TRƯỜNG ĐẠI HỌC CÔNG NGHỆ - ĐHQG HN

KHOA TRÍ TUỆ NHÂN TẠO

BÁO CÁO
BÀI TẬP CUỐI KÌ

Giảng viên: TS. ĐẶNG TRẦN BÌNH

Trợ giảng: CN. Nguyễn Văn Phi

Lớp: K67 – AI1

Sinh viên: Trần Văn Dy_22022523

Đề bài: Phân tích tương tác và nội dung của một/ nhiều tài khoản Facebook

Giới thiệu

Trong thực tế hiện nay việc khai thác dữ liệu từ Facebook là một việc vô cùng cần thiết và có ứng dụng rất rộng rãi. Nhiều công ty truyền thông tại Việt Nam đã tích lũy kinh nghiệm trong việc thu thập dữ liệu từ Facebook để đánh giá tác động của chương trình quảng cáo và tiếp thị. Các doanh nghiệp này thường thực hiện việc cào dữ liệu để tổng hợp thông tin và đo lường mức độ ảnh hưởng của họ. Một số công ty khác thì bán giải pháp trích xuất thông tin từ Facebook với đa dạng các nội dung có thể thu thập được. Hoặc đơn giản nhất với sinh viên hoặc một số shop nhỏ họ thực hiện cào dữ liệu để theo dõi tình trạng phát triển của fanpage, xu hướng phát triển trong thời gian gần đây.

Việc cào dữ liệu FaceBook là một ý tưởng rất hay và cần thiết để nghiên cứu . Chính vì vậy trong bài tập này em đã thực hiện cào dữ liệu Facebook về để nghiên cứu và phân tích dựa vào các trường thông tin đã thu thập được. Do chưa có nhiều kinh nghiệm về làm báo cáo cũng như những hạn chế về mặt kiến thức , trong bài báo cáo chắc chắn cũng không thể tránh khỏi những thiếu sót . Em mong nhận được sự đóng góp , phê bình từ phía thầy dể báo cáo của em được hoàn thiện hơn .

Em xin chân thành cảm ơn thầy!

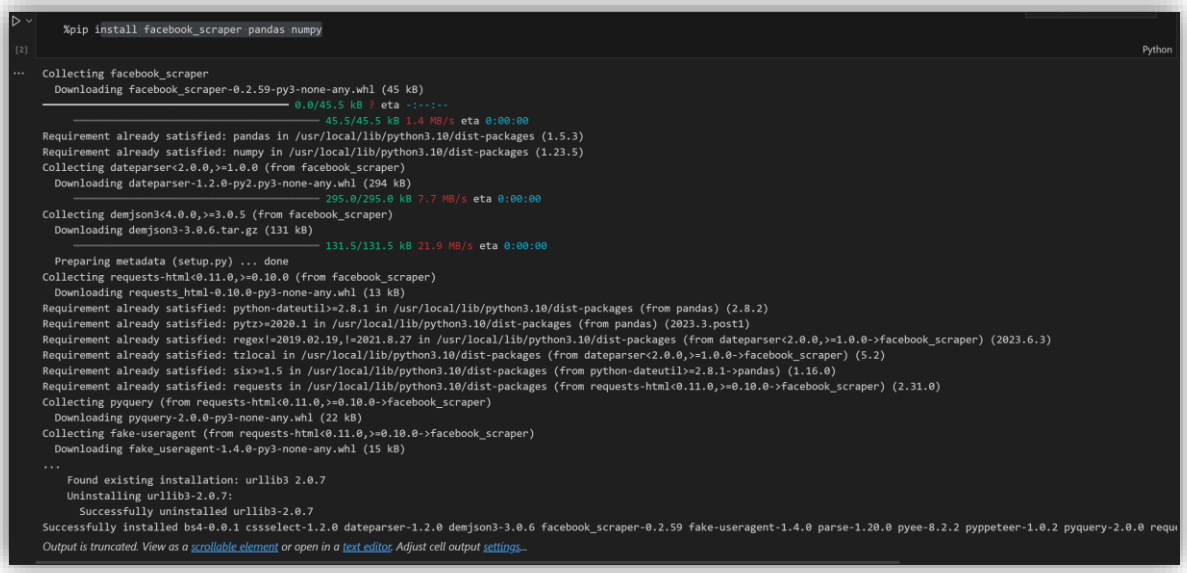
Nội dung

Phần 1: Thu thập dữ liệu

I, Yêu cầu về cài đặt

1.Thư viện dùng

Cài đặt thư viện facebook_scraper để cào thông tin các bài viết



2.File cookies.txt

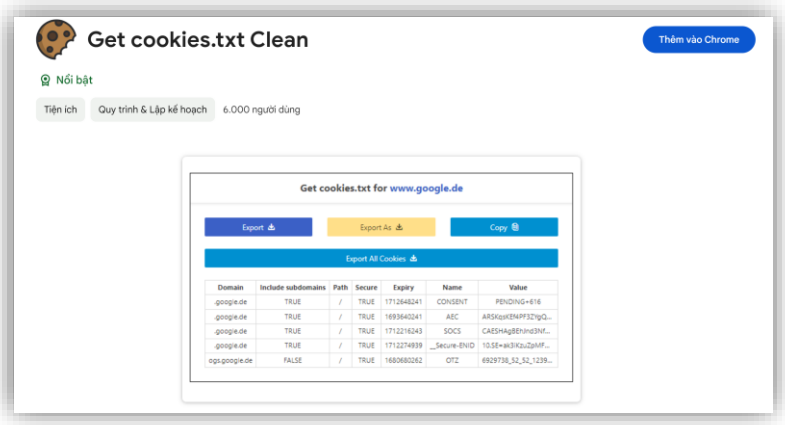
Trong quá trình craw data, việc sử dụng cookies có một số ứng dụng:

- Thu thập dữ liệu công khai: thông tin người dùng, bài đăng, hình ảnh và các thông tin công cộng khác
- Tích hợp với FB API: xác thực và truy cập vào dữ liệu API yêu cầu

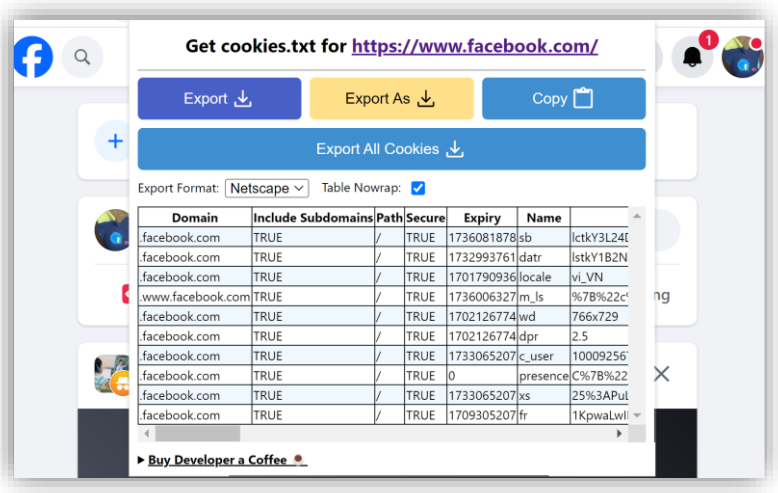
Các bước thực hiện:

