Báo cáo môn học:
Lập trình xử lý dữ liệu với python (AIT2003_1)

I.Giới thiệu

Crawl dữ liệu từ Facebook là một phương pháp quan trọng để thu thập thông tin từ nguồn dữ liệu lớn này. Trong báo cáo này, chúng ta sẽ thảo luận về quy trình crawl dữ liệu từ Facebook bằng cách sử dụng các công cụ và ngôn ngữ lập trình python.

II. Mục tiêu

Thu thập dữ liệu và phân tích và trực quan hoá dữ liệu thu thập được bàng những kiến thức đã được học để hiểu rõ hơn về xu hướng, tương tác của người dùng.

III. Các bước thực hiện

1. Crawling data

Bước 1: Kết nối với google drive để lưu trữ dữ liệu và code:

```
from google.colab import drive
drive.mount('/content/drive')
```

Bước 2: Cài đặt thư viện facebook_scraper:

```
!pip install facebook-scraper
```

Bước 3: Import thư viện cần thiết để crawl dữ liệu

```
from facebook_scraper import get_posts
import pandas as pd
import time
```

Bước 4: Khởi tạo tên page cần lấy, đường dẫn để lữu file dữ liệu và đường dẫn của file cookies:

```
FANPAGE_LINK ="MarinLiekuriva"
FOLDER_PATH = "/content/drive/MyDrive/IAI_Python_final_project"
COOKIE_PATH = "/content/drive/MyDrive/IAI_Python_final_project/www.facebook.com_cookies (5).txt"

PAGES_NUMBER = 20
```

Bước 5: Bắt đầu crawl dữ liệu

```
post_list = []
for post in get_posts(FANPAGE_LINK,
                      options={"comments": True, "reactions": True, "allow_extra_requests": True},
                      extra_info=True, pages=PAGES_NUMBER, cookies=COOKIE_PATH):

    print(post)
    time.sleep(10)
    post_list.append(post)
```

Sử dụng time.sleep(10) để tạm dừng 10s sau mỗi lần lấy 1 post tránh việc bị ban account, tuy nhiên thời gian crawl sẽ lâu hơn.

Bước 6: Xuất dữ liệu đã crawl được thành file csv và lưu tại đường dẫn đã khởi tạo ở trên:

```python
# Initialize dataframe to scrape Facebook post
post_df_full = pd.DataFrame(columns=post_list[0].keys(), index=range(len(post_list)), data=post_list)

# To df
path=FOLDER_PATH + '/' + FANPAGE_LINK + ".csv"
post_df_full.to_csv(path, index=False)
print(path)
```

Link data:

https://drive.google.com/file/d/1kQounC-wbW4D2Rfjlr5priRxY4T573tj/view?usp=drive_link

2. Data cleansing and preprocessing

Bước 1: Đọc và xem thông tin của dữ liệu thô:

Đọc dữ liệu thô:

```python
raw_data = pd.read_csv('/content/drive/MyDrive/IAI_Python_final_project/MarinLiekuriva.csv')
raw_data.info()
```

Thông tin của dữ liệu thô:

0	post_id	200 non-null	int64
1	text	114 non-null	object
2	post_text	113 non-null	object
3	shared_text	7 non-null	object
4	original_text	17 non-null	object
5	time	200 non-null	object
6	timestamp	200 non-null	int64
7	image	189 non-null	object
8	image_lowquality	200 non-null	object
9	images	198 non-null	object
10	images_description	198 non-null	object
11	images_lowquality	200 non-null	object
12	images_lowquality_description	200 non-null	object
13	video	7 non-null	object
14	video_duration_seconds	0 non-null	float64
15	video_height	0 non-null	float64
16	video_id	7 non-null	float64
17	video_quality	0 non-null	float64
18	video_size_MB	0 non-null	float64
19	video_thumbnail	7 non-null	object
20	video_watches	0 non-null	float64
21	video_width	0 non-null	float64
22	likes	200 non-null	int64
23	comments	200 non-null	int64
24	shares	200 non-null	int64
25	post_url	200 non-null	object
26	link	15 non-null	object
27	links	199 non-null	object
28	user_id	200 non-null	int64
29	username	200 non-null	object
30	user_url	200 non-null	object
31	is_live	200 non-null	bool
32	factcheck	0 non-null	float64
33	shared_post_id	15 non-null	float64
34	shared_time	15 non-null	object
35	shared_user_id	15 non-null	float64
36	shared_username	15 non-null	object
37	shared_post_url	15 non-null	object
38	available	200 non-null	bool
39	comments_full	200 non-null	object
40	reactors	200 non-null	object
41	w3_fb_url	200 non-null	object
42	reactions	200 non-null	object
43	reaction_count	200 non-null	int64
44	with	18 non-null	object
45	page_id	200 non-null	int64
46	sharers	0 non-null	float64
47	image_id	185 non-null	float64
48	image_ids	200 non-null	object
49	was_live	200 non-null	bool
50	fetched_time	200 non-null	object

