CRAWL DỮ LIỆU TỪ FACEBOOK

[Phần A : các công cụ được sử dụng 3](#_Toc151296095)

[Phần 1: vấn đề 3](#_Toc151296096)

[Phần 2: công cụ 4](#_Toc151296097)

[1.Crawl 4](#_Toc151296098)

[2.Preprocessing 4](#_Toc151296099)

[3.Analysis 6](#_Toc151296100)

[Phần B : thực thi 7](#_Toc151296101)

[Phần 1 : Crawl dữ liệu 7](#_Toc151296102)

[1.Page dùng để crawl 7](#_Toc151296103)

[2.Phương pháp 8](#_Toc151296104)

[3.Dữ liệu crawl được 8](#_Toc151296105)

[Phần 2 : Phân tích , trực quan hóa dữ liệu 10](#_Toc151296106)

[1.Tiền xử lý dữ liệu 10](#_Toc151296107)

[1.1,Xóa bỏ những cột bị miss dữ liệu 10](#_Toc151296108)

[1.2,Định dạng dữ liệu 12](#_Toc151296109)

[2.Thống kê và trực quan hóa 14](#_Toc151296110)

[2.1, Số lần tương tác, cái nhìn khái quát 14](#_Toc151296111)

[2.1.1, Thời gian thu thập 14](#_Toc151296112)

[2.1.2, Một số dữ liệu đáng chú ý 14](#_Toc151296113)

[2.1.3, Cái nhìn tổng thể 15](#_Toc151296114)

[2.2, Reactions hàng ngày 15](#_Toc151296115)

[2.3, Bức tranh tổng quan 16](#_Toc151296116)

[2.4, Max,min của likes,shares và comments 18](#_Toc151296117)

[2.4.1, Reactions 18](#_Toc151296118)

[2.4.2, Shares 19](#_Toc151296119)

[2.4.3, Comments 20](#_Toc151296120)

[2.5, Regular posting timestamps 22](#_Toc151296121)

[2.6, Tương tác theo giờ 23](#_Toc151296122)

[2.7,Thống kê trong tuần 24](#_Toc151296123)

[2.8,Tương quan thời gian đăng vài viết và lượt tương tác 25](#_Toc151296124)

[3.Phân tích từ khóa trong bài viết 26](#_Toc151296125)

[3.1, Số lượng xuất hiện của một vài từ khóa bất kỳ 26](#_Toc151296126)

[3.2, Tần suất các từ được sử dụng nhiều nhất 27](#_Toc151296127)

[3.3,Tìm mức độ quan trong của các từ trong bài post 27](#_Toc151296128)

[Phần 3 : Phân tích, đánh giá comments 28](#_Toc151296129)

[1.Tổng số comments 28](#_Toc151296130)

[2.Lọc comments 29](#_Toc151296131)

[3.Wordclound comments 30](#_Toc151296132)

[4,Phân tích độ tích cực, tiêu cực của các comments 31](#_Toc151296133)

[5.Phân tích người bình luận trong các bài Post 31](#_Toc151296134)

[5.1, Số người dùng comments 31](#_Toc151296135)

[5.2, Phân tích 32](#_Toc151296136)

[5.2.1, Lấy ID người dùng 32](#_Toc151296137)

[5.2.2, Lấy thông tin người dùng 32](#_Toc151296138)

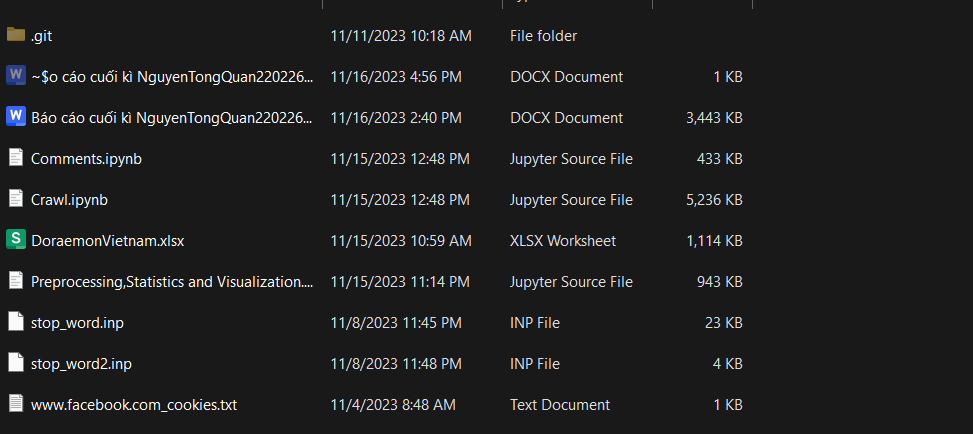
[5.2.3, Phân tích 33](#_Toc151296139)

