CS172 Project Part A Report

Collaboration Details

Hongchao Yu

- Implemented a basic twitter crawler function using Tweepy, a python library for accessing Twitter API
- Separated data storage into .txt files of size 10MB each. Crawler stops once size goal has been reached (2GB)

Saikrishna Reddy

- Assisted in the acquisition of data from teammates and ran a QA role, testing code and ensuring it works in line with the given requirements.
- Worked on a solutions architecture role ensuring proper designs, explanations, descriptions of the engineering done by partners via the report.

Minwhan Oh

- Implemented title crawler for a given URL using BeautifulSoup4.
- Tested the application, and organized screenshots in the report.
- Put the crawled title into the json object (incomplete).

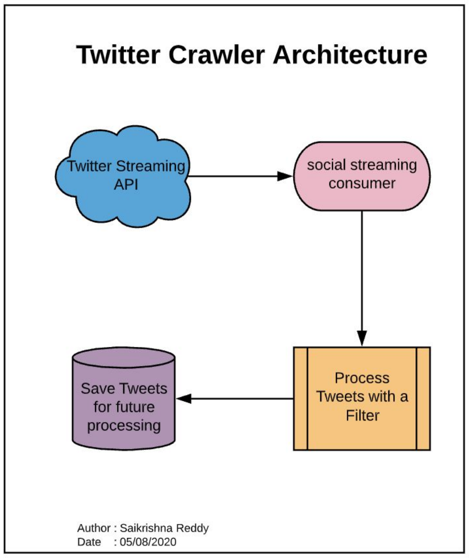
Deven Fafard

Overview of System

We used the Twitter Streaming API to collect geolocated tweets and stored it in large files of about 10 MB while having one tweet per row. With Twitter's Streaming API, users record a set of data (i.e keywords, usernames, locations, named places, etc.). After this, the tweets match the data, which is then pushed to the user. In a sense, this is mutualism between Twitter and the end-user. For example, you inform Twitter whenever they get hold of tweets with a keyword, say "apple pie", they can process it and send it to you in real-time. In other words, Twitter is pushing the data requested by the end-user, rather than a pull by the end-user.

a. **Architecture**

Tweets are streamed in batches using the Twitter API. Tweets are stored, then parsed, and qualifying tweets are stored separately. Tweets that contain HTML links are handled to retrieve the <title> tag of the HTML page.


Twitter Crawler Architecture
Twitter Streaming API
social streaming consumer
Process Tweets with a Filter
Save Tweets for future processing
Author : Saikrishna Reddy
Date : 05/08/2020

b. **Data Collection Strategy**

We stream about 1 million tweets using the Twitter API and store them in a file. We then discard all tweets that do not qualify for our purposes: tweets not in English, tweets with no location, etc. The remaining tweets stay intact after discarding non-qualifying tweets; these are stored in another file. We also used urllib3 to retrieve the <title> tags from each HTML page if there was one in the tweet object. Often, users would put URLs to other tweets or Instagram posts which would essentially be a retweet, so to avoid those, if the URL container Twitter or Instagram, it would automatically be filtered out. The processed title tags would also have trailing hex characters so those were removed as well.

c. Data Structures Employed

Tweets obtained from our crawler are stored in a JSON object. The fields of this JSON object includes the text of the tweet, the user who tweeted, the date of the tweet, urls, etc. At the moment, the JSON fields are the default that the Twitter API returns. We plan to further parse through this data in part B into fields that will be useful to query through.

System Limitation

- Have not rigorously optimized our data collection strategy, thus, crawling efficiency/speed may vary
- If too many users are trying to run the crawler, twitter API may be rate-limited due to too many requests
- A lower percentage of twitter users uses the geolocation feature so data accumulation may be slowed
    - Strategy to counter this is to currently expand our bounding box coordinates
- The standard library of Python cannot perform the various HTML operations(e.g. Request, read URL). Hence the additional library called BeautifulSoup4 is required.

