**Getting old tweets, Instagram posts and Google vision output**

This document shows how to (i) get old tweets, (ii) get Instagram posts, and (iii) use Google vision for image analytics. Please read each line carefully before you run anything. This will help avoid frustration and wastage of time.

**Getting old tweets**

Go to the bash console and install the required library by using the following command:

pip3.6 install --user GetOldTweets3

Now you can get old tweets in different ways.

**Please note that you must run this script directly from the bash console.**

**Getting tweets by username:**

GetOldTweets3 --username "barackobama" --maxtweets 1000

**Getting tweets by a query search**:

GetOldTweets3 --querysearch "europe refugees" --maxtweets 1

**Getting tweets by a username and bound dates:**

GetOldTweets3 --username "barackobama" --since 2015-09-10 --until 2015-09-12 --maxtweets 1

**Getting tweets by a language (you can specify en for English):**

GetOldTweets3 --querysearch "bitcoin" --maxtweets 10 --lang cn

**Get tweets by location:**

GetOldTweets3 --querysearch "bitcoin" --near "Berlin, Germany" --within 15mi

**Of course you can combine the various options:**

GetOldTweets3 –querysearch “Trump midterm” –since 2018-10-12 –until 2018-11-06 –maxtweets 5000

The output will be saved in your files section as before.

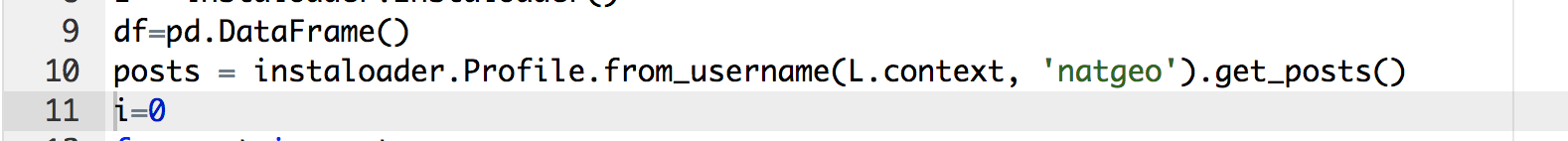
**Accessing Instagram posts**

The script is called **Insta\_withoutComments.py** since it does not fetch the comments (only the number of likes).

Go to bash and install the library by using the following command:

pip3.6 install --user instaloader

The hashtag should be given in line 10 (replace natgeo by your hashtag)



Run the script as any other script you have run. The code gives around 400 posts in one run.

The output will be saved in your files section.

**Google vision**

Upload the json file (Python-03112e22a573.json) to PythonAnywhere. Also upload the python script **gcloud.py**

The second and the third lines in the script are the libraries you have to install in the bash console. Also, if you have to rerun the code, it is important that you go to the consoles and **kill** the previous console. Results are stored as Labels.xlsx.

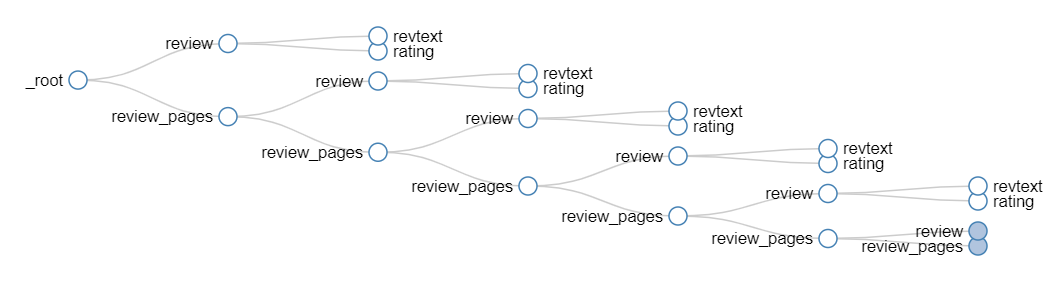
The single column input file is on line 15: loc = ("image\_url.xlsx") This file must contain a single column (no column label) with the URLs to images you got from Instagram or another site.