Ta thu dược bộ dữ liệu với 50 trường dữ liệu và thấy được có những trường dữ liệu bị thiếu thông tin và kiểu dữ liệu của những trường dữ liệu

Bước 2: Kiểm trra sự thiếu hụt của dữ liệu:

Kết quả trả về:


{'post_id': 0.0,
'text': 0.43,
'post_text': 0.435,
'shared_text': 0.965,
'original_text': 0.915,
'time': 0.0,
'timestamp': 0.0,
'image': 0.055,
'image_lowquality': 0.0,
'images': 0.01,
'images_description': 0.01,
'images_lowquality': 0.0,
'images_lowquality_description': 0.0,
'video': 0.965,
'video_duration_seconds': 1.0,
'video_height': 1.0,
'video_id': 0.965,
'video_quality': 1.0,
'video_size_MB': 1.0,
'video_thumbnail': 0.965,
'video_watches': 1.0,
'video_width': 1.0,
'likes': 0.0,
'comments': 0.0,
'shares': 0.0,
'post_url': 0.0,
'link': 0.925,
'links': 0.005,
'user_id': 0.0,
'username': 0.0,
'user_url': 0.0,
'is_live': 0.0,
'factcheck': 1.0,
'shared_post_id': 0.925,
'shared_time': 0.925,
'shared_user_id': 0.925,
'shared_username': 0.925,
'shared_post_url': 0.925,
'available': 0.0,
'comments_full': 0.0,
'reactors': 0.0,
'w3_fb_url': 0.0,
'reactions': 0.0,
'reaction_count': 0.0,
'with': 0.91,
'page_id': 0.0,
'sharers': 1.0,
'image_id': 0.075,
'image_ids': 0.0,
'was_live': 0.0,
'fetched_time': 0.0}

Có thể thấy có một số trường dữ liệu thiếu hoặc không có dữ liệu nhưng cũng có một số trường dữ liệu có đầy đủ thông tin. Ta sẽ lọc ra những trường dữ liệu cần thiết cho việc phân tích và xem xét mất mát ở những trường đó.

Bước 3: Lọc ra trường dữ liệu cần thiết:

```
1 data = raw_data[['post_id', 'text', 'time', 'image', 'reactors', 'reactions', 'reaction_count', 'comments', 'comments_full', 'shares', 'was_live', 'fetched_time']]
```

Ta sẽ lấy ra những trường cần thiết để phục vụ việc phân tích.

Bước 4: Kiểm tra sự thiếu hụt dữ liệu ở những trường đã được lọc ra:

```
1 missing_data(data)
```

```
{'post_id': 0.0,
 'text': 0.0,
 'time': 0.0,
 'image': 0.055,
 'reactors': 0.0,
 'reactions': 0.0,
 'reaction_count': 0.0,
 'comments': 0.0,
 'comments_full': 0.0,
 'shares': 0.0,
 'was_live': 0.0,
 'fetched_time': 0.0}
```

Đa số những trường cần thiết đều có đầy đủ thông tin, trường dữ liệu text và image có thể chấp nhận được do có những bài đăng không kèm theo caption và có những bài đăng không kèm theo ảnh.