System Deployment

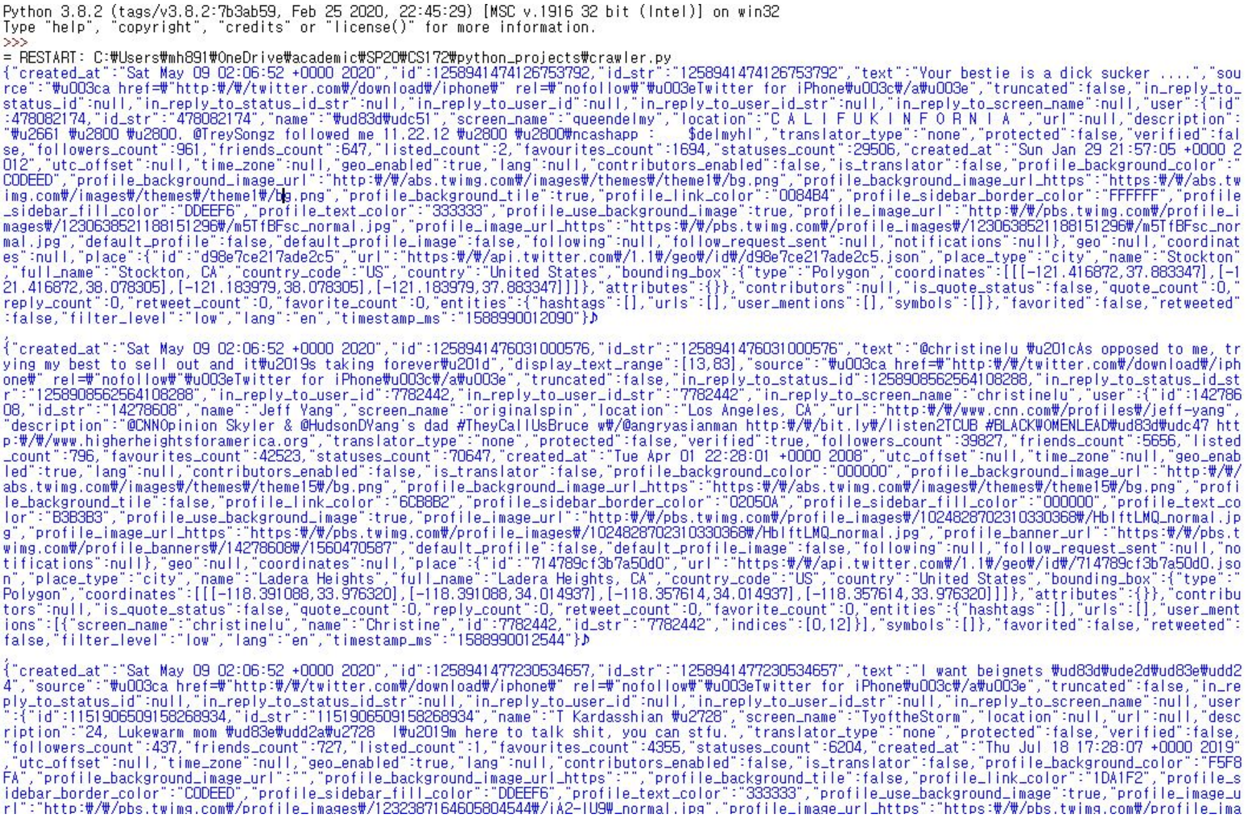
To run the program, you must have the following

- Python 3 (any version but NOT in Windows)
- Tweepy library
    - Pip3 install tweepy
- BeautifulSoup4 library
    - Pip install bf4 (in windows10, type “py -m pip install bf4” in command prompt window)

After downloading the zip file we have uploaded, simply run the following command