Bước 1: Thêm tiện ích get cookies

Tại đây



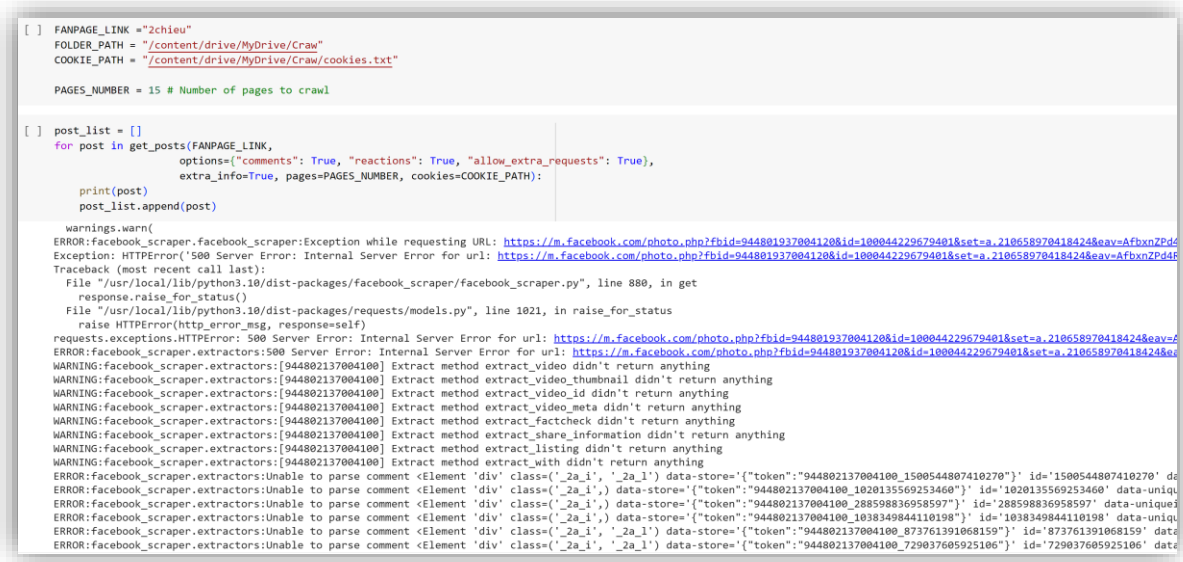
Bước 2: Mở facebook và download cookies



II, Cào dữ liệu

1. Thực hành

Ở bài này em cào dữ liệu của fanpage Hai Chiều trên google colab



```
[ ] FANPAGE_LINK ="2chieu"
    FOLDER_PATH = "/content/drive/MyDrive/Craw"
    COOKIE_PATH = "/content/drive/MyDrive/Craw/cookies.txt"

    PAGES_NUMBER = 15 # Number of pages to crawl

[ ] 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)
        post_list.append(post)

      warnings.warn(
    ERROR:facebook_scraper.facebook_scraper:Exception while requesting URL: https://m.facebook.com/photo.php?fbid=944801937004120&id=100044229679401&set=a.210658970418424&eav=AfbxnZPd4
    Exception: HTTPError('500 Server Error: Internal Server Error for url: https://m.facebook.com/photo.php?fbid=944801937004120&id=100044229679401&set=a.210658970418424&eav=AfbxnZPd4F
    Traceback (most recent call last):
      File "/usr/local/lib/python3.10/dist-packages/facebook_scraper/facebook_scraper.py", line 880, in get
        response.raise_for_status()
      File "/usr/local/lib/python3.10/dist-packages/requests/models.py", line 1021, in raise_for_status
        raise HTTPError(http_error_msg, response=self)
    requests.exceptions.HTTPError: 500 Server Error: Internal Server Error for url: https://m.facebook.com/photo.php?fbid=944801937004120&id=100044229679401&set=a.210658970418424&eav=A
    ERROR:facebook_scraper.extractors:500 Server Error: Internal Server Error for url: https://m.facebook.com/photo.php?fbid=944801937004120&id=100044229679401&set=a.210658970418424&ea
    WARNING:facebook_scraper.extractors:[944802137004100] Extract method extract_video didn't return anything
    WARNING:facebook_scraper.extractors:[944802137004100] Extract method extract_video_thumbnail didn't return anything
    WARNING:facebook_scraper.extractors:[944802137004100] Extract method extract_video_id didn't return anything
    WARNING:facebook_scraper.extractors:[944802137004100] Extract method extract_video_meta didn't return anything
    WARNING:facebook_scraper.extractors:[944802137004100] Extract method extract_factcheck didn't return anything
    WARNING:facebook_scraper.extractors:[944802137004100] Extract method extract_share_information didn't return anything
    WARNING:facebook_scraper.extractors:[944802137004100] Extract method extract_listing didn't return anything
    WARNING:facebook_scraper.extractors:[944802137004100] Extract method extract_with didn't return anything
    ERROR:facebook_scraper.extractors:Unable to parse comment <Element 'div' class=('_2a_i', '_2a_l') data-store='{"token":"944802137004100_1500544807410270"}' id='1500544807410270' da
    ERROR:facebook_scraper.extractors:Unable to parse comment <Element 'div' class=('_2a_i',) data-store='{"token":"944802137004100_1020135569253460"}' id='1020135569253460' data-uniqu
    ERROR:facebook_scraper.extractors:Unable to parse comment <Element 'div' class=('_2a_i',) data-store='{"token":"944802137004100_288598836958597"}' id='288598836958597' data-uniquei
    ERROR:facebook_scraper.extractors:Unable to parse comment <Element 'div' class=('_2a_i',) data-store='{"token":"944802137004100_1038349844110198"}' id='1038349844110198' data-uniqu
    ERROR:facebook_scraper.extractors:Unable to parse comment <Element 'div' class=('_2a_i', '_2a_l') data-store='{"token":"944802137004100_873761391068159"}' id='873761391068159' data
    ERROR:facebook_scraper.extractors:Unable to parse comment <Element 'div' class=('_2a_i', '_2a_l') data-store='{"token":"944802137004100_729037605925106"}' id='729037605925106' data
```

…

Số lượng bài là 150

```
[ ]  len(post_list)

     150
```

Bài đầu tiên

```
[ ]  post_list[0]

     {'post_id': '944802137004100',
      'text': 'Cuốn quá hong dứt ra được \U0001fae2',
      'post_text': 'Cuốn quá hong dứt ra được \U0001fae2',
      'shared_text': '',
      'original_text': None,
      'time': datetime.datetime(2023, 11, 28, 11, 12, 18),
      'timestamp': 1701169938,
      'image': None,
      'image_lowquality': 'https://scontent-ord5-2.xx.fbcdn.net/v/t39.30808-6/405182200_944801930337454_9185539492962934583_n.jpg?stp=cp0_dst-jpg_e15_p320x320_q65&_nc_cat=106&ccb=1-
     7&_nc_sid=ab7367&efg=eyJpIjoidCJ9&_nc_ohc=hfe7pQ-InXYAX9-ZHl1&_nc_ht=scontent-ord5-2.xx&oh=00_AfC_oAFIocWAtPQR0MlPDal4EWk3HrcsNIU0JNksHzoUzQ&oe=656E0FB5',
      'images': [],
      'images_description': [],
      'images_lowquality': ['https://scontent-ord5-2.xx.fbcdn.net/v/t39.30808-6/405182200_944801930337454_9185539492962934583_n.jpg?stp=cp0_dst-jpg_e15_p320x320_q65&_nc_cat=106&ccb=1-
     7&_nc_sid=ab7367&efg=eyJpIjoidCJ9&_nc_ohc=hfe7pQ-InXYAX9-ZHl1&_nc_ht=scontent-ord5-2.xx&oh=00_AfC_oAFIocWAtPQR0MlPDal4EWk3HrcsNIU0JNksHzoUzQ&oe=656E0FB5'],
      'images_lowquality_description': ["Có thể là hình minh họa về văn bản cho biết 'TÔI KHI CHẾ 1 BÀI NHẠC KÌ CỤC CŨNG LÀ TÔI CẢ NGÀY HÃM ĐÓ: Lần đầu tiên Trái thanh long CÓ trong mì
     tôm/ f/2u haichieu suzustudio SUZU STUDIO'"],
      'video': None,
      'video_duration_seconds': None,
      'video_height': None,
      'video_id': None,
      'video_quality': None,
      'video_size_MB': None,
      'video_thumbnail': None,
      'video_watches': None,
      'video_width': None,
      'likes': None,
      'comments': 2202,
      'shares': 671,
      'post_url': 'https://facebook.com/2chieu/posts/944802137004100',
      'link': None,
      'links': [{'link': '/story.php?story_fbid=pfbid033zo8V4bk4TByfeNejJqgrKcmMRhrtCeKzfcYZsNq65mTWc3iE9Fgtyp3qN77FCHrl&id=100044229679401&eav=AfbFQsV9-PFnF1yrGv1avBKrXHKbkrAovsR-
     06WiFC6fLmBI7JHGpxm2xbShYRjACSk&m_entstream_source=timeline&refid=17&_ft_=encrypted_tracking_data.0AY8MRcg1tJ-
     bZqr_gDN0oMQnDt61VERpA9QDwSleIpYN0SQN4WuRWhwR5m4vU6ox88GsAquh2v1ChHYyVmQ_jnV_-uRuO9TaTdDzmf0vH4OYUVuJ85fRnRlrB-
     DlwxUtftJumkmt2Wx7LM8tYxL11oe1xc5gOmgk3thwfTOFS0WQlpBGuNS8CrBPLQA1sUAoBLXt7RGjgEyhCB8VS31lLO0ZzDcwDYMG7BuITlEYJQmN9GJ0RutJ2lSBUP7kp-WqiHMVwRXZGlFq-U50UtzFNg1y-\\-
```

…

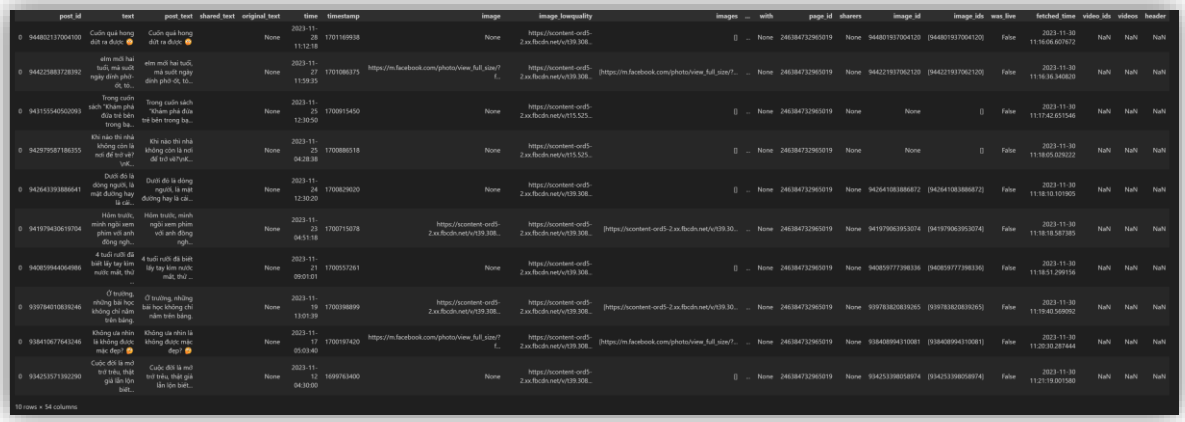
2. Tổ chức dữ liệu thô

```
[ ]  # Initialize dataframe to scrape Facebook post
     post_df_full = pd.DataFrame(columns = [])
     # Start to collect Facebook post data by facebook_scraper library
     for post in post_list:
         post_entry = post
         fb_post_df = pd.DataFrame.from_dict(post_entry, orient='index')
         fb_post_df = fb_post_df.transpose()
         post_df_full = post_df_full.append(fb_post_df)
         print(post['post_id']+' get')

     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
       post_df_full = post_df_full.append(fb_post_df)
     <ipython-input-16-f5ad45f1f0fe>:8: FutureWarning: The frame.append method is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
```