# Phần A : các công cụ được sử dụng

## Phần 1: vấn đề

-Việc crawl dữ liệu từ facebook rất cần thiết và được phổ biến rộng rãi.Việc crawl có thể được thực hiện với mục đích nghiên cứu, phân tích thị trường,phát triển sản phẩm, hoặc hiểu rõ hơn về sự tương tác,ý kiến của cộng đồng.Do đó em đã thực hiện việc crawl dữ liệu của 1 page trên Facebook nhằm đánh giá sự tương tác, yêu thích của mọi người đối với 1 bộ truyện của tuổi thơ đó là Doraemon.Bài làm còn nhiều thiếu sót, rất mong nhận được sự đóng góp , nhận xét từ phía thầy.

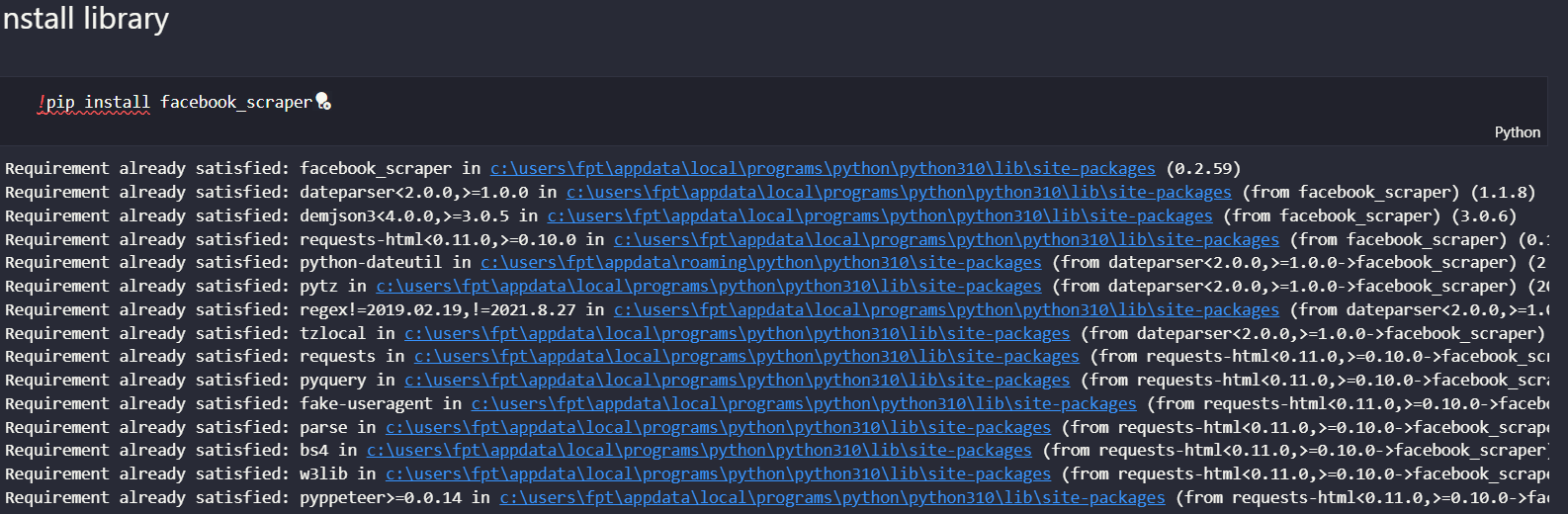
-Project bao gồm các file dữ liệu:



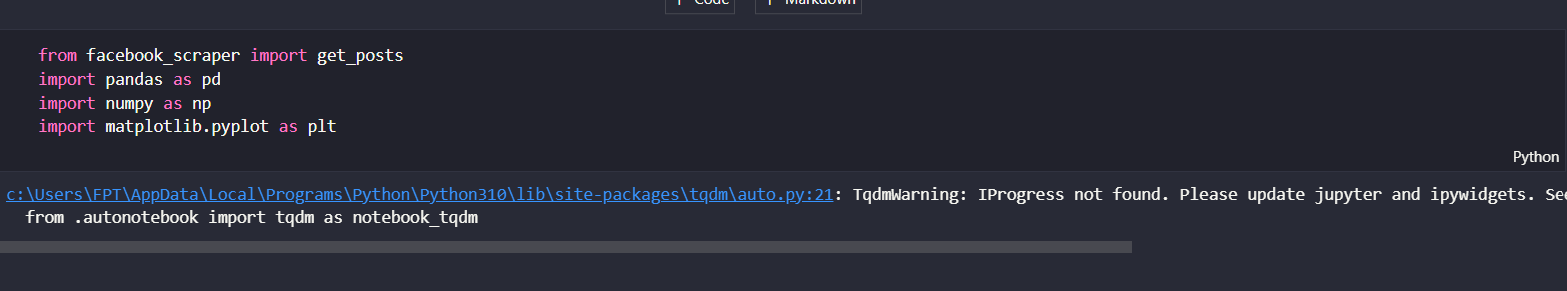
## Phần 2: công cụ

### 1.Crawl

-Sử dụng scraper cho phép cào dữ liệu từ Facebook:

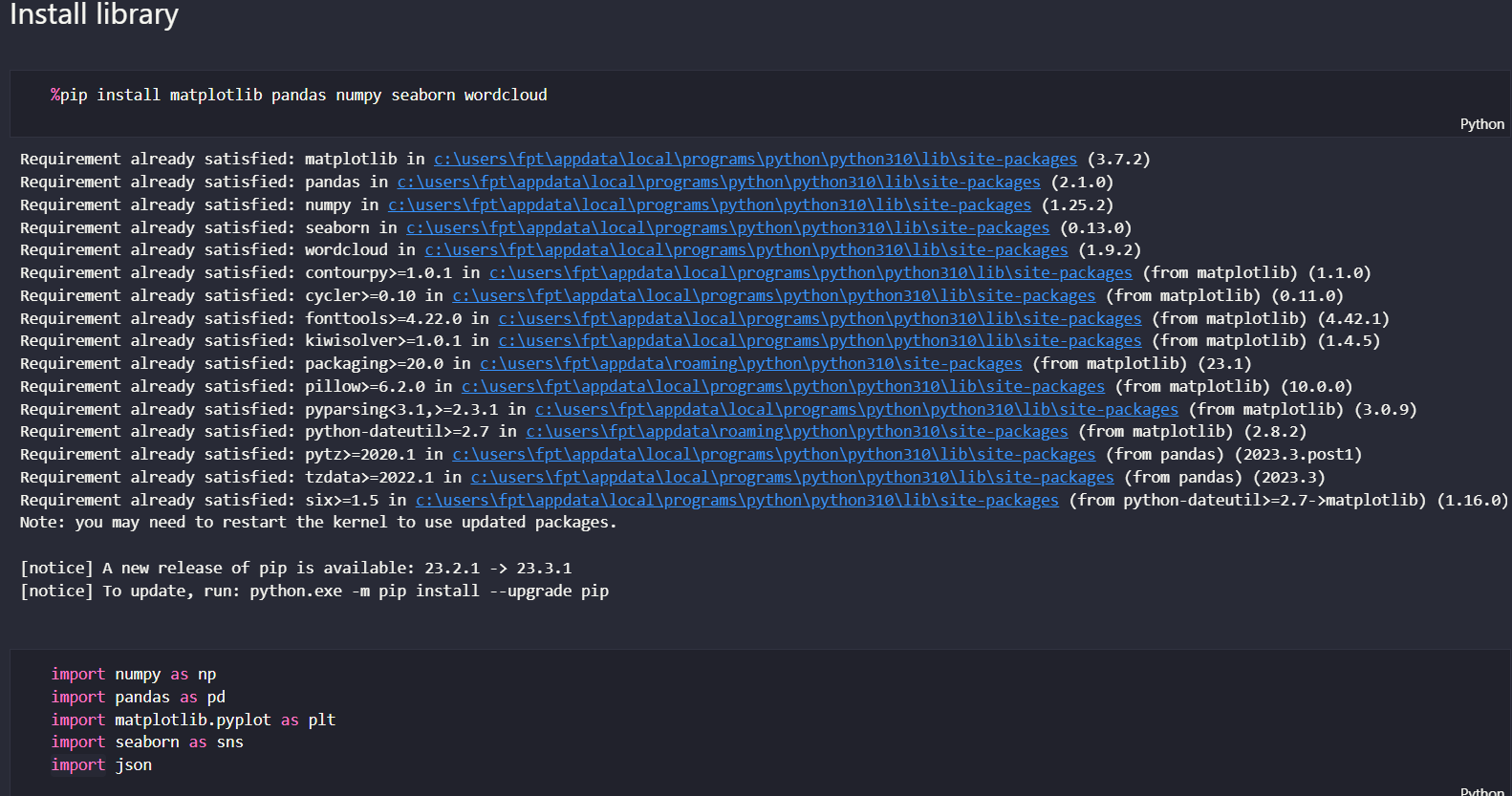


-Import các thư viện cần thiết :



### 2.Preprocessing

-Sử dụng các thư viện để hiện thị,xử lý,phân tích:



+Matplotlib:

Tác dụng: Matplotlib là một thư viện trực quan hóa dữ liệu, giúp tạo ra đồ thị và biểu đồ chất lượng cao.

