Wrangle Report

March 12, 2019

By- Piramanayagam S

1 Introduction

The data wrangling project was very challenging and i learned alot about the gathering process and the Twitter API.

Data wrangling consists of: - Gathering data - Assessing data - cleaning data

2 Gathering

Gathering Data for this project composed 3 pieces of data - The WeRateDogs Twitter archive. We will download this file manually by clicking the following link: twitter archive_enhanced

- The Tweet image predictions i.e, what breed of dog is present in each tweet according to a neural network. We will can downlaod this file manually by clicking the following link : image predictions.tsv
- Each tweet's retweet count and favorite (i.e. "like") count at minimum, and any additional data we will find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, we will query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

2.1 Gather : Summary

Gathering is the first step in the data wrangling process. We could finish the high-level gathering process:

Obtaining data Getting data from an existing file (twitter-archive-enhanced.csv) Reading from csv file using pandas Downloading a file from the internet (image-predictions.tsv) Downloading file using requests Querying an API (tweet_json.txt) Get JSON object of all the tweet_ids using Tweepy importing that data into our programming environment (Jupyter Notebook)

2.2 Assessing data

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues will be our newt step. We will detect and document at quality issues and tidiness issues.

2.3 Quality

Completeness, Validity, Accuracy, Consistency => a.k.a content issues

archive dataset

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be intergs instead of float
- retweeted_status_timestamp, timestamp should be datetime instead of object (string)
- We only want original ratings (no retweets) that have images

images dataset

- Missing values from images dataset (2075 rows instead of 2356)
- Some tweet_ids have the same jpg_url
- Some tweets are have 2 different tweet_id one redirect to the other

Tidness Untidy data => a.k.a structural issues

- No need to all the informations in images dataset, (tweet_id and jpg_url what matters)
- We may want to add a gender column from the text columns in archives dataset
- All tables should be part of one dataset

2.4 Cleanning data

Cleaning our data is the third step in data wrangling. It is where we will fix the quality and tidiness issues that we identified in the assess step.

3 Conclusion

This Project was my biggest challenge to date,specifically using the Twitter API to gather the JSON data.Data wrangling is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on insights, and wasting time.