…

10 post đầu tiên:



	post_id	text	post_text	shared_text	original_text	time	timestamp	image	image_lowquality	images	...	with	page_id	sharers	image_id	image_ids	was_live	fetched_time	video_ids	videos	header
0	9448021370041 00	Cuốn quá hong dứt ra được 😂	Cuốn quá hong dứt ra được 😂		None	2023-11-28 11:12:18	1701169938	None	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[]	...	None	246384732965019	None	9448019370041 20	[9448019370041 20]	False	2023-11-30 11:16:06.607672	NaN	NaN	NaN
0	9442258837283 92	em mới hai tuổi, mà suốt ngày dính phớ- ớt, tô...	em mới hai tuổi, mà suốt ngày dính phớ- ớt, tô...		None	2023-11-27 11:59:35	1701086375	https://m.facebook.com/photo/view_full_size/?f...	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[https://m.facebook.com/photo/view_full_size/?...	...	None	246384732965019	None	9442219370621 20	[9442219370621 20]	False	2023-11-30 11:16:36.340820	NaN	NaN	NaN
0	9431555405020 93	Trong cuốn sách "Khám phá đứa trẻ bên trong ba...	Trong cuốn sách "Khám phá đứa trẻ bên trong ba...		None	2023-11-25 12:30:50	1700915450		https://scontent-ord5-2.xx.fbcdn.net/v/t15.525...	[]	...	None	246384732965019	None	None	[]	False	2023-11-30 11:17:42.651546	NaN	NaN	NaN
0	9429795871863 55	Khi nào thì nhà không còn là nơi để trở về? Việ...	Khi nào thì nhà không còn là nơi để trở về? Việ...		None	2023-11-25 04:28:38	1700886518		https://scontent-ord5-2.xx.fbcdn.net/v/t15.525...	[]	...	None	246384732965019	None	None	[]	False	2023-11-30 11:18:05.029222	NaN	NaN	NaN
0	9426433938866 41	Dưới đó là dòng người, là mặt đường hay là cái...	Dưới đó là dòng người, là mặt đường hay là cái...		None	2023-11-24 12:30:20	1700829020		https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[]	...	None	246384732965019	None	9426410838866 72	[9426410838866 72]	False	2023-11-30 11:18:10.101905	NaN	NaN	NaN
0	9419794306197 04	Hôm trước, mình ngồi xem phim với anh đồng nghi...	Hôm trước, mình ngồi xem phim với anh đồng nghi...		None	2023-11-23 04:51:18	1700715078	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[https://scontent-ord5-2.xx.fbcdn.net/v/t39.30...	...	None	246384732965019	None	9419790639530 74	[9419790639530 74]	False	2023-11-30 11:18:18.587385	NaN	NaN	NaN
0	9408599440649 86	4 tuổi rưỡi đã biết lấy tay kìm nước mắt, thủ ...	4 tuổi rưỡi đã biết lấy tay kìm nước mắt, thủ ...		None	2023-11-21 09:01:01	1700557261		https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[]	...	None	246384732965019	None	9408597777398 336	[9408597777398 336]	False	2023-11-30 11:18:51.299156	NaN	NaN	NaN
0	9397840108392 46	Ở trường, những bài học không chỉ nằm trên bảng.	Ở trường, những bài học không chỉ nằm trên bảng.		None	2023-11-19 13:01:39	1700398899	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[https://scontent-ord5-2.xx.fbcdn.net/v/t39.30...	...	None	246384732965019	None	9397838208392 65	[9397838208392 65]	False	2023-11-30 11:19:40.569092	NaN	NaN	NaN
0	9384106776432 46	Không ủa nhìn là không được mắc đẹp? 😂	Không ủa nhìn là không được mắc đẹp? 😂		None	2023-11-17 05:03:40	1700197420	https://m.facebook.com/photo/view_full_size/?f...	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[https://m.facebook.com/photo/view_full_size/?...	...	None	246384732965019	None	9384089943100 81	[9384089943100 81]	False	2023-11-30 11:20:30.287444	NaN	NaN	NaN
0	9342535713922 90	Cuộc đời là mở trí tiêu, thật giá lắm lộn biết...	Cuộc đời là mở trí tiêu, thật giá lắm lộn biết...		None	2023-11-12 04:30:00	1699763400		https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[]	...	None	246384732965019	None	9342533980589 74	[9342533980589 74]	False	2023-11-30 11:21:19.001580	NaN	NaN	NaN

10 rows × 54 columns

3. Lưu file

Lưu file dưới 2 dạng :

- Csv: để phân tích bài viết (nội dụng, tương tác,…)
- Npy: để phân tích bình luận và người bình luận

```
path1=FOLDER_PATH + FANPAGE_LINK + ".npy"
np.save(path1, arr)    # .npy extension is added if not given
print(path1)
```

```
/content/drive/MyDrive/Craw2chieu.npy
```

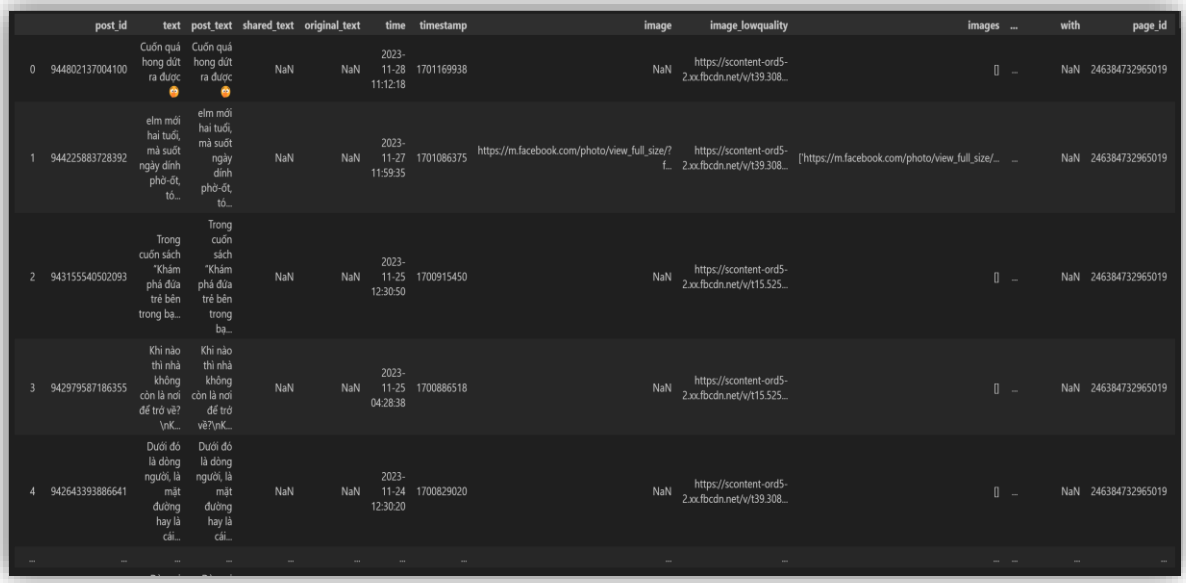
```
# To df
path=FOLDER_PATH + FANPAGE_LINK + "new.csv"
post_df_full.to_csv(path, index=False)
print(path)
```

```
/content/drive/MyDrive/Craw2chieunew.csv
```

Phần 2: Tiền xử lý dữ liệu

I. Bài viết

Dataframe trước



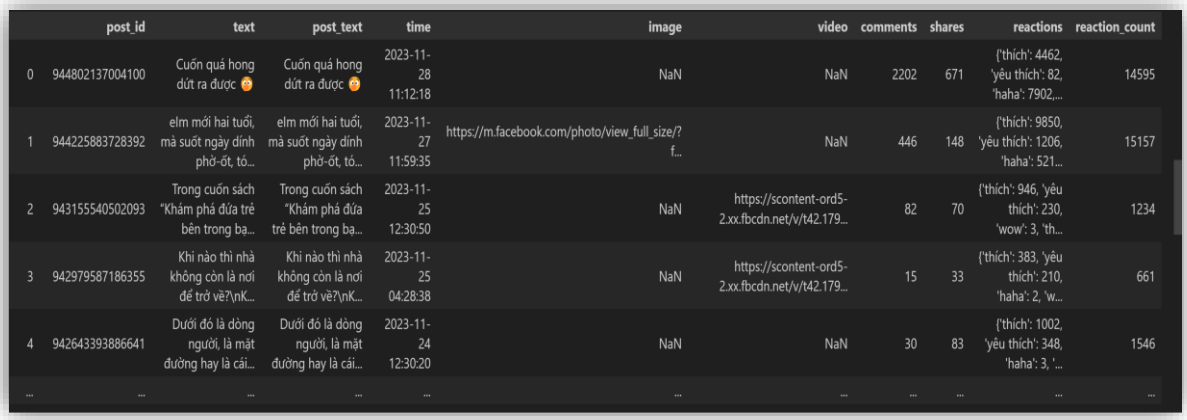
	post_id	text	post_text	shared_text	original_text	time	timestamp	image	image_lowquality	images	...	with	page_id
0	944802137004100	Cuốn quá hong dứt ra được 😍	Cuốn quá hong dứt ra được 😍	NaN	NaN	2023-11-28 11:12:18	1701169938	NaN	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[]	...	NaN	246384732965019
1	944225883728392	elm mới hai tuổi, mà suốt ngày dính phờ-ốt, tô...	elm mới hai tuổi, mà suốt ngày dính phờ-ốt, tô...	NaN	NaN	2023-11-27 11:59:35	1701086375	https://m.facebook.com/photo/view_full_size/?f...	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	['https://m.facebook.com/photo/view_full_size/...	...	NaN	246384732965019
2	943155540502093	Trong cuốn sách "Khám phá đứa trẻ bên trong ba...	Trong cuốn sách "Khám phá đứa trẻ bên trong ba...	NaN	NaN	2023-11-25 12:30:50	1700915450	NaN	https://scontent-ord5-2.xx.fbcdn.net/v/t15.525...	[]	...	NaN	246384732965019
3	942979587186355	Khi nào thì nhà không còn là nơi để trở về? \nK...	Khi nào thì nhà không còn là nơi để trở về?\nK...	NaN	NaN	2023-11-25 04:28:38	1700886518	NaN	https://scontent-ord5-2.xx.fbcdn.net/v/t15.525...	[]	...	NaN	246384732965019
4	942643393886641	Dưới đó là dòng người, là mặt đường hay là cái...	Dưới đó là dòng người, là mặt đường hay là cái...	NaN	NaN	2023-11-24 12:30:20	1700829020	NaN	https://scontent-ord5-2.xx.fbcdn.net/v/t39.308...	[]	...	NaN	246384732965019
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Thông tin

