# Weibo Crawler

Objective

To crawl weibos posted in Singapore POIs and the user profile.

Dependence

Python libraries

pandas

json

csv

urllib3

datetime

time

random

Project Structure

Weibo\_crawler

|-src

|-get\_poi\_list.py-→Write to poi\_list.csv

|-each\_poi.py -→to get the tweets

|-target\_users.py -→read in user\_data.csv (newly crawled) and cleaned\_data.csv (database of stored user). If user\_data not in cleaned\_data, add to the target\_user.csv list

|-clean\_data.py -→Verified and unverified user from weibo\_data.csv

|-verified\_user.py -→require user account and cookie info

|-non\_verified\_user.py -→require user account and cookie info

|-download\_pic.py -→reads in string from pictures column in weibo dataset

|-data

|-poi\_list.csv

|-target\_user.csv

|-target\_verified\_user.csv

|-output

|-pictures

|-weibo.csv -→Contins weibo posts

|-user\_data.csv-→Contains database of users

|cleaned\_data.csv

The project is structured into three main folders which are src, output, and data for storing source code, data crawled by the project and data generated during the process. Below are the description of each file:

src/get\_poi\_list.py: this crawls the SG POIs in Weibo and generate poi\_list.csv for other crawling use.

src/each\_poi.py: this goes through each SG POI webpage in Weibo and crawls all the weibos posted there today, yesterday and the day before yesterday and the crawled data is to be saved in weibo.csv.

src/target\_users.py: this generates a list of users whose profile haven’t been downloaded. The user list stored in the csv file target\_user.csv.

src/clean\_data.py: this removes redundant weibos and save data to cleaned\_data.csv meanwhile it also generates a verified user list as target\_verified\_user.csv.

src/verified\_user.py: this will crawl user profile, who is in the target\_verified\_user.csv, and save them in user\_data.csv

src/non\_verified\_user.py: this will crawl user profile, who is in the target\_user.csv, and save them in user\_data.csv

src/download\_pic.py: this download pictures from weibo

Data Flow