Ứng dụng: Hiển thị và trực quan hóa dữ liệu thống kê, phân tích, và mô phỏng trong Python.

+Pandas:

Tác dụng: Pandas là một thư viện xử lý và phân tích dữ liệu dạng bảng (DataFrame).

Ứng dụng: Đọc, xử lý, và thống kê dữ liệu. Hỗ trợ quản lý và thao tác trên dữ liệu có cấu trúc.

+NumPy:

Tác dụng: NumPy là một thư viện toán học, cung cấp các cấu trúc dữ liệu mảng và các hàm toán học hiệu quả.

Ứng dụng: Thực hiện các phép toán số học trên mảng, xử lý dữ liệu đa chiều.

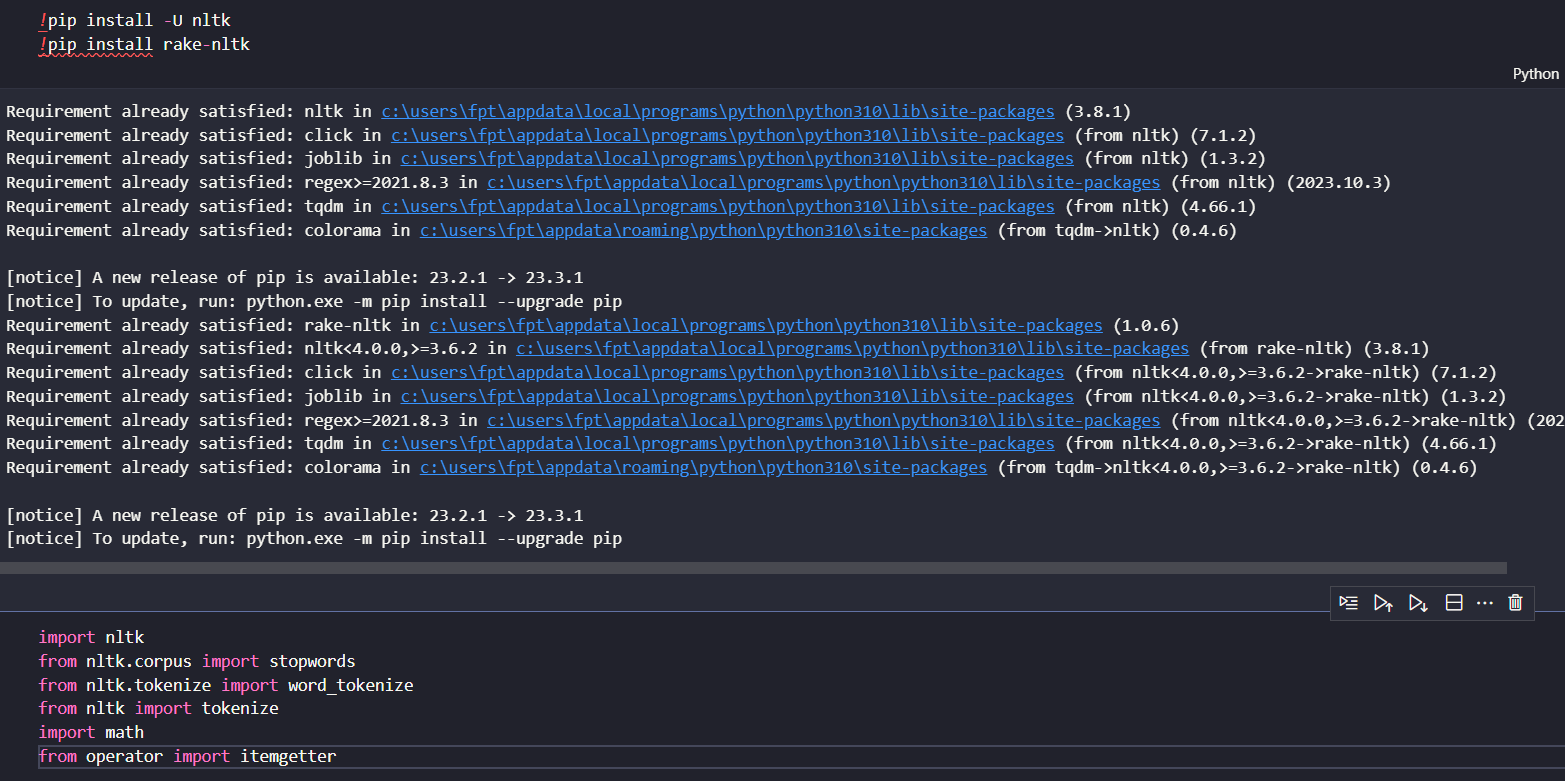
+Seaborn:

Tác dụng: Seaborn là một thư viện trực quan hóa dữ liệu dựa trên Matplotlib, tập trung vào việc tạo ra đồ thị thống kê có thẩm mỹ cao.