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 54 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   post_id                         150 non-null    int64
 1   text                            149 non-null    object
 2   post_text                       148 non-null    object
 3   shared_text                     4 non-null      object
 4   original_text                   0 non-null      float64
 5   time                            150 non-null    object
 6   timestamp                       150 non-null    int64
 7   image                           122 non-null    object
 8   image_lowquality                149 non-null    object
 9   images                          150 non-null    object
 10  images_description              150 non-null    object
 11  images_lowquality               150 non-null    object
 12  images_lowquality_description   150 non-null    object
 13  video                           3 non-null      object
 14  video_duration_seconds          0 non-null      float64
 15  video_height                    0 non-null      float64
 16  video_id                        3 non-null      float64
 17  video_quality                   0 non-null      float64
 18  video_size_MB                   0 non-null      float64
 19  video_thumbnail                 3 non-null      object
...
 52  videos                          6 non-null      object
 53  header                          4 non-null      object
dtypes: bool(3), float64(18), int64(7), object(26)
memory usage: 60.3+ KB
```

Nhìn qua chúng ta thấy nhiều cột bị thiếu dữ liệu(bài đăng không có nội dung, hạn chế quyền truy cập của token,..) vì thế để thuận tiện cho bước phân tích ta sẽ bỏ những cột không cần thiết, và cài lại giá trị những hàng hoặc cột nếu cần thiết

Dataframe sau



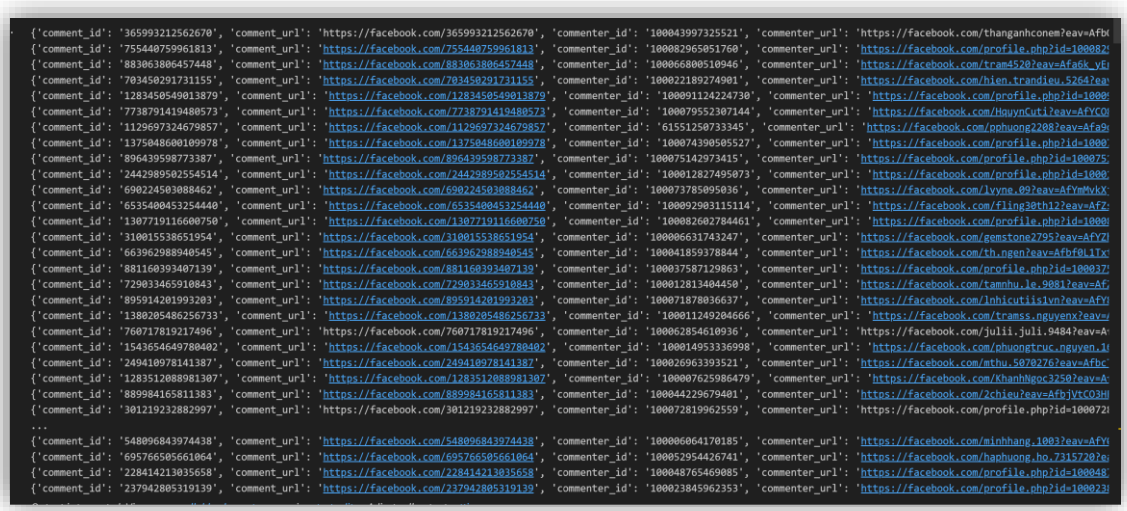
	post_id	text	post_text	time	image	video	comments	shares	reactions	reaction_count
0	944802137004100	Cuốn quá hong dứt ra được	Cuốn quá hong dứt ra được	2023-11-28 11:12:18	NaN	NaN	2202	671	{'thích': 4462, 'yêu thích': 82, 'haha': 7902,...	14595
1	944225883728392	elm mới hai tuổi, mà suốt ngày dính phờ-ốt, tó...	elm mới hai tuổi, mà suốt ngày dính phờ-ốt, tó...	2023-11-27 11:59:35	https://m.facebook.com/photo/view_full_size/?f...	NaN	446	148	{'thích': 9850, 'yêu thích': 1206, 'haha': 521...	15157
2	943155540502093	Trong cuốn sách "Khám phá đứa trẻ bên trong ba...	Trong cuốn sách "Khám phá đứa trẻ bên trong ba...	2023-11-25 12:30:50	NaN	https://scontent-ord5-2.xx.fbcdn.net/v/t42.179...	82	70	{'thích': 946, 'yêu thích': 230, 'wow': 3, 'th...	1234
3	942979587186355	Khi nào thì nhà không còn là nơi để trở về?\nK...	Khi nào thì nhà không còn là nơi để trở về?\nK...	2023-11-25 04:28:38	NaN	https://scontent-ord5-2.xx.fbcdn.net/v/t42.179...	15	33	{'thích': 383, 'yêu thích': 210, 'haha': 2, 'w...	661
4	942643393886641	Dưới đó là dòng người, là mặt đường hay là cái...	Dưới đó là dòng người, là mặt đường hay là cái...	2023-11-24 12:30:20	NaN	NaN	30	83	{'thích': 1002, 'yêu thích': 348, 'haha': 3, '...	1546
...	...	...	...	...	...	...	...	...	...	...

Thông tin

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   post_id         150 non-null    int64
 1   text            150 non-null    object
 2   post_text       150 non-null    object
 3   time            150 non-null    object
 4   image           150 non-null    int32
 5   video           150 non-null    int32
 6   comments        150 non-null    int64
 7   shares          150 non-null    int64
 8   reactions       150 non-null    object
 9   reaction_count  150 non-null    int64
dtypes: int32(2), int64(4), object(4)
memory usage: 10.7+ KB
```

II. Bình luận

Bình luận chưa xử lí là 1 phần trong file HaiChieu.npy



Sau xử lí

	commenter_id	commenter_name	comment_text	comment_time
0	100043997325521	Thắng Anh Con Em	Nghe riết cái thấy cũng thích thích hay hay 🤣	2023-11-23
1	100082965051760	Trang Lê	Bài gì z:))	2023-11-23
2	100066800510946	Bích Trâm	Nghe xong ăn mì 3 miền	2023-11-23
3	100022189274901	Trần Diệu Hiền	Này thiệt nha, kiểu bị cuốn cuốn sao í🐸	2023-11-23
4	100091124224730	Hân Lê	Hay	2023-11-23
...	...	...	...	...
7466	100029720314950	Kiều Hồ Trung Dũng	Nếu nói theo hướng tích cực thì những cái cây ...	2023-03-30
7467	100022488562817	Đức Đức	Đào Lê Tâm An	2023-03-30
7468	100012760427859	Huỳnh Trang Hoàng Mỹ	Minh Duy Trịnh Phan để cho dễ nhận diện thuii,...	2023-03-30
7469	100002686756259	Minh Duy Trịnh Phan	Huỳnh Trang Hoàng Mỹ mình thích ý nghĩ của bạn 😁	2023-03-30
7470	100081237549027	Ngọc Như Ý	Hà Phương sẽ thật tuyệt zời	2023-03-30

