE OF CONTENTS

INTRODUCTION

This report is a project on Data Wrangling as part of the requirements to complete Data Analytics with Python. The project centres on Data Wrangling of dog ratings from a twitter user.

AIM

The aim of this report is to give a roadmap as to how 3 datasets were successfully gathered from different sources, how they were assessed and cleaned.

BACKGROUND

This report gives a walkthrough of wrangling 3 datasets. The main dataset wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs. The ratings are actually over 10 but most of the ratings in the dataset exceed the number of 10. The tweet archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. The tweet archive is filtered to only tweets with rating making 2356 tweets.

The second dataset is an image prediction dataset, in which an algorithm was used to predict the breed of the dog from the images in each tweet. The dataset contains the tweet_id, 3 predictions for each id, their confidence and a Boolean column stating if the prediction is a dog breed.

The third dataset contains the favorite count and tweet count for each tweet.

STEPS SPECIFICATION

The wrangling process was in 3 steps starting from gathering, assessing and cleaning.

GATHER

Each of the datasets were gathered differently.

For dataset 1, a csv file called twitter_archive_enhanced.csv was provided, which was downloaded manually and opened in jupyter notebook using pandas.

Dataset 2 was downloaded programmatically using the requests library and a url.

Dataset 3 was obtained by querying data from Twitter API using the tweet_ids in dataset 1. Each tweet was queried using python's tweepy library. Each tweet's entire set of JSON data was stored in a .txt file. Each tweet's JSON data was written on its own line and then this .txt file was read line by line into a pandas DataFrame .

ASSESS

Each of the dataset was assessed both visually and programmatically to find out their quality and tidiness issue.

CLEAN

Each dataset was cleaned using a 3 step process, define, code and test. They were cleaned based on the assessments documented. A couple of pandas and numpy methods were used to achieve this.

CONCLUSION

After all quality and tidiness issues were resolved, the 3 datasets were merged into 1 and saved into a single csv file. The file was thereafter analyzed and a couple of insights were obtained from it.