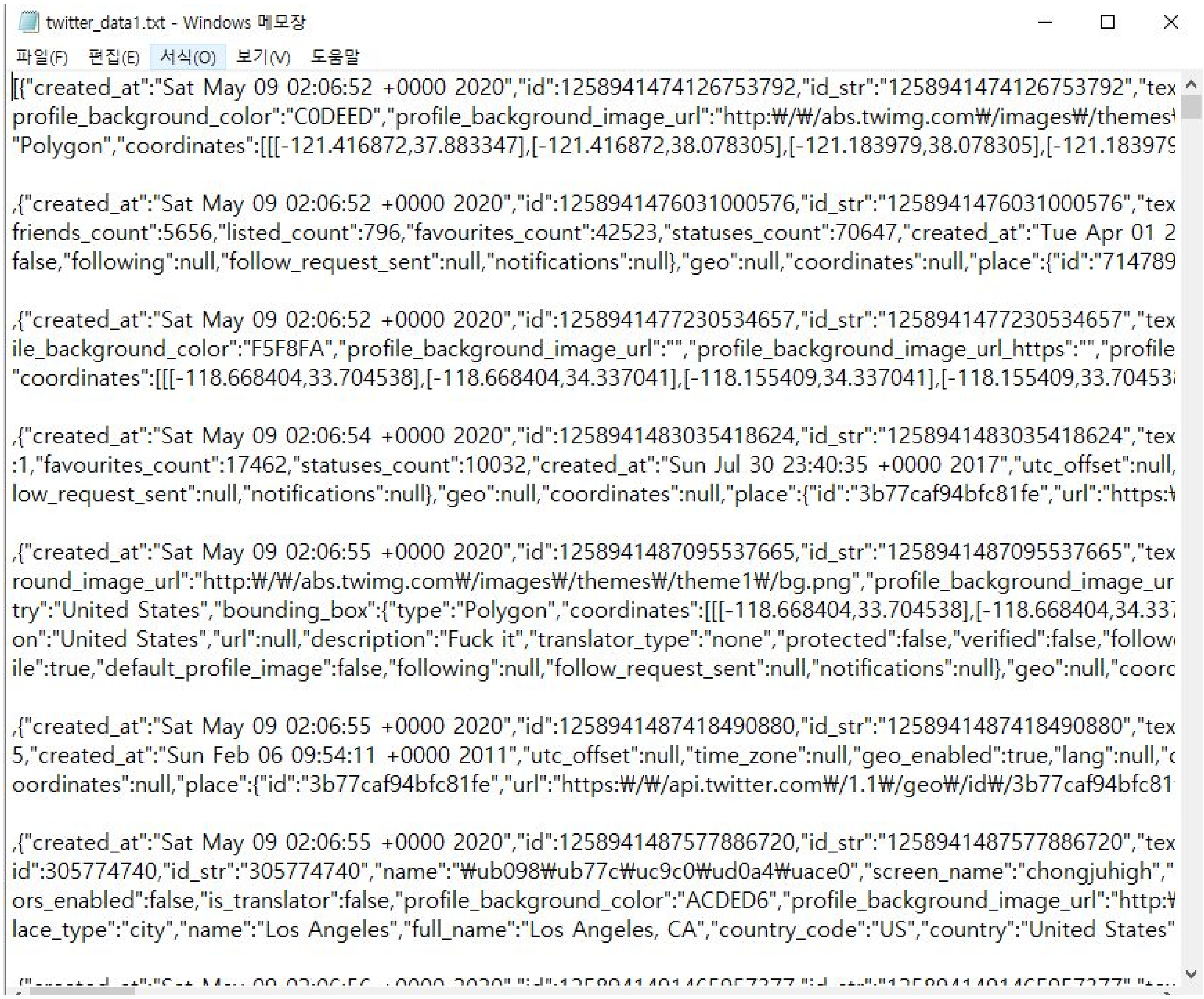
- ./twitter_crawler.sh
    - You may need permission to run the bash script so before the command above run chmod+x twitter_crawler.sh
- For windows
    - Run IDLE(Python 3.x 32bit), navigate to the directory of extracted zip to open crawler.py
    - Run - Run Module, or Press F5 to run
    - You can check the tweet_data1.txt in the newly created folder inside of the project folder.