Now run the gcloud.py script. The output is provided in Labels.xlsx

**Web scraper sitemaps**

**Tripadvisor – uses links for pages, but can’t click on them to select (put cursor on link, and use S button)**

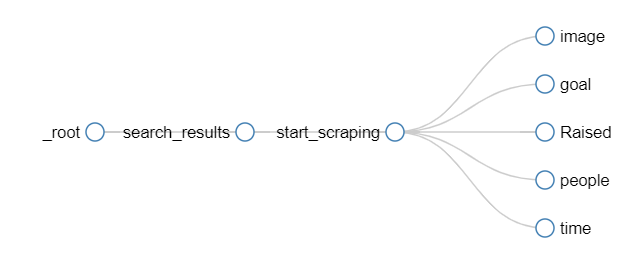
**Reviews of the Taj Mahal Palace Hotel, Mumbai, India**



{"\_id":"taj","startUrl":["https://www.tripadvisor.com/Hotel\_Review-g304554-d302179-Reviews-The\_Taj\_Mahal\_Palace\_Mumbai-Mumbai\_Maharashtra.html"],"selectors":[{"id":"review","type":"SelectorElement","parentSelectors":["\_root","review\_pages"],"selector":"div.rev\_wrap div.ui\_column.is-9:nth-of-type(n+2)","multiple":true,"delay":0},{"id":"revtext","type":"SelectorText","parentSelectors":["review"],"selector":"p.partial\_entry","multiple":true,"regex":"","delay":0},{"id":"rating","type":"SelectorElementAttribute","parentSelectors":["review"],"selector":"span.ui\_bubble\_rating","multiple":false,"extractAttribute":"class","delay":0},{"id":"review\_pages","type":"SelectorLink","parentSelectors":["\_root","review\_pages"],"selector":"div.mobile-more a.pageNum:nth-of-type(n+2)","multiple":true,"delay":0}]}

**GoFundMe (uses Load More/Show More)**

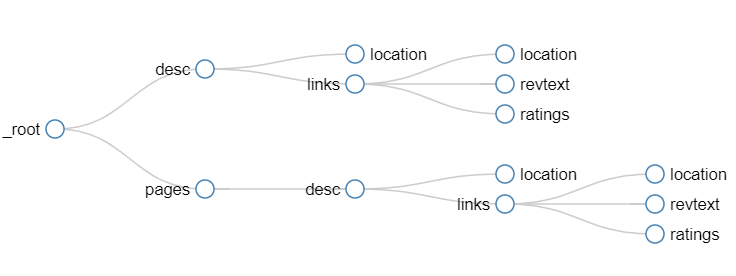
**Search terms: “Golden pups”**



{"\_id":"gofundme","startUrl":["https://www.gofundme.com/mvc.php?route=homepage\_norma/search&term=golden%20pups"],"selectors":[{"id":"start\_scraping","type":"SelectorLink","parentSelectors":["search\_results"],"selector":"a.campaign-tile-img--contain","multiple":true,"delay":0},{"id":"search\_results","type":"SelectorElementClick","parentSelectors":["\_root"],"selector":"div.content-section.content-section--tan div.grid-container","multiple":false,"delay":0,"clickElementSelector":"div.cell div.cell a.button","clickType":"clickOnce","discardInitialElements":false,"clickElementUniquenessType":"uniqueText"},{"id":"image","type":"SelectorImage","parentSelectors":["start\_scraping"],"selector":"div.main-column img.campaign-img","multiple":false,"delay":0},{"id":"goal","type":"SelectorText","parentSelectors":["start\_scraping"],"selector":"div.main-column span.smaller","multiple":false,"regex":"","delay":0},{"id":"Raised","type":"SelectorText","parentSelectors":["start\_scraping"],"selector":"div.main-column h2.goal strong","multiple":false,"regex":"","delay":0},{"id":"people","type":"SelectorText","parentSelectors":["start\_scraping"],"selector":"div.main-column div.campaign-status span","multiple":false,"regex":"","delay":0},{"id":"time","type":"SelectorText","parentSelectors":["start\_scraping"],"selector":"div.main-column div.campaign-status","multiple":false,"regex":"","delay":0}]}

**Yelp**

**Apartments near 78705**



{"\_id":"aus\_apt","startUrl":["https://www.yelp.com/search?find\_desc=apartments&find\_loc=Austin,+TX+78705"],"selectors":[{"id":"desc","type":"SelectorElement","parentSelectors":["\_root","pages"],"selector":"li.regular-search-result:nth-of-type(n+5)","multiple":true,"delay":0},{"id":"location","type":"SelectorText","parentSelectors":["desc","links"],"selector":"address","multiple":true,"regex":"","delay":0},{"id":"links","type":"SelectorLink","parentSelectors":["desc"],"selector":"a.biz-name","multiple":true,"delay":0},{"id":"revtext","type":"SelectorText","parentSelectors":["links"],"selector":"div.review-content p","multiple":true,"regex":"","delay":0},{"id":"ratings","type":"SelectorHTML","parentSelectors":["links"],"selector":"div.biz-rating div div.i-stars","multiple":false,"regex":"","delay":0},{"id":"pages","type":"SelectorLink","parentSelectors":["\_root"],"selector":"a.pagination-links\_anchor","multiple":true,"delay":0}]}