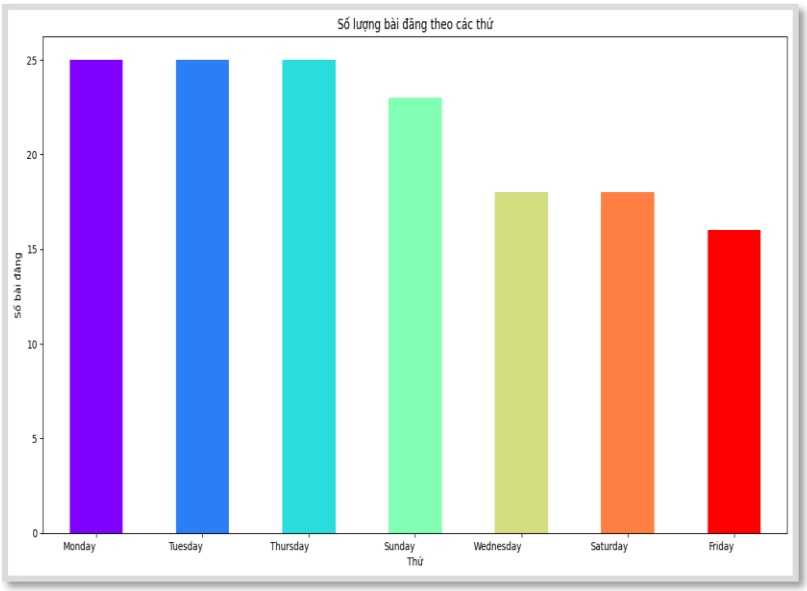
7471 rows × 4 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7471 entries, 0 to 7470
Data columns (total 4 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   commenter_id    7471 non-null   object
 1   commenter_name  7471 non-null   object
 2   comment_text    7471 non-null   object
 3   comment_time    7471 non-null   datetime64[ns]
dtypes: datetime64[ns](1), object(3)
memory usage: 233.6+ KB
```

Phần 3: Phân tích

I. Bài viết

1. Bài đăng theo các thứ



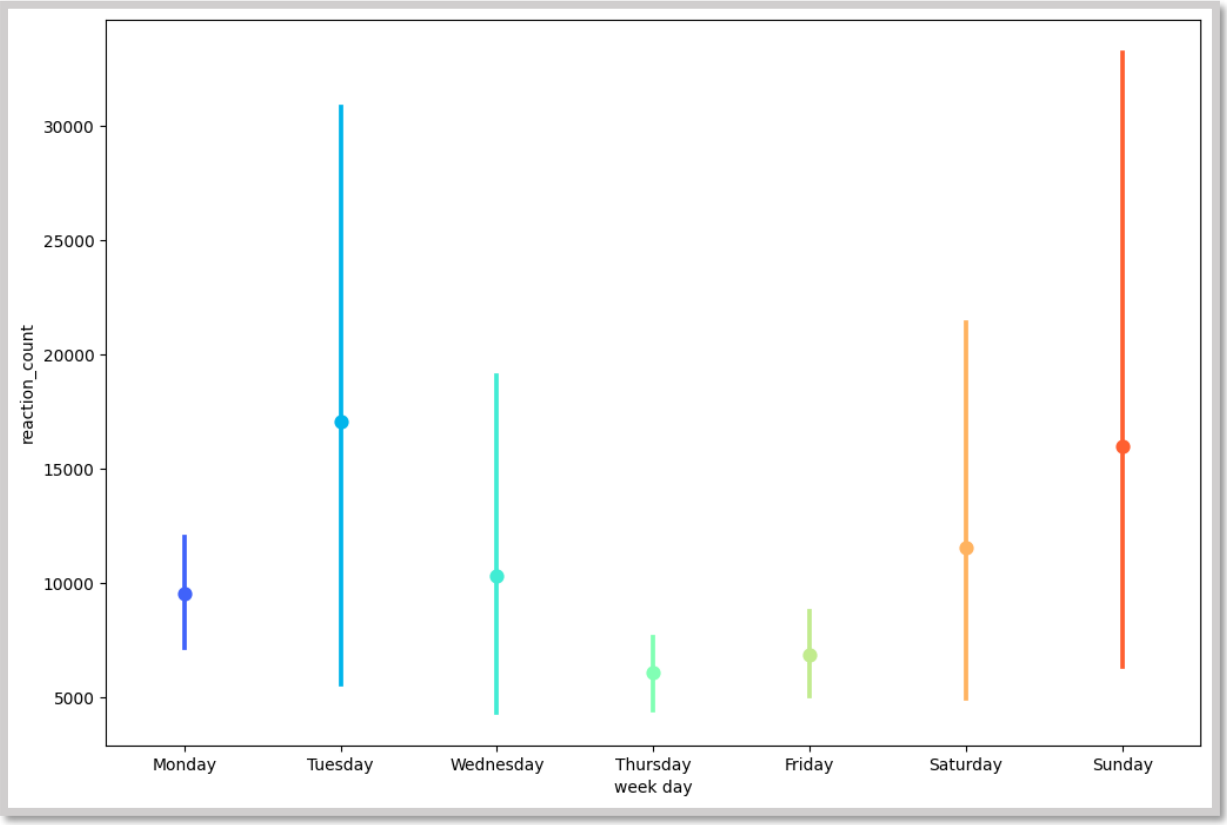
Ta thấy thứ **hai, ba và năm** cùng có số bài đăng là **nhiều nhất**, 25 bài ; ngày có số bài đăng **ít nhất** là thứ **sáu**

⇨ Như vậy fanpage này chủ yếu hoạt động vào đầu tuần và kém hoạt động hơn vào cuối tuần

Các bài đăng vào hôm thứ 2

	post_id	text	post_text	time	image	video	comments	shares	reaction_count	week_day
1	944225883728392	elm mới hai tuổi, mà suốt ngày dính phờ-ốt, tó...	elm mới hai tuổi, mà suốt ngày dính phờ-ốt, tó...	2023-11-27 11:59:35	1	0	446	148	15157	Monday
20	919536056197375	Ước gì có cái công tắc để bật tắt mọi suy nghĩ...	Ước gì có cái công tắc để bật tắt mọi suy nghĩ...	2023-10-16 12:00:48	1	0	296	603	7312	Monday
33	902544574563190	Có những tổn thương chỉ lộ ra khi người ta nhậ...	Có những tổn thương chỉ lộ ra khi người ta nhậ...	2023-09-18 12:10:38	1	0	304	1210	15843	Monday
43	890615962422718	Khi bạn còn trẻ, họ mặc định bạn không biết gì...	Khi bạn còn trẻ, họ mặc định bạn không biết gì...	2023-08-28 12:46:14	1	0	2558	3174	6015	Monday
47	886134539537527	Có biết vì sao bạn hay yêu nhầm người không? V...	Có biết vì sao bạn hay yêu nhầm người không? V...	2023-08-21 01:33:54	1	0	747	239	14899	Monday

Đăng bài vào thứ mấy thì nhận được nhiều tương tác ?



- Trung bình thứ ba, chủ nhật sẽ nhận được nhiều tương tác nhất, thứ năm,sáu nhận được ít nhất

- Thứ năm và sáu đăng tương đối nhiều (41/150 bài) nhưng nhận lại không được nhiều tương tác

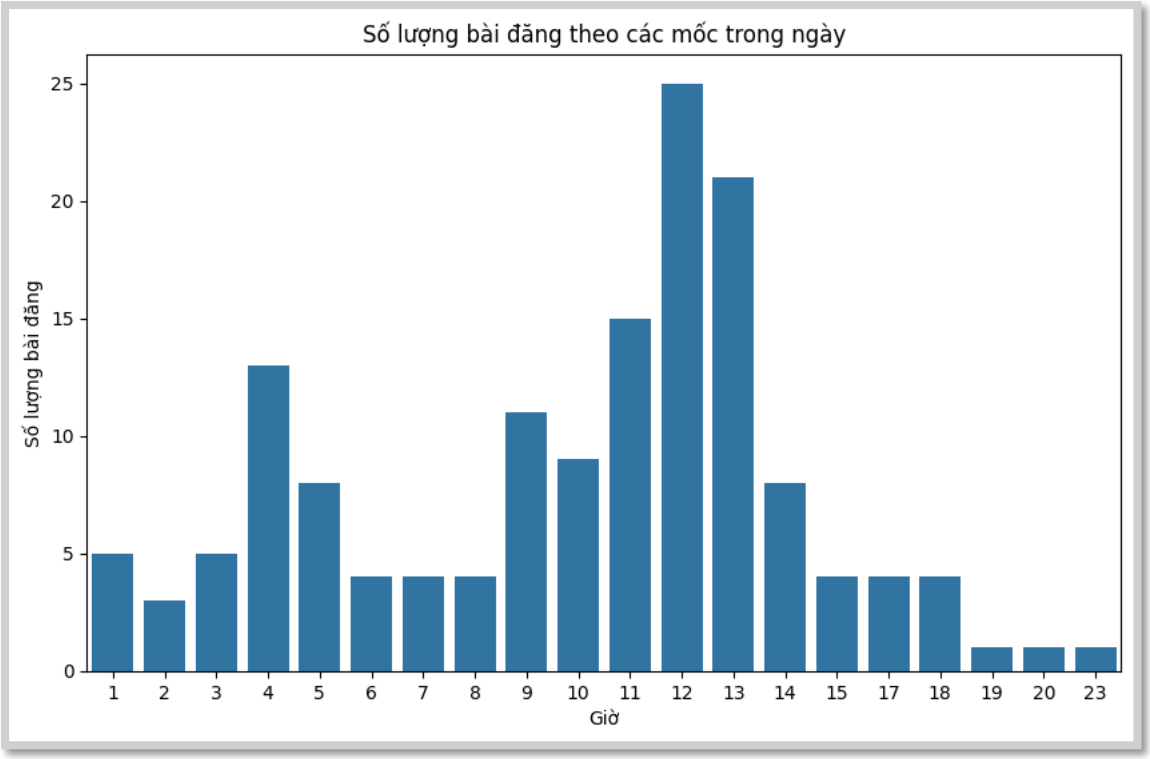
Bốn ngày có nhiều bài viết nhất

```
Monday: 25
Sunday: 23
Friday: 16
Thursday: 25
```

-

⇨ Đăng bài vào thứ ba hoặc chủ nhật để nhận được nhiều tương tác

Các mốc thời gian trong ngày hay đăng bài

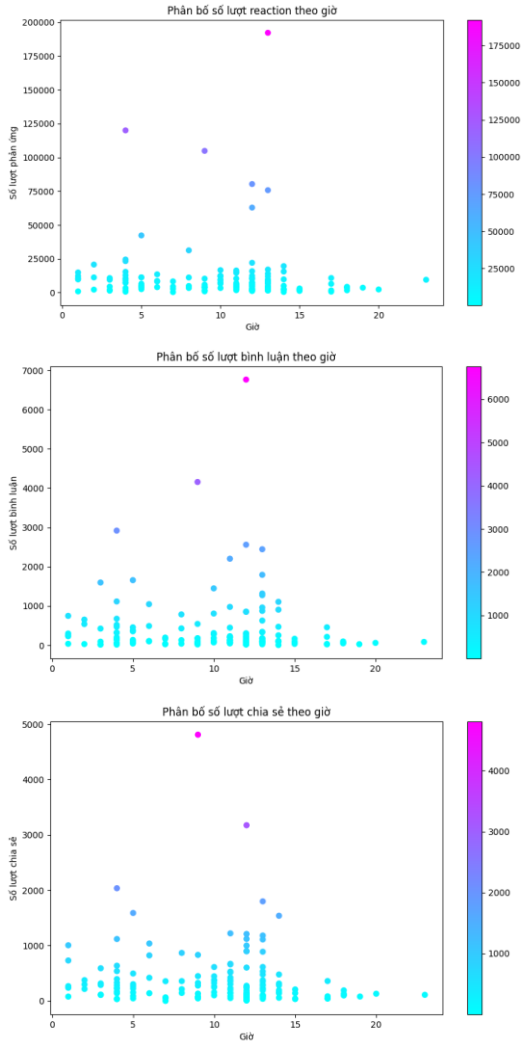


- Mốc 11-13h đăng nhiều bài nhất là 61 bài
- Mốc 19-2h đăng ít bài nhất là 11 bài

