Data Wrangling

The objective of data wrangling is quite clear: structure the dataset and correct erroneous contents. The project is achieved by three main steps: Gather, Assess, and Clean. I will elaborate on the process I went through in three bullet points.

- Gather: there are plenty of sources from which you can gather the information for your analysis. In order to collect all necessary information, I use three different ways to download all necessary tables.
  - ✓ CSV: I download manually Enhanced Twitter Archive and upload the archive to Jupyter Notebook Workplace. Then I simply use “read_csv” function in pandas to load the data.
  - ✓ Internet: a part of the data is hosted on Udacity server. Therefore, I have to download the information programmatically using Requests library and save the information in a tsv file.
  - ✓ API: to gather information via API is the most challenging in this step. Unlike MediaWiki, you can only retrieve data via API of twitter after you are granted the access and create API key and token. Below is the syntax I used to retrieve data from Twitter API.

In [12]:
```
# Query Twitter's API for JSON data for each tweet ID in the Twitter archive
count = 0
fails_dict = {}
start = timer()
tweet_ids = twitter_archive.tweet_id

# Save each tweet's returned JSON as a new line in a .txt file
with open('tweet_json.txt', 'w') as outfile:
    # This loop will likely take 20-30 minutes to run because of Twitter's rate limit
    for tweet_id in twitter_archive['tweet_id']:
        count += 1
        print(str(count) + ": " + str(tweet_id))
        try:
            tweet = api.get_status(tweet_id, tweet_mode='extended')
            print("Success")
            json.dump(tweet._json, outfile)
            outfile.write('\n')
        except tweepy.TweepError as e:
            print("Fail")
            fails_dict[tweet_id] = e
            pass
end = timer()
print(end - start)
print(fails_dict)
```

- Assess: data assessment is integral to data analysis. There are two different types of data issues you would encounter in assessment stage: quality and tidiness issues. The quality issue is related to any problems from content. On the other hand, the tidiness issue is related to any structural problems.
  - ✓ Quality issues: I start with visual assessment and already spot some data errors, such as denominators of rating greater than 10, misrepresentation of dog names like ‘a’ or ‘the’. Then I use pandas functions to assess data programmatically to identify other quality issues like wrong data type and retweet data in Enhanced Twitter Archive.
  - ✓ Tidiness issues: visual assessment is more efficient to spot tidiness issues. By browsing rapidly through the dataset, I already noticed the messy presentation of source column. Then I also check if I can merge these three datasets to make the data more readable and condense.

- Clean: this is the last step in data wrangling process. The idea is to use codes to solve any quality and tidiness issues identified in the assessment stage. I follow three steps for data cleaning.
  - ✓ Define: based on the assessment, I have to define how to solve those issues with concrete steps. Different from assessments, which are observations, the definition here has verbs. Not following the procedure used in the course, I cluster three steps of data cleaning by each issue.

Define

Remove html tags in the source.

Code

```
In [34]: import re

         def remove_html_tags(df):
             clean = re.compile('<.*?>')
             return re.sub(clean, '', df['source'])

         twitter_archive_clean['source'] = twitter_archive_clean.apply(remove_html_tags, axis=1)
         twitter_archive_clean['source'] = twitter_archive_clean['source'].astype('category')
```

Test

```
In [35]: twitter_archive_clean['source'].value_counts()

Out[35]: Twitter for iPhone     2042
         Vine - Make a Scene      91
         Twitter Web Client       31
         TweetDeck                11
         Name: source, dtype: int64
```

  - ✓ Code: this is a difficult step in data cleaning. Since there are countless problems you could come across, and each problem requires different sorts of code to cope with, I spend most of my time on this step. After several times of try and error, I could find the optimal solution for each issue I enlisted in the assessment.
  - ✓ Test: the test serves to test the code you write in code section. If the code does not work appropriately, I simply look for some solutions on the internet and try other syntax.
- Store: after iterate the wrangling process, I save the cleaned data into a master file for my analysis.

Store

```
In [125]: twitter_archive_clean_final.to_csv('twitter_archive_master.csv', index=False)
```

- Data Visualization and Analysis: I take data from the master file to conduct analysis and visualization.

Data wrangling takes plenty of time and energy, since you need to find solutions to many problems. When you clean the dataset, you could encounter other problems which have to be addressed in order to proceed. However, data wrangling enables me to have good base to conduct my analysis and more likely to find more insights. The process is laborious but fundamental to good data analysis.