Bước 5: Kiểm tra và chuẩn hoá dữ liệu:

Kiểm tra kiểu dữ liệu của các trường dữ liệu:

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   post_id         200 non-null    int64
 1   text            200 non-null    string
 2   time            200 non-null    object
 3   image           189 non-null    object
 4   reactors        200 non-null    object
 5   reactions       200 non-null    object
 6   reaction_count  200 non-null    int64
 7   comments        200 non-null    int64
 8   comments_full   200 non-null    object
 9   shares          200 non-null    int64
 10  was_live        200 non-null    object
 11  fetched_time    200 non-null    object
dtypes: int64(4), object(7), string(1)
memory usage: 18.9+ KB
```

Biểu diễn dữ liệu dưới dạng bảng:

```
1 data.head()
```

	post_id	text	time	image	reactors	reactions	reaction_count	comments	shares	was_live
0	920510076113642	Bwonya	2023-11-29 03:43:41	https://m.facebook.com/photo/view_full_size/?f...	[]	{'thích': 124, 'yêu thích': 117, 'haha': 2, 'w...	247	15	22	0
1	920052726159377		2023-11-28 05:26:20	https://scontent-lga3-1.xx.fbcdn.net/v/t39.308...	[{'name': 'Nhân Trương', 'link': 'https://face...	{'thích': 785, 'yêu thích': 806, 'haha': 25, '...	1652	64	91	0
2	919508826213767	#HonkaiStarRail	2023-11-27 03:10:08	https://scontent-lga3-1.xx.fbcdn.net/v/t39.308...	[{'name': 'Nhân Trương', 'link': 'https://face...	{'thích': 1218, 'yêu thích': 32, 'haha': 2174,...	3450	60	312	0
3	918486696315980	Kallen watching HoTr debut ;))\n\nMarin Liekur...	2023-11-25 02:21:39	NaN	[{'name': 'Đào Tấn Hòa', 'link': 'https://face...	{'thích': 415, 'yêu thích': 11, 'haha': 841, '...	1289	28	40	0
4	918453912985925	Hello, I'm just opened my ko-fi shop to sell t...	2023-11-25 00:58:42	NaN	[{'name': 'New Eyes', 'link': 'https://faceboo...	{'thích': 812, 'yêu thích': 464, 'haha': 19, '...	1384	27	33	0

Ta sẽ loại bỏ những giá trị NaN trong trường dữ liệu text:

```
for i in range(len(data)):
  if type(data['text'][i]) == float:
    data['text'][i] = ''
```

Biến những giá trị NaN về thành string rỗng.

Bước 6: Xuất file dữ liệu sau xử lý

