**Reddit Political Posts Scapred by PRAW and Reddit Api**

Proposed by Yanru Jiang

**Task Description**

Both PRAW and Reddit Api are the commonly used APIs for scraping Reddit Posts. While PRAW has a higher maximum limit of requests that can be sent to Reddit (N = 1000), Reddit Api only allows developers to send 100 requests each time. Overall, PRAW is more stable and powerful, while Reddit Api could provide more variety of data, especially the url of images that are associated with the posts.

In this task, I would like you to combine two datasets together. I have used Python Pandas to transfer JSON files into dataframes, and saved as csv files. I expect you to use Python to merge two files together (PRAW and Reddit Api datasets), perform some data transformation, store images at your local, and transfer the final dataframe in MySQL.

**Tasks to complete**

1. Merge two csv files by joining PRAW to Reddit Api (deduplicate the repetitive posts)
2. Use datetime to transform the UTC time to readable timestamps
3. Use requests to download images using their urls and store them in your local. Please notice that the images urls collected by Reddit Api cannot be used for download directly. You need to perform some data transformation of urls for them to be usable (Hint: replace “amp;” in urls to blank, ie. delete all “amp;”).
4. Transfer data frame to MySQL

**Bonus**

1. Use Spacy and NLTK to perform name entity recognition (NER) on each title in the “title” column. Store the labels and their associated entities to a new data frame. Make sure each entity can be allocated to their titles by using “id” of the posts.
2. Store the NER data frame in MySQL.
3. Use Geopy to visualize all GPE elements (locations) from NER based on the frequency