IP REPORT 03

A static html file is available for download here for the better reading experience in a browser. Highly recommend.

1 Recap with New DB Tables

Continue from last report, where I ended with DataBase Table design. I've made slight modifications since. I shall paste below the results of mySQL quiries after a few runs, which aligns with what I committed to my repository.

1.1 tag_like

```
mysql> select * from tag_like;
+----------------------------+---------+
| tagname                    | numLike |
+----------------------------+---------+
| animalcrossing             | 5335247 |
| automaticreferencecounting |       1 |
| copyonwrite                |       1 |
| daskapital                 |   12441 |
| rwby                       |  794138 |
| splatoon                   | 1043544 |
+----------------------------+---------+
6 rows in set (0.00 sec)
```

1.2 tag_toppost

```
mysql> select * from tag_toppost;
+----------------------------+-------------+
| tagname                    | postId      |
+----------------------------+-------------+
| animalcrossing             | CMSUj3eJ4wD |
| animalcrossing             | CMU41grp-W0 |
| animalcrossing             | CMUan5tJf38 |
| animalcrossing             | CMUbprgJAGO |
| animalcrossing             | CMUhVi8pkXn |
| animalcrossing             | CMUmB7TpBvA |
| animalcrossing             | CMV2haDpvZe |
| animalcrossing             | CMVBd7wJlGr |
| animalcrossing             | CMVChBXJ9UX |
| automaticreferencecounting | B1v3vpajGk_ |
| copyonwrite                | BuoQQmVnpa5 |
| daskapital                 | B_0LxT6AJgk |
| daskapital                 | B_qf5ZdpQOw |
| daskapital                 | B8OjRF-JR51 |
| daskapital                 | BifMI1zgAUd |
| daskapital                 | BrVjIXMARQ9 |
| daskapital                 | CAcoy-DIPF3 |
| daskapital                 | CDmEdrejJnK |
| daskapital                 | CEIsHK1Alhu |
| rwby                       | CL21DQ7FMAj |
| rwby                       | CL5Q-gwlTqO |
| rwby                       | CMAVu62gPDr |
| rwby                       | CMEHrUdrTeQ |
| rwby                       | CMITG8ulXvp |
| rwby                       | CMK_-imAftu |
| rwby                       | CMStEdJBUW4 |
| rwby                       | CMSUcxTpjNf |
| rwby                       | CMVkbzwBu9O |
| splatoon                   | CL-B72BJxOO |
| splatoon                   | CL-B72BJxQO |
| splatoon                   | CL2Ma8AJv7a |
| splatoon                   | CMF6DZYFs5l |
| splatoon                   | CMFabycp3_v |
| splatoon                   | CMOBd82JKYQ |
| splatoon                   | CMQaFi4JgmS |
| splatoon                   | CMRfS9KJYG3 |
| splatoon                   | CMU0sHGJWcw |
| splatoon                   | CMVC-7OJQVR |
+----------------------------+-------------+
37 rows in set (0.00 sec)
```