```
path = '/content/drive/MyDrive/IAI_Python_final_project' + '/clean_data.csv'
data.to_csv(path, index=False)
```

3. Data analysis

Trước tiên ta sẽ phân tích caption trong mỗi post của trang:

Có thể thấy những từ khoá như Honkai Impact, HonkaiStarRail và BlueArchive nên người theo dõi của trang chủ yếu là những người chơi của những tựa game gacha nổi tiếng. Ta còn thấy được tên của những nền tảng mạng xã hội khác như twitter, pixiv.

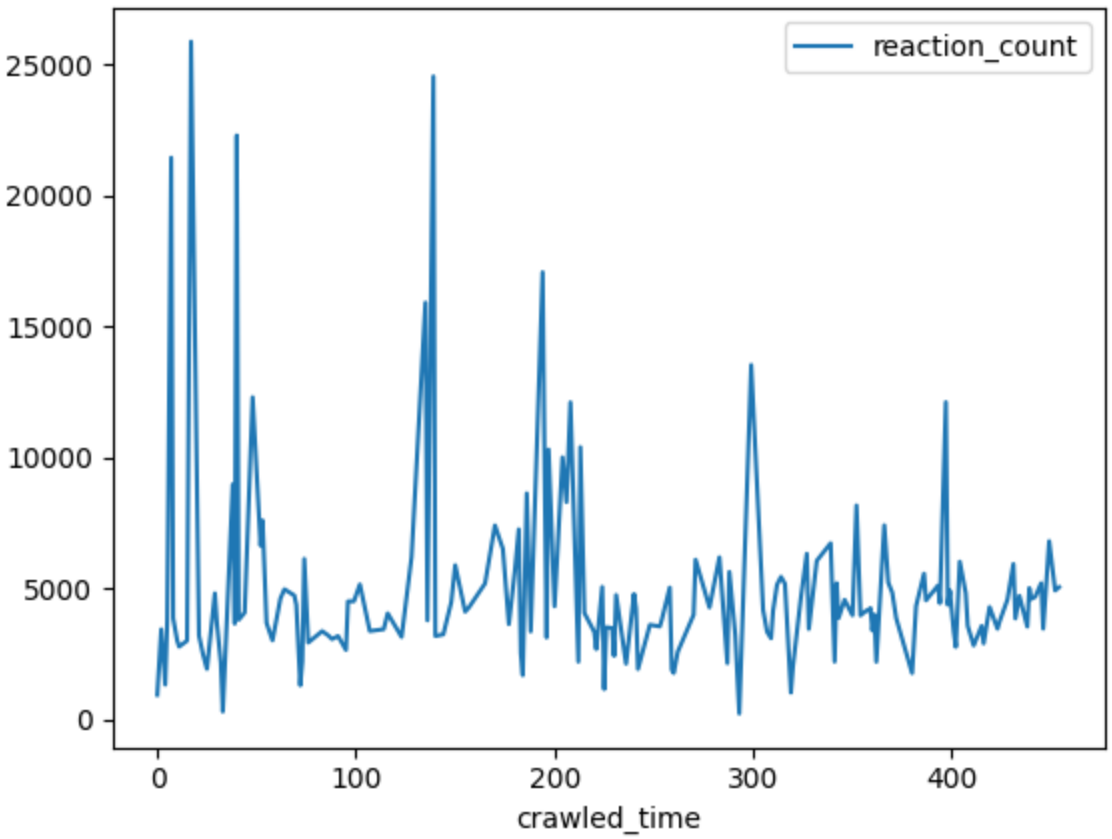
Tiếp theo ta sẽ tính thời gian từ lúc crawl dữ liệu đến lúc bài đăng được đăng lên facebook:

```
1 data['crawled_time'] = data['fetched_time'].astype(np.datetime64) - data['time'].astype(np.datetime64)
```

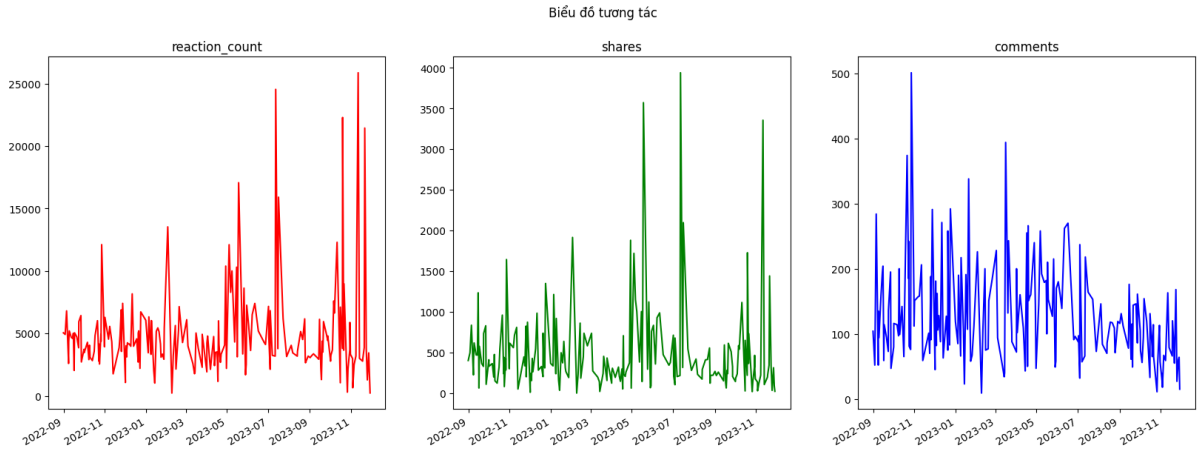
```
1 data['crawled_time'].describe()
```