Screenshots



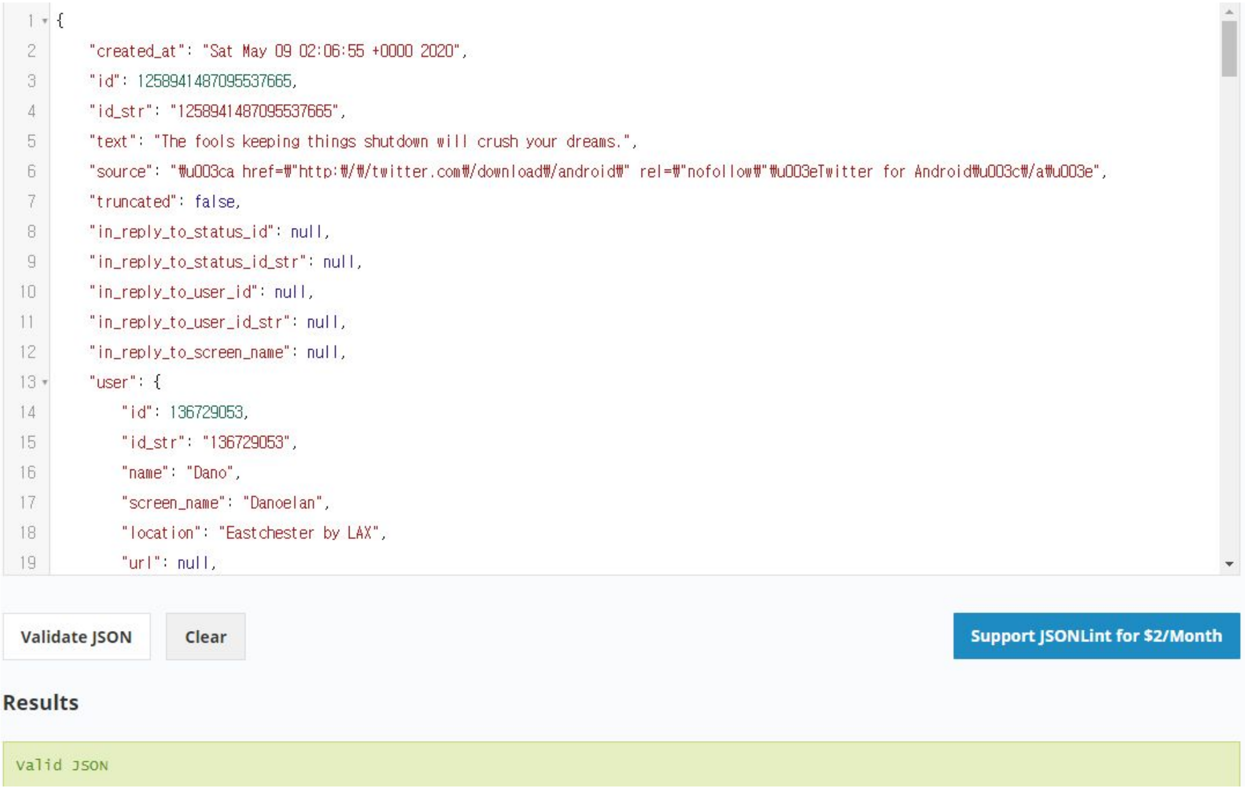
```
Python 3.8.2 (tags/v3.8.2:7b3ab59, Feb 25 2020, 22:45:29) [MSC v.1916 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: C:₩Users₩mh891₩OneDrive₩academic₩SP20₩CS172₩python_projects₩crawler.py
{"created_at":"Sat May 09 02:06:52 +0000 2020","id":1258941474126753792,"id_str":"1258941474126753792","text":"Your bestie is a dick sucker ....","sou
rce":"₩u003ca href=₩"http:₩/₩/twitter.com₩/download₩/iphone₩" rel=₩"nofollow₩"₩u003eTwitter for iPhone₩u003c₩/a₩u003e","truncated":false,"in_reply_to_
status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id"
:478082174,"id_str":"478082174","name":"₩ud83d₩udc51","screen_name":"queendelmy","location":"C A L I F U K I N F O R N I A ","url":null,"description":
"₩u2661 ₩u2800 ₩u2800. @TreySongz followed me 11.22.12 ₩u2800 ₩u2800₩ncashapp :    $delmyhl","translator_type":"none","protected":false,"verified":fal
se,"followers_count":961,"friends_count":647,"listed_count":2,"favourites_count":1694,"statuses_count":29506,"created_at":"Sun Jan 29 21:57:05 +0000 2
012","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":null,"contributors_enabled":false,"is_translator":false,"profile_background_color":"
C0DEED","profile_background_image_url":"http:₩/₩/abs.twimg.com₩/images₩/themes₩/theme1₩/bg.png","profile_background_image_url_https":"https:₩/₩/abs.tw
img.com₩/images₩/themes₩/theme1₩/bg.png","profile_background_tile":true,"profile_link_color":"0084B4","profile_sidebar_border_color":"FFFFFF","profile
_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:₩/₩/pbs.twimg.com₩/profile_i
mages₩/1230638521188151296₩/m5TfBFsc_normal.jpg","profile_image_url_https":"https:₩/₩/pbs.twimg.com₩/profile_images₩/1230638521188151296₩/m5TfBFsc_nor
mal.jpg","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinat
es":null,"place":{"id":"d98e7ce217ade2c5","url":"https:₩/₩/api.twitter.com₩/1.1₩/geo₩/id₩/d98e7ce217ade2c5.json","place_type":"city","name":"Stockton"
,"full_name":"Stockton, CA","country_code":"US","country":"United States","bounding_box":{"type":"Polygon","coordinates":[[[-121.416872,37.883347],[-1
21.416872,38.078305],[-121.183979,38.078305],[-121.183979,37.883347]]]},"attributes":{}},"contributors":null,"is_quote_status":false,"quote_count":0,"
reply_count":0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted"
:false,"filter_level":"low","lang":"en","timestamp_ms":"1588990012090"}♪
,
{"created_at":"Sat May 09 02:06:52 +0000 2020","id":1258941476031000576,"id_str":"1258941476031000576","text":"@christinelu ₩u201cAs opposed to me, tr
ying my best to sell out and it₩u2019s taking forever₩u201d","display_text_range":[13,83],"source":"₩u003ca href=₩"http:₩/₩/twitter.com₩/download₩/iph
one₩" rel=₩"nofollow₩"₩u003eTwitter for iPhone₩u003c₩/a₩u003e","truncated":false,"in_reply_to_status_id":1258908562564108288,"in_reply_to_status_id_st
r":"1258908562564108288","in_reply_to_user_id":7782442,"in_reply_to_user_id_str":"7782442","in_reply_to_screen_name":"christinelu","user":{"id":142786
08,"id_str":"14278608","name":"Jeff Yang","screen_name":"originalspin","location":"Los Angeles, CA","url":"http:₩/₩/www.cnn.com₩/profiles₩/jeff-yang",
"description":"@CNNOpinion Skyler & @HudsonDYang's dad #TheyCallUsBruce w₩/@angryasianman http:₩/₩/bit.ly₩/listen2TCUB #BLACKWOMENLEAD₩ud83d₩udc47 htt