1.3 toppost_info

```
mysql> select * from toppost_info;
+-------------+---------+------------+------------+
| postId      | numLike | numComment | pdate      |
+-------------+---------+------------+------------+
| B_0LxT6AJgk |    1898 |         44 | 2020-05-05 |
| B_qf5ZdpQOw |     265 |         13 | 2020-05-01 |
| B1v3vpajGk_ |       0 |          0 | 2019-08-29 |
| B8OjRF-JR51 |     169 |         40 | 2020-02-06 |
| BifMI1zgAUd |     137 |          7 | 2018-05-07 |
| BrVjIXMARQ9 |     117 |          3 | 2018-12-13 |
| BuoQQmVnpa5 |       5 |          0 | 2019-03-05 |
| CAcoy-DIPF3 |     116 |          3 | 2020-05-21 |
| CDmEdrejJnK |    4969 |        127 | 2020-08-07 |
| CEIsHK1Alhu |     111 |          6 | 2020-08-21 |
| CL-B72BJxQO |    1910 |         18 | 2021-03-03 |
| CL21DQ7FMAj |     915 |         44 | 2021-03-01 |
| CL2Ma8AJv7a |    4197 |         99 | 2021-02-28 |
| CL5Q-gwlTqO |    3412 |        124 | 2021-03-01 |
| CMAVu62gPDr |    1467 |          4 | 2021-03-04 |
| CMEHrUdrTeQ |     544 |         26 | 2021-03-06 |
| CMF6DZYFs5l |     336 |         14 | 2021-03-06 |
| CMFabycp3_v |    7133 |        283 | 2021-03-06 |
| CMITG8ulXvp |     409 |          9 | 2021-03-07 |
| CMK_-imAftu |     411 |          3 | 2021-03-08 |
| CMOBd82JKYQ |    3015 |         49 | 2021-03-10 |
| CMQaFi4JgmS |     611 |          7 | 2021-03-10 |
| CMRfS9KJYG3 |    1511 |         74 | 2021-03-11 |
| CMStEdJBUW4 |    1390 |         16 | 2021-03-11 |
| CMSUcxTpjNf |    1571 |         39 | 2021-03-11 |
| CMSUj3eJ4wD |    1143 |         18 | 2021-03-11 |
| CMU0sHGJWcw |    2559 |         30 | 2021-03-12 |
| CMU41grp-W0 |    9307 |         29 | 2021-03-12 |
| CMUan5tJf38 |     212 |         52 | 2021-03-12 |
| CMUbprgJAGO |    4444 |         64 | 2021-03-12 |
| CMUhVi8pkXn |    1506 |         49 | 2021-03-12 |
| CMUmB7TpBvA |    3639 |         11 | 2021-03-12 |
| CMV2haDpvZe |   12419 |         31 | 2021-03-13 |
| CMVBd7wJlGr |    5026 |        135 | 2021-03-12 |
| CMVC-7OJQVR |    1187 |         29 | 2021-03-12 |
| CMVChBXJ9UX |    5366 |         31 | 2021-03-12 |
| CMVkbzwBu9O |    1558 |         43 | 2021-03-12 |
+-------------+---------+------------+------------+
37 rows in set (0.00 sec)
```