```
count                            200
mean     230 days 00:59:21.875612748
std      141 days 08:42:48.915199358
min        0 days 00:13:35.449123
25%       95 days 23:40:20.151922250
50%      231 days 06:31:26.082445500
75%      358 days 23:39:29.588896752
max        454 days 16:10:31.062803
Name: crawled_time, dtype: object
```

Bài viết lâu nhất có thể lấy được cách đây 454 ngày trước. Nên ta sẽ đánh giá lượng tương tác trung bình của trang trong khoảng 1 năm trở lại đây:



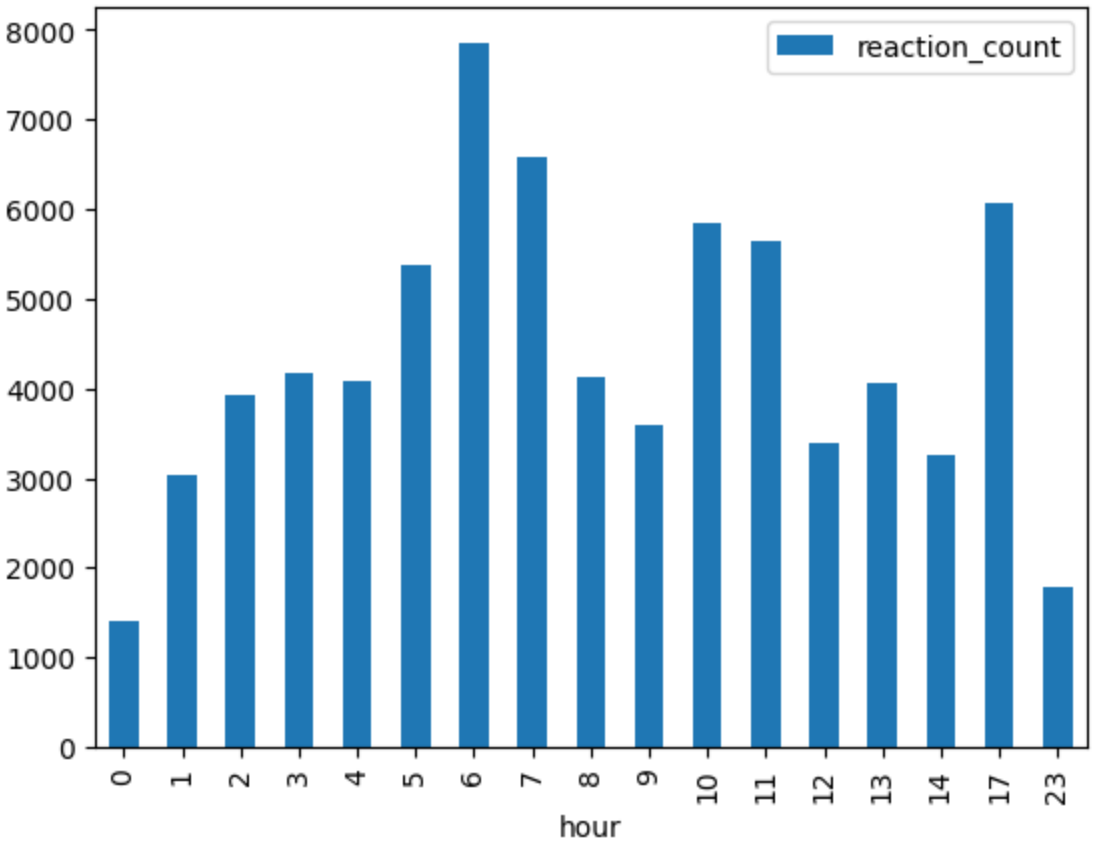
Trong khoảng 150 ngày trở lại đây trang có sự tăng vượt bậc về lượng tương tác. Để đánh giá xem lượng tương tác về cảm xúc đó có phải là ảo hay là dùng tool thì ta sẽ đánh giá cả comment với lượng người share:



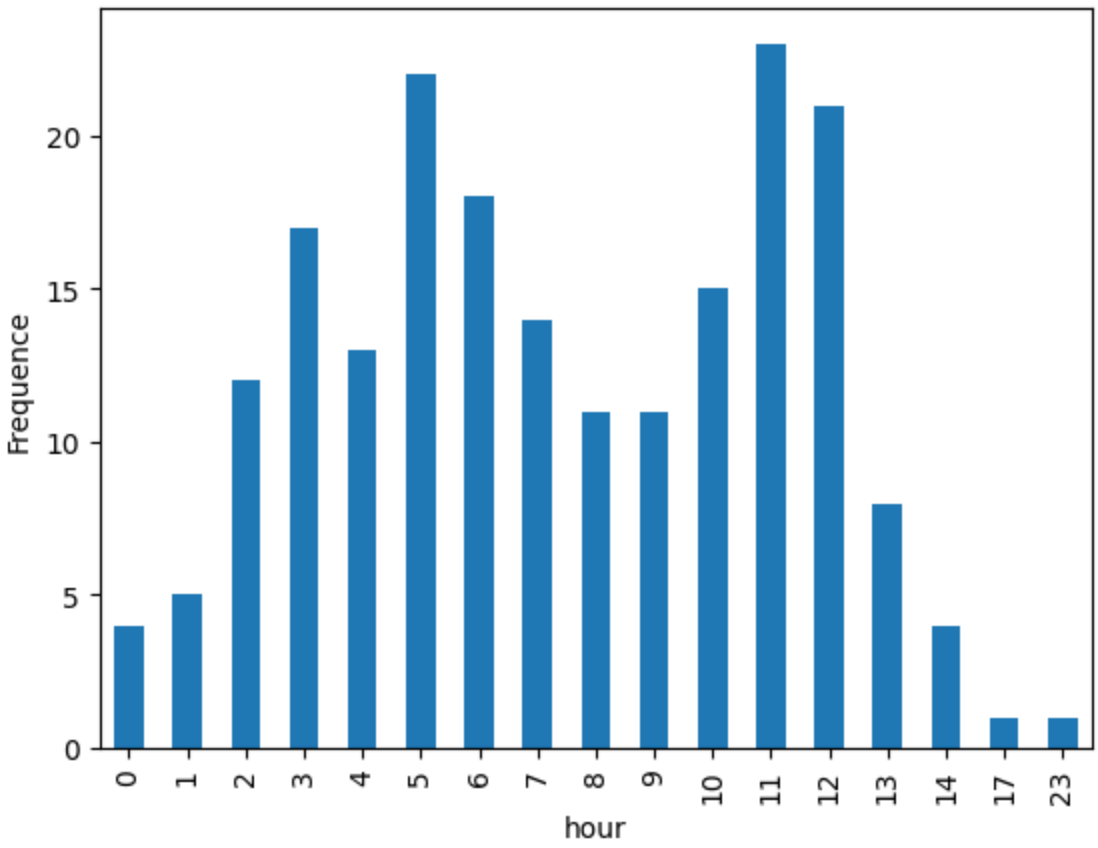