p:₩/₩/www.higherheightsforamerica.org","translator_type":"none","protected":false,"verified":true,"followers_count":39827,"friends_count":5656,"listed
_count":796,"favourites_count":42523,"statuses_count":70647,"created_at":"Tue Apr 01 22:28:01 +0000 2008","utc_offset":null,"time_zone":null,"geo_enab
led":true,"lang":null,"contributors_enabled":false,"is_translator":false,"profile_background_color":"000000","profile_background_image_url":"http:₩/₩/
abs.twimg.com₩/images₩/themes₩/theme15₩/bg.png","profile_background_image_url_https":"https:₩/₩/abs.twimg.com₩/images₩/themes₩/theme15₩/bg.png","profi
le_background_tile":false,"profile_link_color":"6CB8B2","profile_sidebar_border_color":"02050A","profile_sidebar_fill_color":"000000","profile_text_co
lor":"B3B3B3","profile_use_background_image":true,"profile_image_url":"http:₩/₩/pbs.twimg.com₩/profile_images₩/1024828702310330368₩/HblftLMQ_normal.jp
g","profile_image_url_https":"https:₩/₩/pbs.twimg.com₩/profile_images₩/1024828702310330368₩/HblftLMQ_normal.jpg","profile_banner_url":"https:₩/₩/pbs.t
wimg.com₩/profile_banners₩/14278608₩/1560470587","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"no
tifications":null},"geo":null,"coordinates":null,"place":{"id":"714789cf3b7a50d0","url":"https:₩/₩/api.twitter.com₩/1.1₩/geo₩/id₩/714789cf3b7a50d0.jso
n","place_type":"city","name":"Ladera Heights","full_name":"Ladera Heights, CA","country_code":"US","country":"United States","bounding_box":{"type":"
Polygon","coordinates":[[[-118.391088,33.976320],[-118.391088,34.014937],[-118.357614,34.014937],[-118.357614,33.976320]]]},"attributes":{}},"contribu
tors":null,"is_quote_status":false,"quote_count":0,"reply_count":0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user_ment
ions":[{"screen_name":"christinelu","name":"Christine","id":7782442,"id_str":"7782442","indices":[0,12]}],"symbols":[]},"favorited":false,"retweeted":
false,"filter_level":"low","lang":"en","timestamp_ms":"1588990012544"}♪
,
{"created_at":"Sat May 09 02:06:52 +0000 2020","id":1258941477230534657,"id_str":"1258941477230534657","text":"I want beignets ₩ud83d₩ude2d₩ud83e₩udd2
4","source":"₩u003ca href=₩"http:₩/₩/twitter.com₩/download₩/iphone₩" rel=₩"nofollow₩"₩u003eTwitter for iPhone₩u003c₩/a₩u003e","truncated":false,"in_re
ply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user
":{"id":1151906509158268934,"id_str":"1151906509158268934","name":"T Kardasshian ₩u2728","screen_name":"TyoftheStorm","location":null,"url":null,"desc
ription":"24, Lukewarm mom ₩ud83e₩udd2a₩u2728  I₩u2019m here to talk shit, you can stfu.","translator_type":"none","protected":false,"verified":false,
"followers_count":437,"friends_count":727,"listed_count":1,"favourites_count":4355,"statuses_count":6204,"created_at":"Thu Jul 18 17:28:07 +0000 2019"
,"utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":null,"contributors_enabled":false,"is_translator":false,"profile_background_color":"F5F8
FA","profile_background_image_url":"","profile_background_image_url_https":"","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_s
idebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_u
rl":"http:₩/₩/pbs.twimg.com₩/profile_images₩/1232387164605804544₩/iA2-IU9W_normal.jpg","profile_image_url_https":"https:₩/₩/pbs.twimg.com₩/profile_ima
```