```
Mốc 1 giờ: 5 bài viết
Mốc 2 giờ: 3 bài viết
Mốc 3 giờ: 5 bài viết
Mốc 4 giờ: 13 bài viết
Mốc 5 giờ: 8 bài viết
Mốc 6 giờ: 4 bài viết
Mốc 7 giờ: 4 bài viết
Mốc 8 giờ: 4 bài viết
Mốc 9 giờ: 11 bài viết
Mốc 10 giờ: 9 bài viết
Mốc 11 giờ: 15 bài viết
Mốc 12 giờ: 25 bài viết
Mốc 13 giờ: 21 bài viết
Mốc 14 giờ: 8 bài viết
Mốc 15 giờ: 4 bài viết
Mốc 16 giờ: 0 bài viết
Mốc 17 giờ: 4 bài viết
Mốc 18 giờ: 4 bài viết
Mốc 19 giờ: 1 bài viết
Mốc 20 giờ: 1 bài viết
Mốc 21 giờ: 0 bài viết
Mốc 22 giờ: 0 bài viết
Mốc 23 giờ: 1 bài viết
```

⇨ Fanpage chủ yếu đăng vào lúc trưa, đăng rất ít vào buổi tối

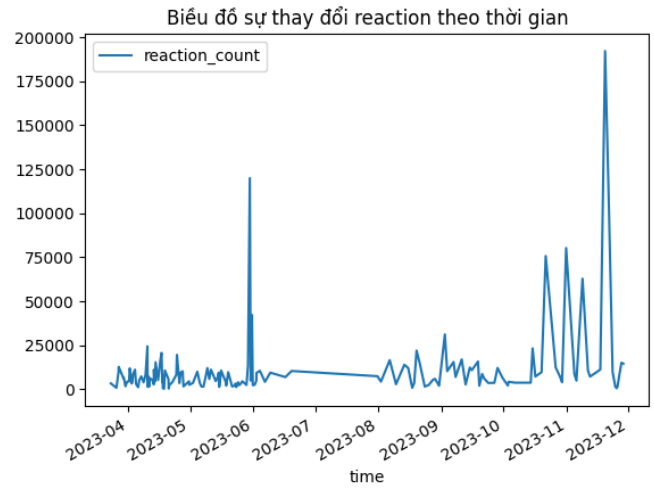
Phân bố các lượt tương tác theo giờ



- Trong khung giờ từ 11-13h, lượt tương tác nhận lại được rất nhiều ngược lại buổi tối giao đoạn 19-22h rất ít tương tác. Sự khác biệt này phản ánh mẫu quen thuộc hành vi trực tuyến của giới trẻ. Buổi trưa là thời gian nghỉ trưa, mọi người có xu hướng tìm đến những thứ giải trí sau giờ học hoặc làm việc căng thẳng. Trong khi đó, buổi tối có thể là thời gian họ dành cho các hoạt động khác như học tập, gia đình, bạn bè, …

⇨ Tóm lại, thời gian thích hợp nhất để Fanpage Hai Chiều đăng bài là trưa của thứ ba hoặc chủ nhật, hạn chế đăng vào các tối của thứ năm và sáu vì khi ấy có đăng nhiều cũng không nhận được nhiều phản hồi

Phân tích về reaction




```
Độ giao động của reaction_count là: 22014.571394267226
```

- Số lượng react tăng đột ngột vào cuối tháng 5 và cuối tháng 11
- Có nhiều biến động , độ giao động lớn ~ 22000
- Giai đoạn tháng 6-8 hầu như không có sự phát triển

Bài viết có nhiều lượt reaction nhất

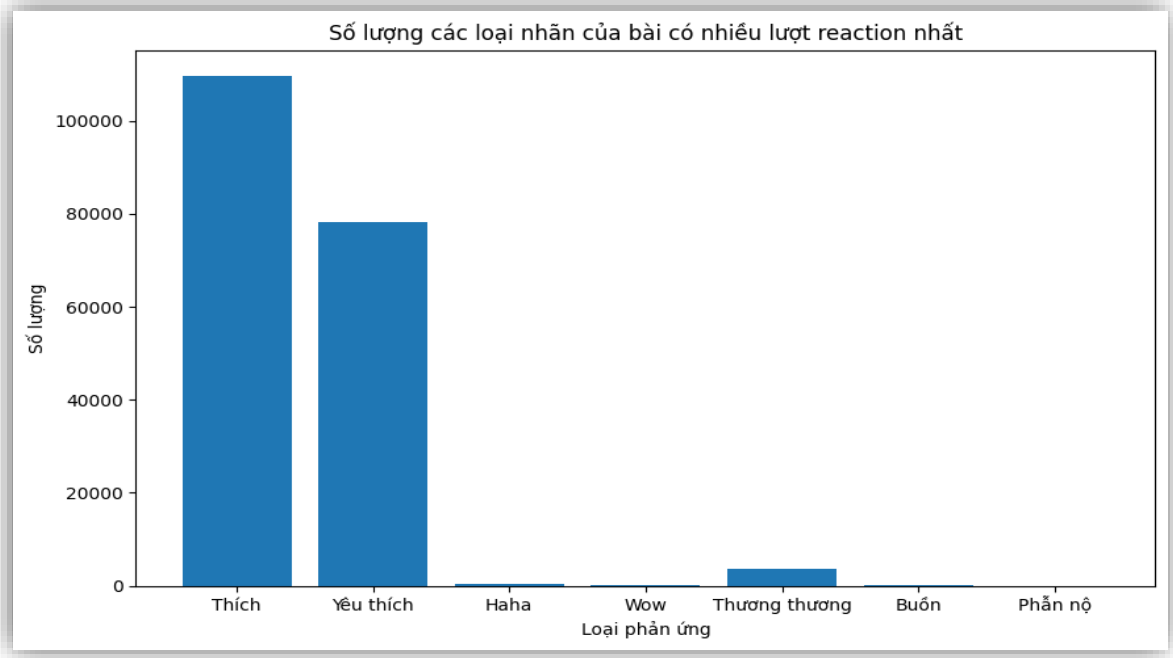
Thông tin bài viết có nhiều reaction nhất

post_id	939784010839246
text	Ở trường, những bài học không chỉ nằm trên bảng.
post_text	Ở trường, những bài học không chỉ nằm trên bảng.
time	2023-11-19 13:01:39
image	1
video	0
comments	1315
shares	1799
comments_full	[{'comment_id': '897054048437725', 'comment_ur...
reaction_count	192186
week_day	Sunday
hour	13
thích	109684.0
yêu thích	78073.0
haha	464.0
wow	103.0
thương thương	3731.0
buồn	125.0
phẫn nộ	6.0

Bài viết có ít lượt reaction nhất

Thông tin bài viết có ít lượt reaction nhất

post_id	809295750554740
text	Poster của team, ad dễ cho page công ty =)) lụ...
post_text	Poster của team, ad dễ cho page công ty =)) lụ...
time	2023-04-18 07:43:49
image	1
video	0
comments	24
shares	1
comments_full	[{'comment_id': '537577775215269', 'comment_ur...
reaction_count	345
week_day	Tuesday
hour	7
thích	226.0
yêu thích	108.0
haha	5.0
wow	1.0
thương thương	5.0
buồn	0.0
phẫn nộ	0.0

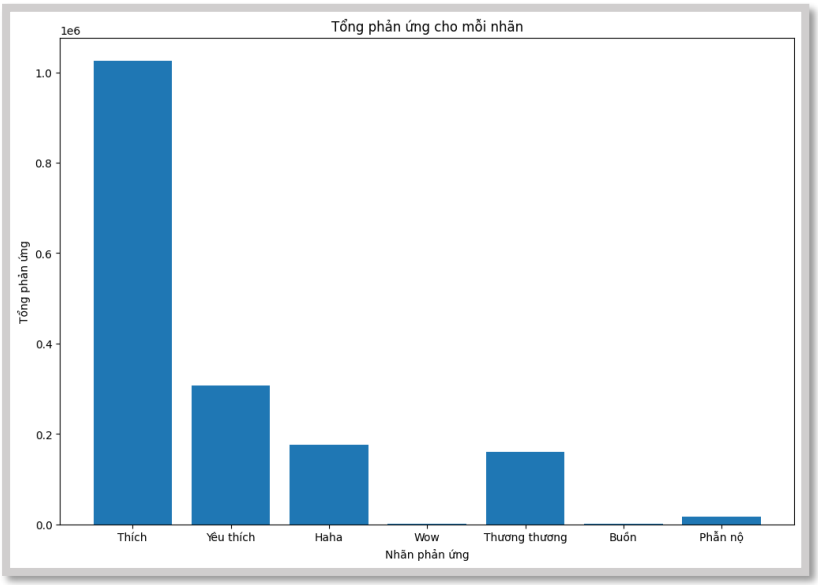


Số lượng reaction bài cao nhất này chủ yếu đến từ nhãn Thích và Yêu thích

⇨ Nhận xét về các ảnh trên:

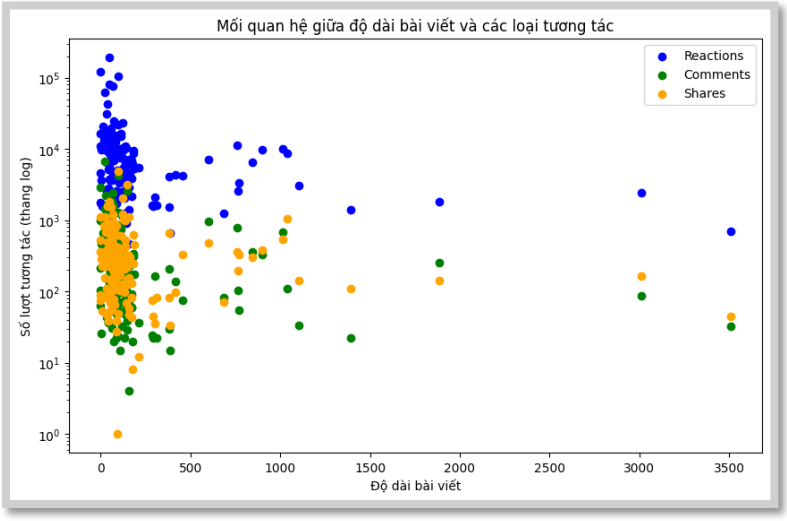
Số lượng reaction tăng đột ngột vào giai đoạn cuối tháng 5 và cuối tháng 11 , đạt đỉnh là bài đăng vào lúc 13h ngày 19-11 với 192186 reaction, chủ yếu đến từ các nhãn Thích và Yêu thích; đạt đáy vào lúc 7h ngày 18-4 với 345 reaction. Bài cao nhất gấp 557 lần bài thấp nhất

Tổng react cho mỗi nhãn



⇨ Lượt react của các bài viết chủ yếu đến từ nhãn Thích , điều này cho thấy người dùng có xu hướng thể hiện sự ủng hộ hoặc sự thích với nội dung một cách đơn giản, thay vì biểu lộ ra cảm xúc thật của mình thông qua các nhãn khác

Mối quan hệ giữa các loại tương tác và độ dài bài viết



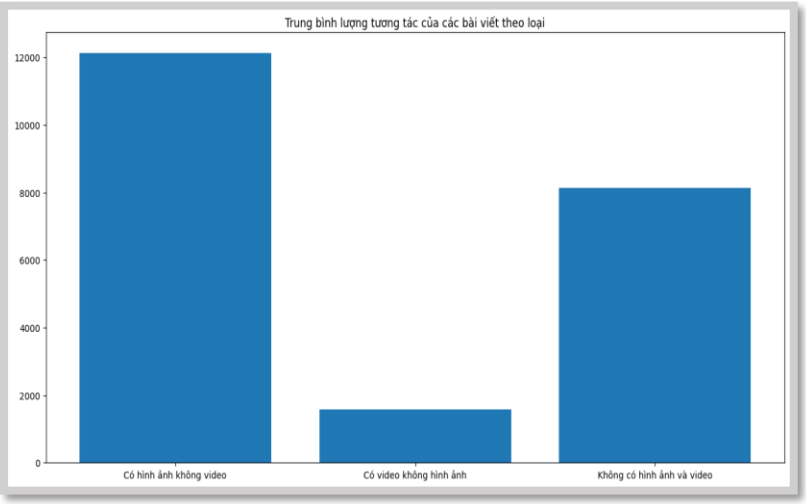
- Văn bản càng ngắn lượt tương tác càng nhiều

⇨ Người dùng facebook có xu hướng ưa chuộng những nội dung ngắn gọn, dễ tiếp cận, và nhanh chóng

Việc có hình ảnh hoặc video trong bài viết liệu thu hút người dùng hơn không?