Ứng dụng: Trực quan hóa mối quan hệ thống kê phức tạp, phân tích mối quan hệ giữa nhiều biến.

+WordCloud:  
Tác dụng: WordCloud là một thư viện trực quan hóa dữ liệu, tạo ra hình ảnh từ các từ xuất hiện trong văn bản.  
Ứng dụng: Hiển thị các từ quan trọng trong văn bản, đặc biệt thường được sử dụng trong phân tích ý kiến và trực quan hóa từ khóa.

-Sử dụng stopword để tìm keyword:



### 3.Analysis

-Sử dụng SentimentAnalyzer để đánh giá cảm xúc, độ tích cực của comments:

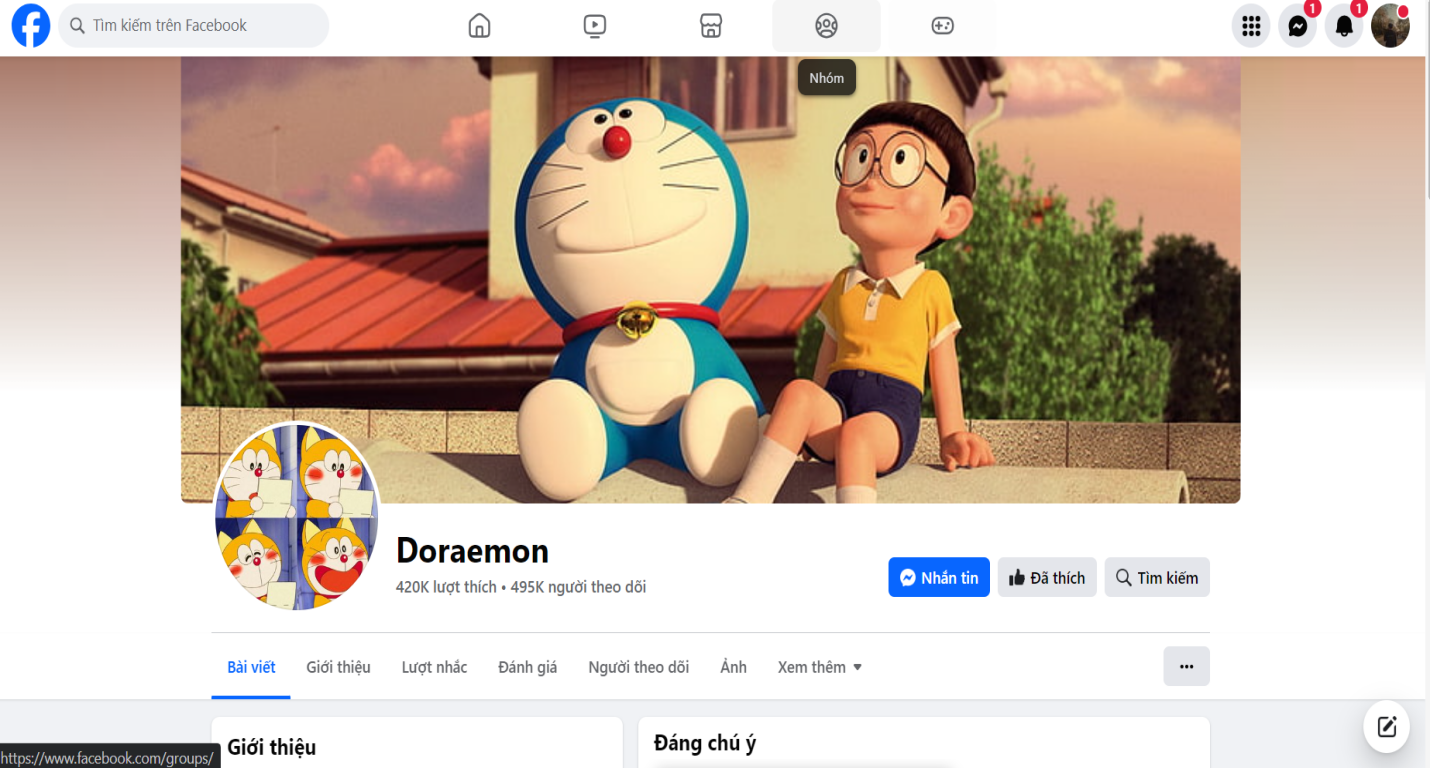


# Phần B : thực thi

## Phần 1 : Crawl dữ liệu

### 1.Page dùng để crawl

-Trang fanpage được chọn là DoraemonVietNam :



### 2.Phương pháp

-FANPAGE\_LINK: Liên kết đến trang fan hâm mộ mà chúng tôi muốn thu thập dữ liệu từ đó.

