Wrangling Description

Jared D’souza

This document describes the wrangling process used in the project. The process was built around the framework: Gathering, Assessing, Cleaning

1. Gathering: For the project, two files were downloaded on the jupyter notebook. One file is a .csv file. The other is a .tsv file with the tweet image predictions present in each tweet. One file was downloaded onto the notebook by querying Twitter’s API for tweets’ JSON data using the Tweepy library This file was loaded as a .txt file which was later made into a data frame. To extract the .txt file, I first created a developer’s account with Twitter and used the necessary codes to query Twitter’s API to extract data. The code that used Tweepy was available to me from Udacity’s project page.
2. Assessing: Using pandas, I assessed the data on various aspects such as data type, number of missing elements, unwanted rows, content etc and as many data quality and tidiness issues as I could. The function that I mostly used to assess the data was .info(). In addition to this function, I visually browsed through the data using the .head() function.
3. Cleaning: After assessing all 3 data frames, I listed out those quality and tidiness issues that I would rectify for the scope of this project. Even though there are a lot of issues present, I stuck to the minimum project requirements. This I rectified each issue in the list using the ‘Define, Code, Test’ framework. First, I cleaned the data frames one by one. After, I merged the three frames into one data frame using the left join function. With this one data frame, I tidied and cleaned everything else on the list.

On completion of the above, I saved the data frame as a .csv file. One visualization was developed for analytical purpose using the numpy library.