1.4 tag_tag

```
mysql> select * from tag_tag;
+----------------------------+-----------------------------+
| tagname                    | alttag                      |
+----------------------------+-----------------------------+
| animalcrossing             | acnh                        |
| animalcrossing             | acnhaesthetics              |
| animalcrossing             | acnhcodes                   |
| animalcrossing             | acnhcommunity               |
| animalcrossing             | acnhdesigns                 |
| animalcrossing             | acnhdesigns                 |
| animalcrossing             | acnhdreamcode               |
| animalcrossing             | acnhidea                    |
| animalcrossing             | acnhinspo                   |
| animalcrossing             | acnhisland                  |
| animalcrossing             | acpc                        |
| animalcrossing             | AMIIBO                      |
| animalcrossing             | animalcrossing              |
| animalcrossing             | animalcrossingaddict        |
| animalcrossing             | animalcrossingcommunity     |
| animalcrossing             | animalcrossingdesigns       |
| animalcrossing             | animalcrossingedit          |
| animalcrossing             | animalcrossingfandom        |
| animalcrossing             | animalcrossingideas         |
| animalcrossing             | animalcrossinginspiration   |
| animalcrossing             | animalcrossinginspo         |
| animalcrossing             | animalcrossinginspos        |
| animalcrossing             | animalcrossingnewhorizons   |
| animalcrossing             | animalcrossingnewleaf       |
| animalcrossing             | animalcrossingpocketcamp    |
| animalcrossing             | animalcrossingqrcodes       |
| animalcrossing             | animalcrossingswitch        |
| animalcrossing             | aninalcrossingnewhorizons   |
| animalcrossing             | bamboo                      |
| animalcrossing             | chinese                     |
| animalcrossing             | cityfolk                    |
| animalcrossing             | crossingcreations           |
| animalcrossing             | gamer                       |
| animalcrossing             | gaming                      |
| animalcrossing             | happyyhorizons              |
| animalcrossing             | horizonsinspo               |
| animalcrossing             | instagamer                  |
| animalcrossing             | island                      |
| animalcrossing             | japan                       |
| animalcrossing             | jungle                      |
| animalcrossing             | mario                       |
| animalcrossing             | newhorizons                 |
| animalcrossing             | Newleaf                     |
| animalcrossing             | nintendo                    |
| animalcrossing             | nintendoswitch              |
| animalcrossing             | pocketcamp                  |
| animalcrossing             | red                         |
| animalcrossing             | spring                      |
| animalcrossing             | switch                      |
| animalcrossing             | theme                       |
| animalcrossing             | tropical                    |
| animalcrossing             | どうぶつの森               |
| animalcrossing             | 動物森友會                 |
| automaticreferencecounting | Arc                         |
| automaticreferencecounting | AutomaticReferenceCounting  |
| automaticreferencecounting | Blog                        |
| automaticreferencecounting | InnovationM                 |
| automaticreferencecounting | Swift                       |
| automaticreferencecounting | SwiftMemoryManagement       |
| automaticreferencecounting | Technical                   |
| copyonwrite                | apple                       |
| copyonwrite                | copyonwrite                 |
| copyonwrite                | cow                         |
| copyonwrite                | google                      |
| copyonwrite                | güvenlikaçığı               |
| copyonwrite                | macos                       |
| copyonwrite                | projectzero                 |
| daskapital                 | 1demaio                     |
| daskapital                 | artspace                    |
| daskapital                 | berlinkreuzberg             |
| daskapital                 | berlinstagram               |
| daskapital                 | bibliophile                 |
| daskapital                 | bookaholic                  |
| daskapital                 | bookish                     |
| daskapital                 | bookphotography             |
| daskapital                 | books                       |
| daskapital                 | bookshelf                   |
| daskapital                 | booksofinstagram            |
| daskapital                 | bookstagram                 |
| daskapital                 | bookstagrammer              |
| daskapital                 | bookworm                    |
| daskapital                 | catherineroseevans          |
| daskapital                 | contemporaryart             |
| daskapital                 | dasartyberlin               |
| daskapital                 | daskapital                  |
| daskapital                 | DE                          |
| daskapital                 | deepblue                    |
| daskapital                 | deutscheweine               |
| daskapital                 | deutschland                 |
| daskapital                 | diadotrabalhador            |
| daskapital                 | drawing                     |
| daskapital                 | edebiütopya                 |
| daskapital                 | freidrichengels             |
| daskapital                 | groupshow                   |
| daskapital                 | hegel                       |
| daskapital                 | Kapital                     |
| daskapital                 | karlmarx                    |
| daskapital                 | kitap                       |
| daskapital                 | kitapiyikivar               |
| daskapital                 | kitapkurdu                  |
| daskapital                 | kitaplar                    |
| daskapital                 | mammalia                    |
| daskapital                 | markers                     |
| daskapital                 | marx                        |
| daskapital                 | maximingrünhaus             |
| daskapital                 | may1st                      |
| daskapital                 | mylife                      |
| daskapital                 | nonfiction                  |
| daskapital                 | nonfictionbooks             |
| daskapital                 | okudumbitti                 |