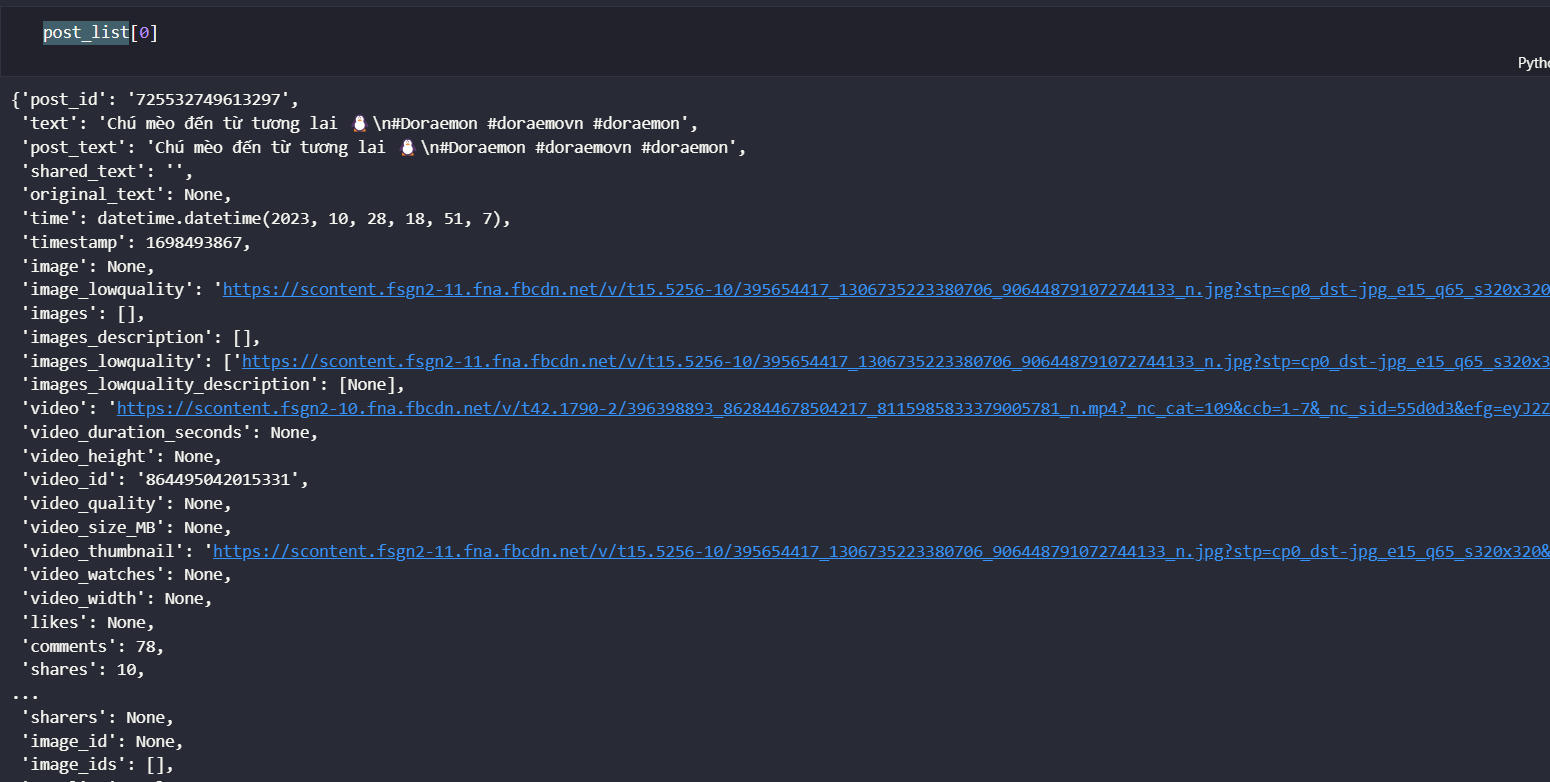
-COOKIE\_PATH: Đường dẫn đến tệp cookie mà chúng tôi sẽ sử dụng để xác thực với Facebook. Tệp cookie này có thể lấy được bằng cách đăng nhập vào Facebook và sao chép cookie từ trình duyệt.

-FOLDER\_NAME: Tên của thư mục mà chúng ta sẽ lưu dữ liệu vào. Thư mục này sẽ được tạo trong cùng thư mục với sổ ghi chép này.