Biểu đồ số lượng reaction và số lượng share có sự tương đồng khá lớn và có hình dáng khá giống nhau còn biểu đồ số lượng comment có xu hướng giảm dần. Có thể

tạm kết luận lượng tương tác của trang là thật.

Tiếp theo ta sẽ xem xét sự tương quan giữa số lượng reaction và thời gian đăng bài trong ngày:

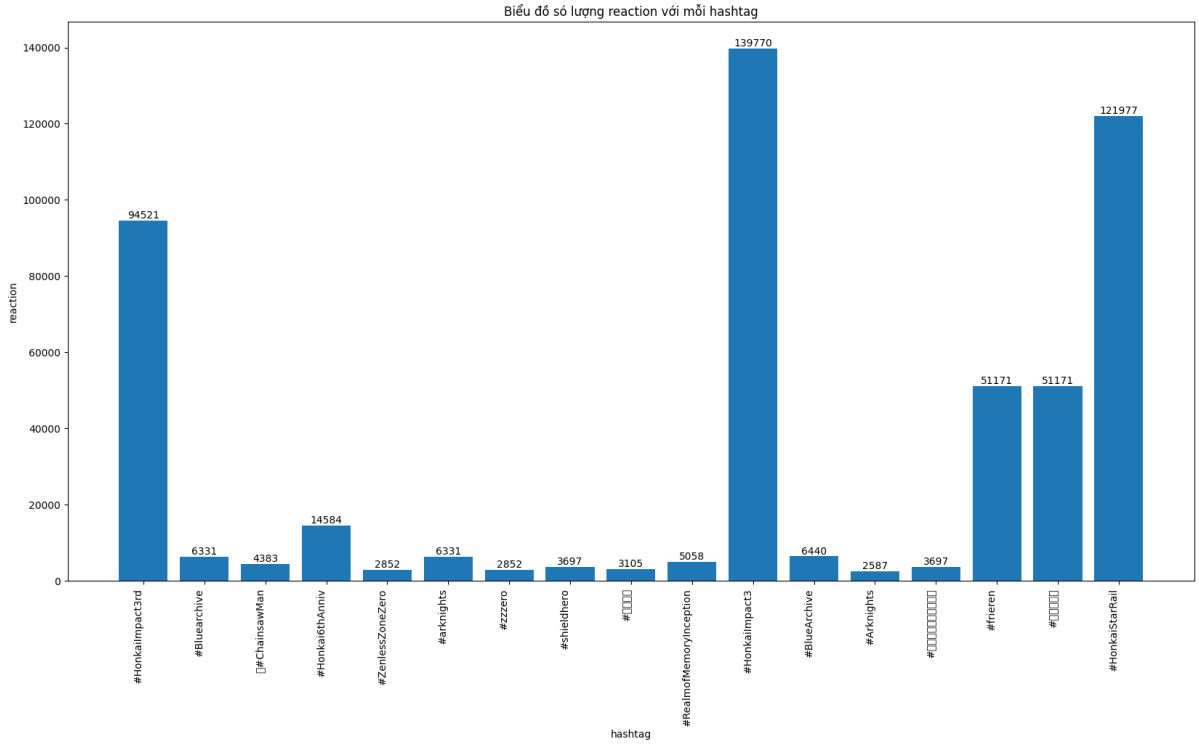


Ta sẽ xem mốc thời gian mà trang thường xuyên đăng bài:



Page thường đăng bài chủ yếu vào lúc 5 giờ sáng, 11 giờ trưa và 12 giờ trưa nhưng lượng reaction chủ yếu lại rơi vào lúc 6, 7 giờ sáng và 10, 11 giờ trưa.

Do đây là trang của 1 artist nên trong mỗi bài viết thường kèm theo hashtag tên game, ta sẽ xem những hashtag thường xuyên xuất hiện và lượng reaction của từng hashtag:



Thấy được lượng reaction cao nhất thuộc về HonkaiImpact3, Honkai: Star Rail và frieren.

4. Conclusion.

Có thể thấy page chủ yếu dành cho người chơi của những tựa game nổi tiếng như Honkai Impact 3, Honkai: Star Rail, Blue Archive, …. Do sự ra mắt của tựa game Honkai: Star Rail ra mắt vào tháng 4 năm 2023 nên từ tháng 5 đổ đi lượng tương tác của page tăng cao hơn hẳn, điều đó được thể hiện qua bảng xếp hạng reaction dựa theo hashtag ở trên.

IV. Kết luận

Crawl dữ liệu từ Facebook là một quá trình phức tạp và đòi hỏi sự cẩn trọng để đảm bảo việc thu thập thông tin một cách hiệu quả và hợp pháp. Việc tuân thủ các quy định và điều khoản sử dụng của Facebook là rất quan trọng để tránh các vấn đề pháp lý. Đồng thời, chúng ta cũng cần chú ý đến etika trong việc sử dụng dữ liệu thu thập được để đảm bảo tính minh bạch và công bằng trong quá trình nghiên cứu hoặc phân tích.