```
Số bài viết có hình ảnh không video: 122
Số bài viết có video không hình ảnh: 3
Số bài viết không có hình ảnh và video: 25
```

Fanpage chủ yếu đăng bài kèm theo hình ảnh



Lượt tương tác khi có hình ảnh là rất cao, trong khi có video thì lại rất thấp. Điều này có thể cho thấy rằng người dùng thích sự ngắn gọn, dễ tiếp cận. Ngoài ra, nội dung hình ảnh có thể cung cấp thông tin một cách trực quan và dễ tiếp nhận hơn video, làm tăng khả năng thu hút sự chú ý và tương tác từ phía người dùng

	comments	shares	reaction_count
0	2202	671	14595
1	446	148	15157
2	82	70	1234
3	15	33	661
4	30	83	1546
...	...	...	...
146	322	637	12811
147	254	1006	9757
148	36	79	891
149	39	145	3409
mean	414	397	11259

⇨ Trung bình mỗi bài viết sẽ có khoảng 400 comments, 400 lượt shares và 11300 lượt react. Dựa trên số liệu này ta có thể khẳng định fanpage có sức hút khá lớn

Phân tích nội dung bài viết

Các từ thông dụng

Các từ có thể dễ dàng nhìn thấy như là: có, bạn, không, người, …

```
10 từ xuất hiện nhiều nhất trong bài viết
[('có', 141), ('là', 127), ('không', 115), ('bạn', 100), ('một', 96), ('người', 91), ('của', 69), ('những', 65), ('và', 63), ('sẽ', 58)]
```

Tìm thử một từ khóa bất kì

```
Tổng số 150 bài viết
3 bài viết chứa từ khóa 'xã hội'

['Mạng xã hội đang dần biến con người thành một cỗ máy vô cảm khi nhiều người tự nhốt bản thân vào đó mà mất dần khả năng kết nối với thế giới thực.',
 'Là một sinh vật xã hội, chúng ta luôn khao khát một danh tính, một diện mạo nhất định trong khi không ai thực sự hiểu bản thể của mình như thế nào.',
 'Mạng xã hội thực chất là nơi kết nối...những kẻ cô đơn?\n\nNghe có vẻ cực đoan ha, nhưng một nghiên cứu năm 2021 chỉ ra trong 10 ngày sử dụng các nền tảng mạng :
```

Trong bài viết có emoji không?

```
Có 72 emoji được sử dụng trong 150 bài viết
Trong đó có 46 bài viết sử dụng emoji
```

⇨ Có tận 46 bài viết chứa emoji, 72 emoji đã được sử dụng. Chủ fanpage có thể thuộc thế hệ Gen Z hoặc có xu hướng sử dụng các phong cách giao tiếp phổ biến trong thế hệ này. Việc sử dụng rộng rãi emoji của tác giả phản ánh sự sáng tạo và mong muốn tạo ra nội dung thú vị, đáng yêu và gần gũi với cộng đồng facebook

II. Bình luận

1. Phân tích comment

Độ dài trung bình của comment?

```
Độ dài trung bình của bình luận là: 82 kí tự
```

Các từ thông dụng

Các từ thông dụng có thể dễ dàng nhìn thấy là: là, thì, mình, có, mà, …

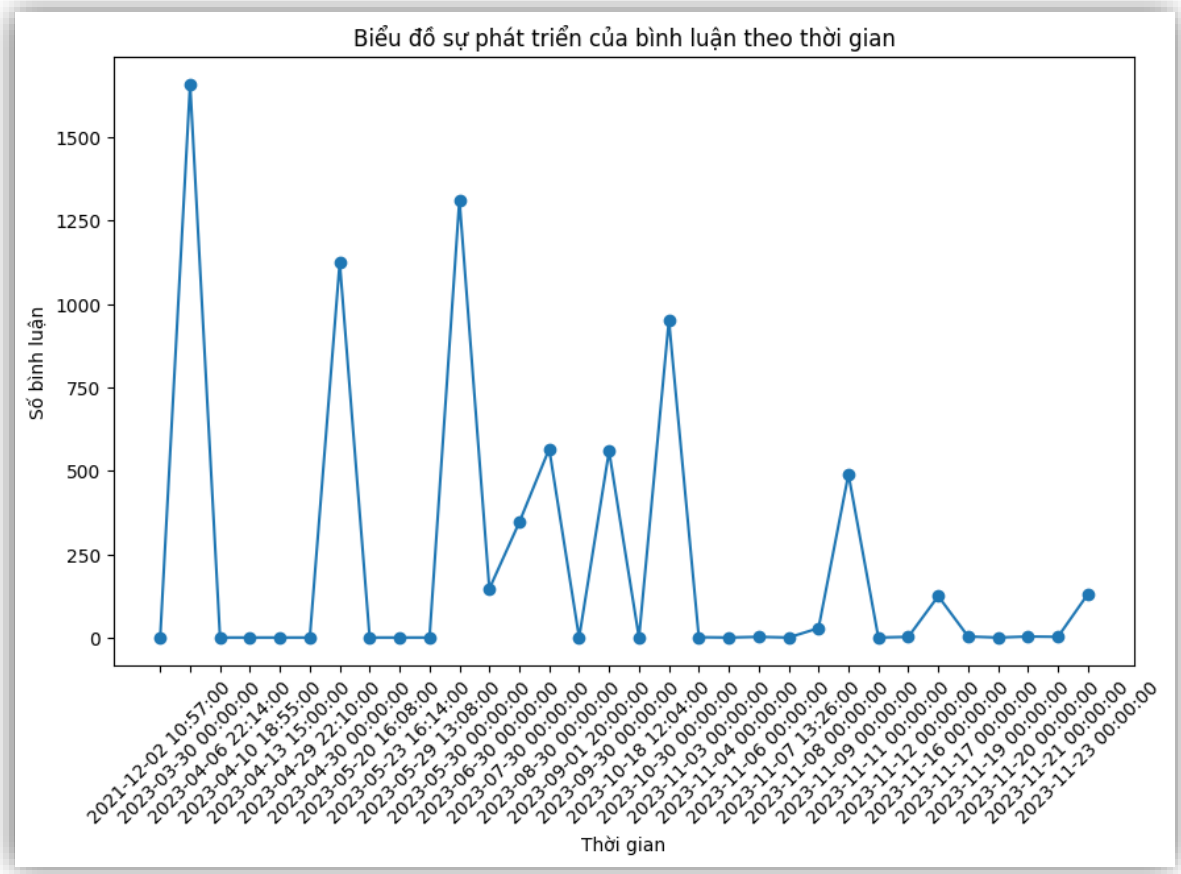
```
10 từ xuất hiện nhiều nhất trong nội luận
 [('là', 2483), ('có', 1951), ('thì', 1610), ('mình', 1565), ('không', 1238), ('người', 1208), ('bạn', 1158), ('mà', 1124), ('cũng', 1041), ('con', 955)]
```

Emoji trong bình luận

Ở phần phân tích nội dung bài viết, ta đã thấy xuất hiện rất nhiều emoji rồi, nên từ đó có thể chắc chắn rằng phần bình luận cũng có.

