# Instagram Crawler

Objective

To crawl ins posted in Singapore POIs.

Dependence

Python libraries

pandas  
csv

time

datetime

json

urllib3

urllib

Directory structure

|-area\_list.csv : filled up manually. Used to feed into get\_poi.py and write out to poi\_url.csv

|-cleaned\_poi.csv: Renamed from cleaned\_poi.csv. poi\_rul column used as inputs into ins\_crawler.py

|-ins\_data.csv: To write out post data

|-get\_poi.py: To crawl poi

|-ins\_crawler.py: To crawl instagram posts

Description

The only condition that you will need chromedriver and selenium is that you wish to crawl poi list again by yourself. For that case, you will also need scrapy

Run get\_poi.py to get the POI list from instagram. Requires area list (area\_list.csv filled up manually).

- To add new location for crawling, go to cleaned\_poi.csv and add additional row with the associated url (e.g. /explore/locations/1740420/suntec-singapore-convention-exhibition-centre/). The crawler will append with www.instagram.com with the above extension

Run ins\_crawler.py will crawl ins posted in SG POIs.

Data

Owner column refers to instagram id from JSON file

Instagram data crawled does not differentiate between locals and tourists. Further differentiation requires further data cleaning through text.

**Command line arguments**

source activate <environment\_for\_crawling>

python3.7 get\_poi.py

python3.7 ins\_crawler.py