Data Wrangling of Twitter Data

Qi Zhao

March 28, 2018

1 Overview

I have encountered many difficulties in this project. The data is neither clean nor tidy. Therefore, there are many things that need to be done before we can visualize the data and have some insights. Fortunately, following the steps of gathering, accessing and cleaning data, I created a clean dataset which makes it much effortless for us to do some exploratory data analysis.

2 Gathering the Data

The biggest problem when gathering the data was that my computer and my Internet connection is too slow which made it painful for me to wait. Unfortunately, I made a mistake when I tried to write the json objects to the txt file. I forgot to write each json object in one single line. However, I used some pattern in the file to successfully read the number of "likes" and "retweets".

3 Accessing the Data

When accessing the data, I immediately realized that the type of the ID's are all numerical. Then I found that there are a lot of "a" and "the" in the name column. And the some rating denominator looks like outliers. Besides, I see the columns like "doggo", "puppo", "floofer" and "pupper", which are mutual exclusive so that they should be put in one column.

4 Cleaning the Data

Some data are easier to clean. We can manage to get what we want simply through a function and the "apply" function in the "dataframe" class of

pandas. However, there are some columns that I have to personally read through the text to figure out a way to clean them. For example, it took me great effort to get all the names right because the name are not all mentioned after "This is".