| daskapital                 | okumak                      |
| daskapital                 | okumakgüzeldir              |
| daskapital                 | philosophy                  |
| daskapital                 | politicaleconomy            |
| daskapital                 | reader                      |
| daskapital                 | readersofig                 |
| daskapital                 | readersofinstagram          |
| daskapital                 | reading                     |
| daskapital                 | reads                       |
| daskapital                 | relapse                     |
| daskapital                 | schaufenster                |
| daskapital                 | schaufensterberlin          |
| daskapital                 | tbt                         |
| daskapital                 | temporaryexhibition         |
| daskapital                 | thelibraryofdidsus          |
| daskapital                 | thephenomenologyofspirit    |
| daskapital                 | trier                       |
| daskapital                 | voyeur                      |
| daskapital                 | weingut                     |
| daskapital                 | window                      |
| daskapital                 | windowshopping              |
| daskapital                 | Капитал                     |
| daskapital                 | Карл_Маркс                  |
| rwby                       | animeedit                   |
| rwby                       | animeedits                  |
| rwby                       | art                         |
| rwby                       | blakebelladonna             |
| rwby                       | digiartist                  |
| rwby                       | digitalart                  |
| rwby                       | Disney                      |
| rwby                       | drawing                     |
| rwby                       | emeraldsustari              |
| rwby                       | emeraldsustrai              |
| rwby                       | emercury                    |
| rwby                       | fanart                      |
| rwby                       | generalironwood             |
| rwby                       | ironwood                    |
| rwby                       | jaunearc                    |
| rwby                       | lafirechickenart            |
| rwby                       | lieren                      |
| rwby                       | lierenedit                  |
| rwby                       | Neo                         |
| rwby                       | Neopolitan                  |
| rwby                       | norarwby                    |
| rwby                       | noravalkyrie                |
| rwby                       | noravalkyrieedit            |
| rwby                       | noravalkyrierwby            |
| rwby                       | pyrrhanikos                 |
| rwby                       | renora                      |
| rwby                       | roosterteeth                |
| rwby                       | rubyrose                    |
| rwby                       | rwby                        |
| rwby                       | rwby8                       |
| rwby                       | rwbyart                     |
| rwby                       | rwbyblake                   |
| rwby                       | rwbyedit                    |
| rwby                       | rwbyedits                   |
| rwby                       | rwbyemerald                 |
| rwby                       | rwbyfan                     |
| rwby                       | rwbyfanart                  |
| rwby                       | rwbyironwood                |
| rwby                       | rwbyneo                     |
| rwby                       | rwbyneopolitan              |
| rwby                       | rwbynora                    |
| rwby                       | rwbyren                     |
| rwby                       | rwbyrenora                  |
| rwby                       | rwbyrubyrose                |
| rwby                       | rwbyv8                      |
| rwby                       | rwbyvol8                    |
| rwby                       | rwbyvolume8                 |
| rwby                       | rwbyweiss                   |
| rwby                       | rwbyyang                    |
| rwby                       | teamjnpr                    |
| rwby                       | teamrwby                    |
| rwby                       | weissschnee                 |
| rwby                       | yangxiaolong                |
| splatoon                   | ａｅｓｔｈｅｔｉｃ          |
| splatoon                   | animegirl                   |
| splatoon                   | artist                      |
| splatoon                   | artistoninstgram            |
| splatoon                   | artistsoninstagram          |
| splatoon                   | cartoonstyle                |
| splatoon                   | characterconcept            |
| splatoon                   | characterdesign             |
| splatoon                   | conceptart                  |
| splatoon                   | conceptartwork              |
| splatoon                   | cute                        |
| splatoon                   | cuteart                     |
| splatoon                   | digitalart                  |
| splatoon                   | digitaldrawing              |
| splatoon                   | fanartdrawing               |
| splatoon                   | inkling                     |
| splatoon                   | inklingoc                   |
| splatoon                   | kawaii                      |
| splatoon                   | nintendo                    |
| splatoon                   | nintendoart                 |
| splatoon                   | nintendoswitch              |
| splatoon                   | nintendoswitchlite          |
| splatoon                   | oc                          |
| splatoon                   | originalcharacter           |
| splatoon                   | punkgirls                   |
| splatoon                   | splatoon                    |
| splatoon                   | splatoon2                   |
| splatoon                   | splatoon2art                |
| splatoon                   | splatoon3                   |
| splatoon                   | splatoonart                 |
| splatoon                   | splatoonartwork             |
| splatoon                   | splatooncharacterart        |
| splatoon                   | splatoonfanart              |
| splatoon                   | splatoonidol                |
| splatoon                   | splatoonoc                  |
| splatoon                   | splatoonurchin              |
| splatoon                   | tenshi                      |
+----------------------------+----------------------------+
222 rows in set (0.00 sec)
```