System in action



```
[{"created_at":"Sat May 09 02:06:52 +0000 2020","id":1258941474126753792,"id_str":"1258941474126753792","tex
profile_background_color":"C0DEED","profile_background_image_url":"http:₩/₩/abs.twimg.com₩/images₩/themes
"Polygon","coordinates":[[[-121.416872,37.883347],[-121.416872,38.078305],[-121.183979,38.078305],[-121.18397

,{"created_at":"Sat May 09 02:06:52 +0000 2020","id":1258941476031000576,"id_str":"1258941476031000576","tex
friends_count":5656,"listed_count":796,"favourites_count":42523,"statuses_count":70647,"created_at":"Tue Apr 01 2
false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":{"id":"714789

,{"created_at":"Sat May 09 02:06:52 +0000 2020","id":1258941477230534657,"id_str":"1258941477230534657","tex
ile_background_color":"F5F8FA","profile_background_image_url":"","profile_background_image_url_https":"","profile
"coordinates":[[[-118.668404,33.704538],[-118.668404,34.337041],[-118.155409,34.337041],[-118.155409,33.70453

,{"created_at":"Sat May 09 02:06:54 +0000 2020","id":1258941483035418624,"id_str":"1258941483035418624","tex
:1,"favourites_count":17462,"statuses_count":10032,"created_at":"Sun Jul 30 23:40:35 +0000 2017","utc_offset":null,
low_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":{"id":"3b77caf94bfc81fe","url":"https:

,{"created_at":"Sat May 09 02:06:55 +0000 2020","id":1258941487095537665,"id_str":"1258941487095537665","tex
round_image_url":"http:₩/₩/abs.twimg.com₩/images₩/themes₩/theme1₩/bg.png","profile_background_image_ur
try":"United States","bounding_box":{"type":"Polygon","coordinates":[[[-118.668404,33.704538],[-118.668404,34.33
on":"United States","url":null,"description":"Fuck it","translator_type":"none","protected":false,"verified":false,"follow
ile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coor

,{"created_at":"Sat May 09 02:06:55 +0000 2020","id":1258941487418490880,"id_str":"1258941487418490880","tex
5,"created_at":"Sun Feb 06 09:54:11 +0000 2011","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":null,"
oordinates":null,"place":{"id":"3b77caf94bfc81fe","url":"https:₩/₩/api.twitter.com₩/1.1₩/geo₩/id₩/3b77caf94bfc81

,{"created_at":"Sat May 09 02:06:55 +0000 2020","id":1258941487577886720,"id_str":"1258941487577886720","tex
id":305774740,"id_str":"305774740","name":"₩ub098₩ub77c₩uc9c0₩ud0a4₩uace0","screen_name":"chongjuhigh","
ors_enabled":false,"is_translator":false,"profile_background_color":"ACDED6","profile_background_image_url":"http:
lace_type":"city","name":"Los Angeles","full_name":"Los Angeles, CA","country_code":"US","country":"United States"

,{"created_at":"Sat May 09 02:06:56 +0000 2020","id":1258941491465957377,"id_str":"1258941491465957377","tex
```