```
Có 3249 emoji được sử dụng trong 7471 bình luận
Trong đó có 2024 bình luận sử dụng emoji
```

Độ phát triển của comment theo thời gian



- Ở biểu đồ này, do đã chuẩn hóa một số giá trị nên trông nó không được liên tục đẹp mắt, nhưng nhìn chung có thể thấy được số lượng comment phát triển rất mạnh thời gian đầu tháng 2 – 4 năm 2023 và giảm mạnh về cuối năm
- Có lẽ fanpage này nổi do các bài viết tạo trend ở thời gian đầu và hiện tại không còn nhiều nữa nên mới gây ra sự suy giảm như này

Những ai là fan cứng của fanpage?

Ở đây em dựa theo tiêu chí số lượng comment của mỗi cá nhân để đánh giá

Top 5 người bình luận nhiều nhất là

	commenter_id	commenter_name	count
0	100044229679401	Hai Chiều	321
1	100043997325521	Thằng Anh Con Em	61
2	100065791324776	Xuân Nghi Vương	44
3	100028374075249	Reanzares Limuel	30
4	1829242625	Phan Hồng Đức	28

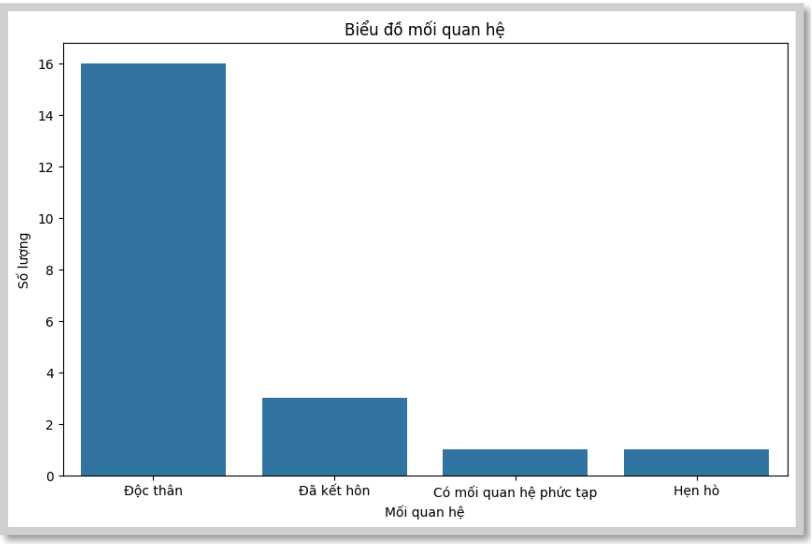
- Xuất hiện tên fanpage “Hai Chiều ” với số lượng nhiều nhất , điều này cho thấy họ có sự tôn trọng đối với cộng đồng của mình, hoặc có thể là họ đang giải đáp thắc của người hâm mộ
- Như vậy, ‘Thằng Anh Con Em’, ‘Xuân Nghi Vương’, ‘Reanzares Limuel’, ‘Phan Hồng Đức’ là các fan cứng.

2. Phân tích commenter
   Do không đủ kiến thức để crawl lại dữ liệu bị thiếu nên phần này em chỉ phân tích 66 profile nhận được từ dữ liệu cũ

	Follower_count	id	Name	Thông tin cơ bản	Học vấn	Mối quan hệ	Nơi từng sống
0	None	100043997325521	Thằng Anh Con Em	Ngày sinh	NaN	NaN	NaN
1	None	100082965051760	Trang Lê	NaN	NaN	NaN	NaN
2	None	100066800510946	Bích Trâm	NaN	Đại học Văn Hiến\nCao đẳng/Đại học\n1 tháng 10...	Độc thân	NaN
3	None	100022189274901	Trần Diệu Hiền	Nữ\nGiới tính	Trường Đại học Mở TP. HCM - Tư vấn tuyển sinh\...	Độc thân	NaN
4	None	100091124224730	Hân Lê	NaN	NaN	Lucie Tranová\nĐã kết hôn	Thành phố Hồ Chí Minh\nTỉnh/Thành phố hiện tại
...	...	...	...	...	...	...	...
61	None	61550611575810	The Lyricalogy	Ngày sinh	NaN	NaN	NaN
62	None	100090007038592	Quoc Anh Ng	NaN	NaN	Độc thân	NaN
63	None	100041235171670	L'une	Ngày sinh	NaN	NaN	NaN
64	None	100024095117283	Đặng Bạch Phát	Nam\nGiới tính	Đại học Bách khoa Hà Nội - Hanoi University of...	Độc thân	Hà Nội\nTỉnh/Thành phố hiện tại\nHà Nội\nQuê quán
65	None	100071394758382	Xóm Xi Xăm	Ngày sinh	NaN	NaN	NaN

66 rows × 7 columns

Những người độc thân



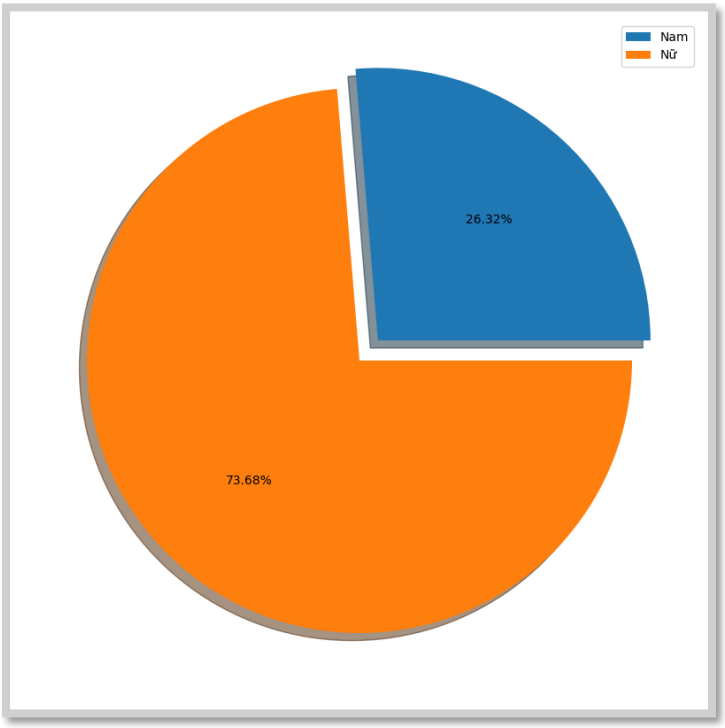
Phần lớn những commenter đều là người độc thân, khoảng 25%

```
Tỷ lệ số người độc thân chiếm 0.24242424242424243
```

⇨ Bài viết có thể chứa nội dung hoặc chủ đề mà người độc thân thường quan tâm hoặc muốn thảo luận

⇨ Fanpage có thể đã triển khai một chiến lược tương tác đặc biệt để kích thích sự tham gia từ nhóm đối tượng này

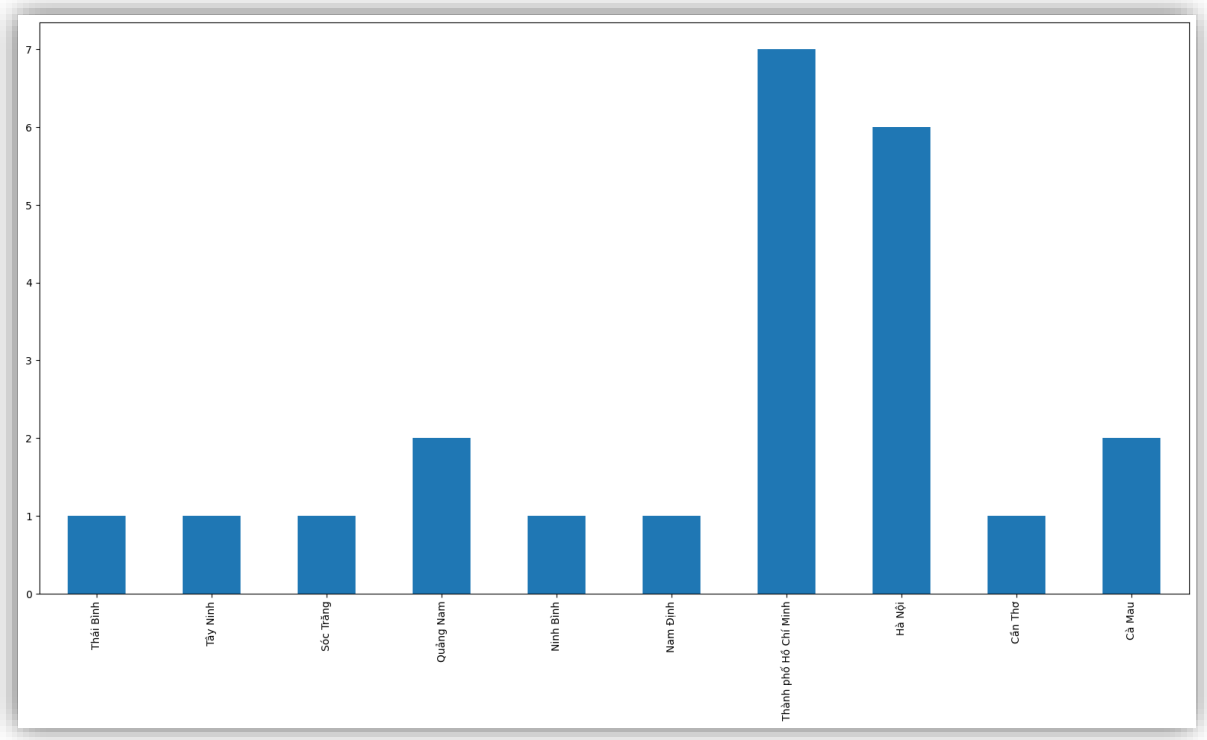
Tỷ lệ giới tính



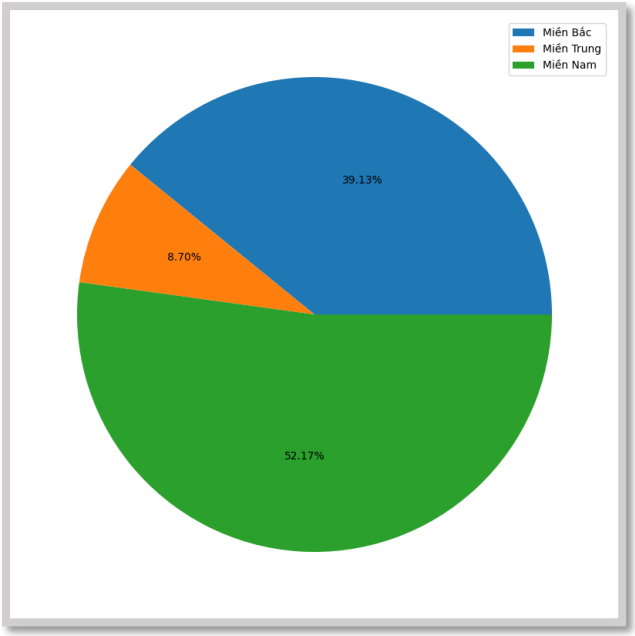
Những người comment chủ yếu là nam, số lượng gấp 3 lần nữ giới

⇨ Nội dung bài viết phù hợp với đối tượng nam

Nơi ở



- Commenter đến từ chủ yếu 2 thành phố lớn của Việt Nam là TP.HCM và TP. Hà Nội



- Miền Nam chiếm chủ yếu ~ 2 miền Bắc và Trung gộp lại