2 Updated Python Scrapper

It is difficult to continue adding more logics and coding blocks when it hits to 200 lines. I have to start thinking refactor the codespairer and abstract functions to other auxiliary files/modules. However, I didn't get enough time to plan it through to make Classes. The structure is at most usable, I found some coupling issues which can only be solved by large restructuring.

2.1 Create new modules

In the `mainLogic.py`, current dependencies as below:

```
# -*- coding: utf-8 -*-

import time
from datetime import datetime, date

import mysql.connector
from mysql.connector import errorcode

from webreader import get_tagpage_json
from jparser import total_like_of
from jparser import create_top9infolist
from db_operation import create_database
from db_operation import create_tables
from db_operation import insert_tag_like
from db_operation import insert_tag_toppost
from db_operation import insert_toppost_info
from db_operation import insert_tag_tag
```

New modules `webreader.py`, `jparser.py`, `db_operation.py` were created to abstract the function definitions from the main logic body. So that logic is cleaner to just make a bunch of function calls. (The parameters passing is not designed to my satisfaction yet, not easy to read)

`webreader.py` provides `get_tagpage_json` to abstract away the getting webpage parts of logic, now it simplified to:

```
jsonstr = get_tagpage_json(tag_url)
```

Which contains the info that we really need and filter out the extra info that we are not interested in.

Then for the parsing, `jparser.py` offers various functions to retrieve the detailed items that we'd like to store into database.

```
total_like_count : int = total_like_of(jsonstr) # not elegent
top9infolist = create_top9infolist(jsonstr)
```

To make the certain data consolidate together, a array of dictionary `top9infolist` is created by `jparser.py` which contains the info of the top 9 posts, each post can have `"postId"`, `"numLike"`, `"numComment"` and `"tags"` mentioned in the caption.

```
    for jdict in toppost_dicts["edges"]:
        postId = jdict["node"]["shortcode"]
        numLike = str(jdict["node"]["edge_liked_by"]["count"])
        numComment = str(jdict["node"]["edge_media_to_comment"]["count"])
        date = datetime.utcfromtimestamp(jdict["node"]["taken_at_timestamp"]).strftime('%Y-%m-%

        comment = jdict["node"]["edge_media_to_caption"]["edges"][0]["node"]["text"]
        tags = re.findall('\B#\w\w+', comment)

        newdict = {}
        newdict["postId"] = postId
        newdict["numLike"] = numLike
        newdict["numComment"] = numComment
        newdict["date"] = date
        newdict["tags"] = tags

        returndictlist.append(newdict)
```

These attributes are ultilized in the functions in the `db_operation.py` to store into relevant tables.

`db_operation.py` also has a single function `create_tables` for creating all the tables needed in our `explore` database.

So our mainlogic of database operations are pretty much simplified to:

```
    cursor = conn.cursor()

    # Create Table 01:     `tag_like`
    create_tables(cursor)

    insert_tag_like(cursor, tag_name, total_like_count)

    insert_tag_toppost(cursor, tag_name, top9infolist)

    insert_toppost_info(cursor, top9infolist)

    insert_tag_tag(cursor, tag_name, top9infolist)


    conn.commit()
    cursor.close()
    conn.close()
```

2.2 Difficulties

A lot time spent on debug for SQL quiries. Often problems can be the table creation and data insertion. The error message is pretty useless which only indicats the location and says that SQL syntax error, so it’s difficult to debug. I spent an hour to find one miss bracket `)` in my DDL statement.

For some reasons, I spent days still couldn’t get it work for `ON DUPLICATE KEY UPDATE`:

```
        add_tag = ("INSERT INTO toppost_info "
            "(postId, numLike, numComment, pdate) "
            "VALUES (%(postId)s, %(numLike)s, %(numComment)s, %(pdate)s)")
#            "ON DUPLICATE KEY UPDATE"
#            "numLike = VALUES(numLike),"
#            "numComment = %s,"
#            "pdate = %s")
```

3 Future planning and challenges

DataBase is structured. I left one function called `filter_strategy` unimplemented, whoes responsibility is to filter out the unrelevant tags in tag_tag table. The tag_tag table stores all the tags mentioned in post caption section of the top 9 posts. Some of them are repeated or their popularity is even higher than the original searched hashtag which is less useful to us. By design certain strategy or algorism, we can replace whatever inside by our liking without affecting other parts of the system.