Crawled data

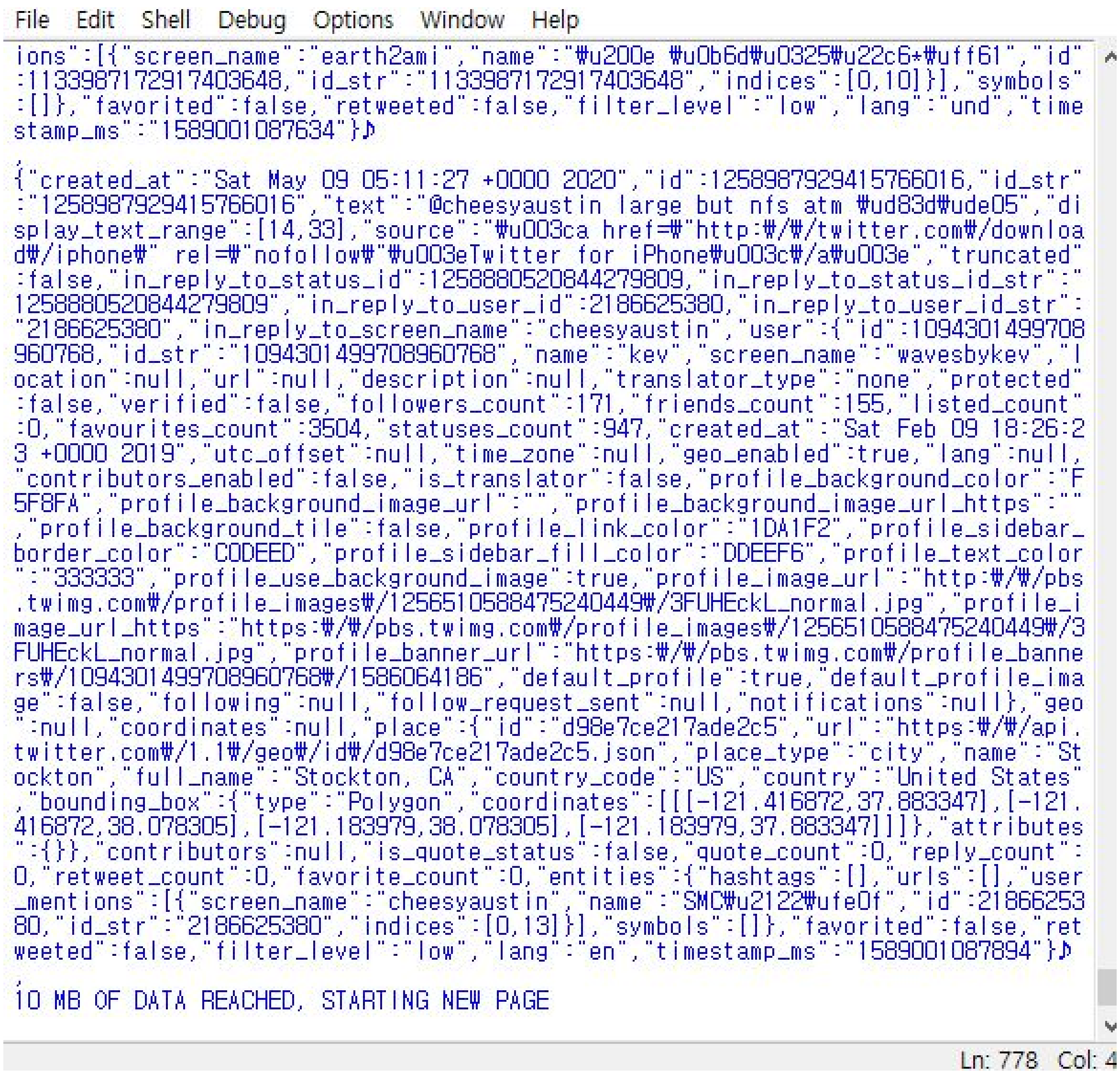


```
{
    "created_at": "Sat May 09 02:06:55 +0000 2020",
    "id": 1258941487095537665,
    "id_str": "1258941487095537665",
    "text": "The fools keeping things shutdown will crush your dreams.",
    "source": "₩u003ca href=₩"http:₩/₩/twitter.com₩/download₩/android₩" rel=₩"nofollow₩"₩u003eTwitter for Android₩u003c₩/a₩u003e",
    "truncated": false,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
        "id": 136729053,
        "id_str": "136729053",
        "name": "Dano",
        "screen_name": "Danoelan",
        "location": "Eastchester by LAX",
        "url": null,
```

Results

```
Valid JSON
```

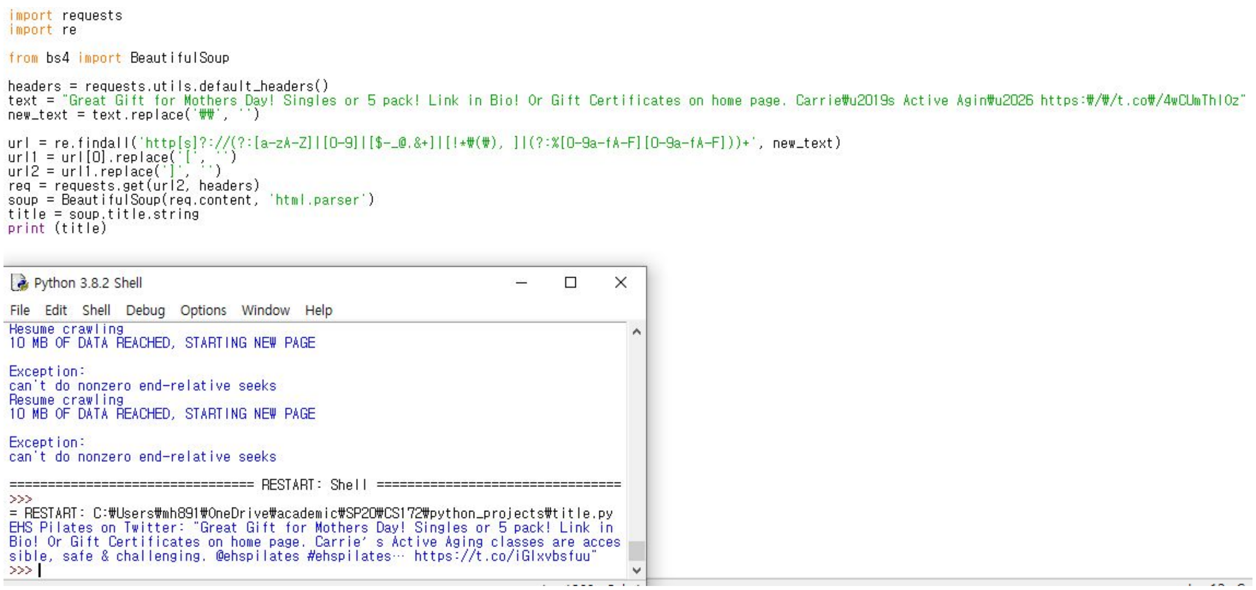
Data validation



```
ions":[{"screen_name":"earth2ami","name":"₩u200e ₩u0b6d₩u0325₩u22c6*₩uff61","id"
:1133987172917403648,"id_str":"1133987172917403648","indices":[0,10]}],"symbols"
:[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"und","time
stamp_ms":"1589001087634"}♪

,
{"created_at":"Sat May 09 05:11:27 +0000 2020","id":1258987929415766016,"id_str"
:"1258987929415766016","text":"@cheesyaustin large but nfs atm ₩ud83d₩ude05","di
splay_text_range":[14,33],"source":"₩u003ca href=₩"http:₩/₩/twitter.com₩/downloa
d₩/iphone₩" rel=₩"nofollow₩"₩u003eTwitter for iPhone₩u003c₩/a₩u003e","truncated"
:false,"in_reply_to_status_id":1258880520844279809,"in_reply_to_status_id_str":"
1258880520844279809","in_reply_to_user_id":2186625380,"in_reply_to_user_id_str":
"2186625380","in_reply_to_screen_name":"cheesyaustin","user":{"id":1094301499708
960768,"id_str":"1094301499708960768","name":"kev","screen_name":"wavesbykev","l
ocation":null,"url":null,"description":null,"translator_type":"none","protected"
:false,"verified":false,"followers_count":171,"friends_count":155,"listed_count"
:0,"favourites_count":3504,"statuses_count":947,"created_at":"Sat Feb 09 18:26:2
3 +0000 2019","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":null,
"contributors_enabled":false,"is_translator":false,"profile_background_color":"F
5F8FA","profile_background_image_url":"","profile_background_image_url_https":""
,"profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sidebar_
border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color
":"333333","profile_use_background_image":true,"profile_image_url":"http:₩/₩/pbs
.twimg.com₩/profile_images₩/1256510588475240449₩/3FUHEckL_normal.jpg","profile_i
mage_url_https":"https:₩/₩/pbs.twimg.com₩/profile_images₩/1256510588475240449₩/3
FUHEckL_normal.jpg","profile_banner_url":"https:₩/₩/pbs.twimg.com₩/profile_banne
rs₩/1094301499708960768₩/1586064186","default_profile":true,"default_profile_ima
ge":false,"following":null,"follow_request_sent":null,"notifications":null},"geo
":null,"coordinates":null,"place":{"id":"d98e7ce217ade2c5","url":"https:₩/₩/api.
twitter.com₩/1.1₩/geo₩/id₩/d98e7ce217ade2c5.json","place_type":"city","name":"St
ockton","full_name":"Stockton, CA","country_code":"US","country":"United States"
,"bounding_box":{"type":"Polygon","coordinates":[[[-121.416872,37.883347],[-121.
416872,38.078305],[-121.183979,38.078305],[-121.183979,37.883347]]]},"attributes
":{}},"contributors":null,"is_quote_status":false,"quote_count":0,"reply_count":
0,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[],"user
_mentions":[{"screen_name":"cheesyaustin","name":"SMC₩u2122₩ufe0f","id":21866253
80,"id_str":"2186625380","indices":[0,13]}],"symbols":[]},"favorited":false,"ret
weeted":false,"filter_level":"low","lang":"en","timestamp_ms":"1589001087894"}♪

,
```

```
10 MB OF DATA REACHED, STARTING NEW PAGE
```

Reached 10 MB of data, starting another page



```
import requests
import re

from bs4 import BeautifulSoup

headers = requests.utils.default_headers()
text = "Great Gift for Mothers Day! Singles or 5 pack! Link in Bio! Or Gift Certificates on home page. Carrie₩u2019s Active Agin₩u2026 https:₩/₩/t.co₩/4wCUmThl0z"
new_text = text.replace('₩₩', '')

url = re.findall('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*₩(₩), ]|(?:%[0-9a-fA-F][0-9a-fA-F]))+', new_text)
url1 = url[0].replace('[', '')
url2 = url1.replace(']', '')
req = requests.get(url2, headers)
soup = BeautifulSoup(req.content, 'html.parser')
title = soup.title.string
print (title)
```

```
Hesume crawling
10 MB OF DATA REACHED, STARTING NEW PAGE

Exception:
can't do nonzero end-relative seeks
Resume crawling
10 MB OF DATA REACHED, STARTING NEW PAGE

Exception:
can't do nonzero end-relative seeks

================================ RESTART: Shell ================================
>>>
= RESTART: C:₩Users₩mh891₩OneDrive₩academic₩SP20₩CS172₩python_projects₩title.py
EHS Pilates on Twitter: "Great Gift for Mothers Day! Singles or 5 pack! Link in
Bio! Or Gift Certificates on home page. Carrie’ s Active Aging classes are acces
sible, safe & challenging. @ehspilates #ehspilates… https://t.co/iGlxvbsfuu"
>>> 
```

Title crawler for a single json object element.