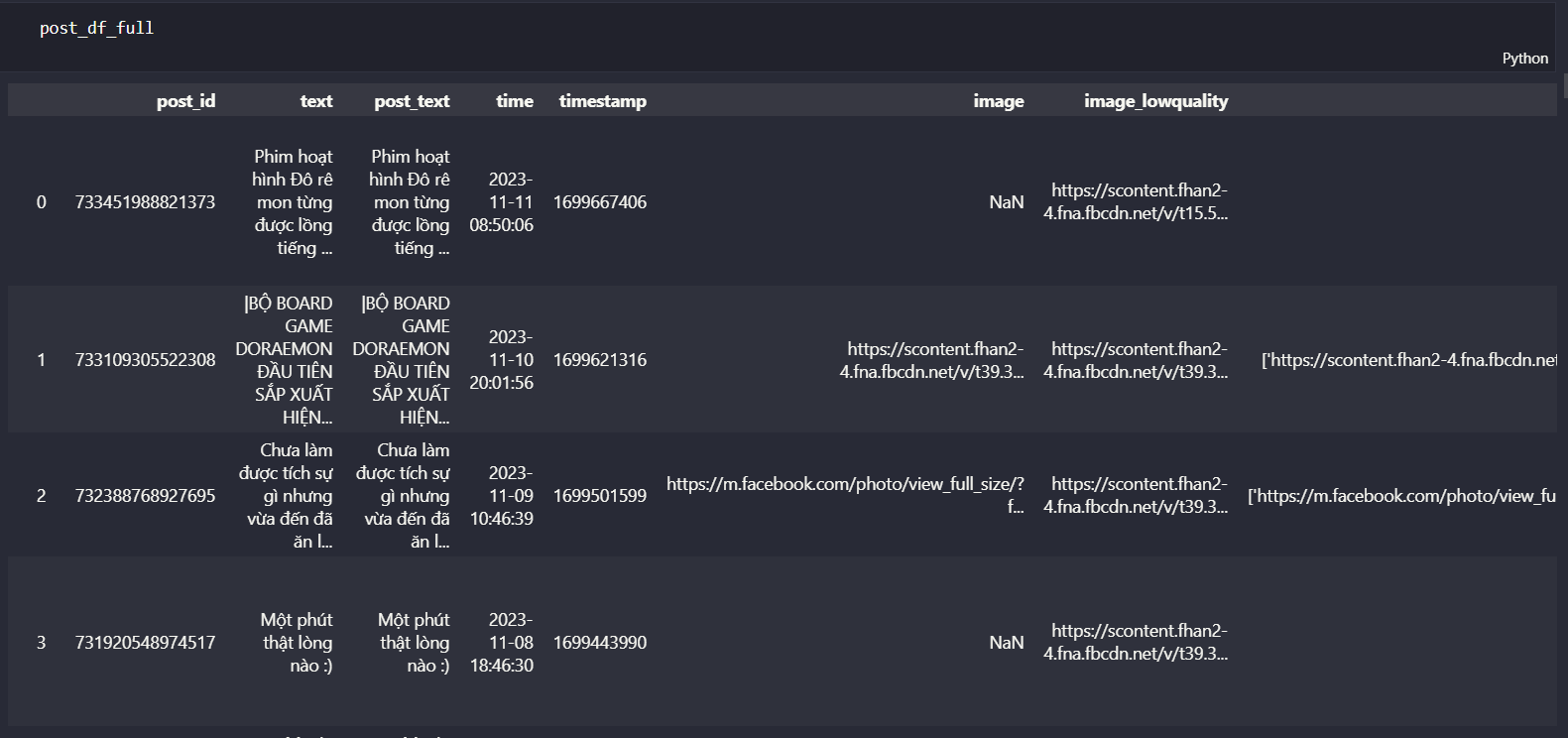


### 3.Dữ liệu crawl được

-Thông tin của 1 post:



-Thông tin crawl được:

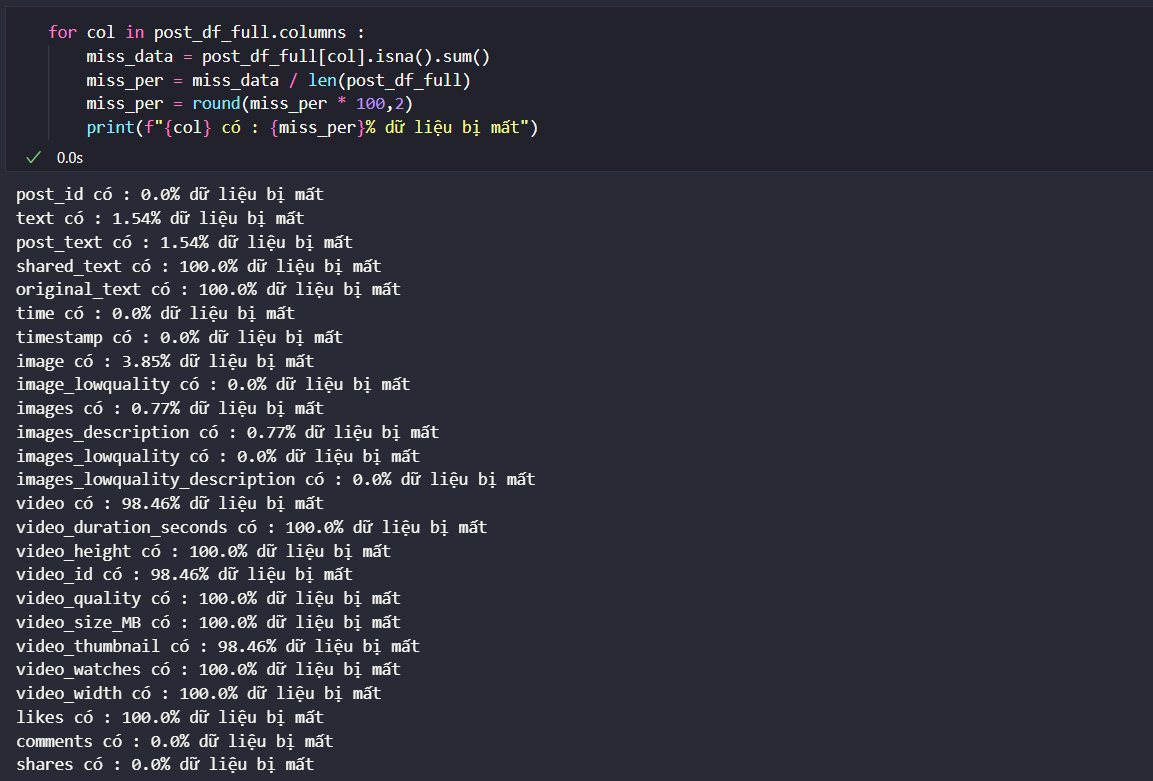


## Phần 2 : Phân tích , trực quan hóa dữ liệu

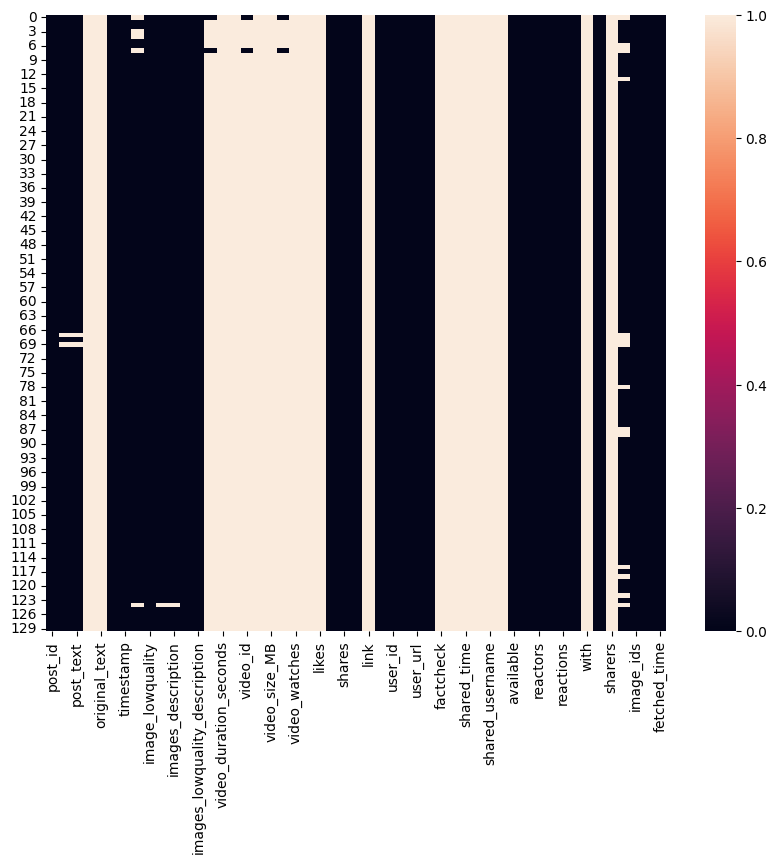
### 1.Tiền xử lý dữ liệu

#### 1.1,Xóa bỏ những cột bị miss dữ liệu

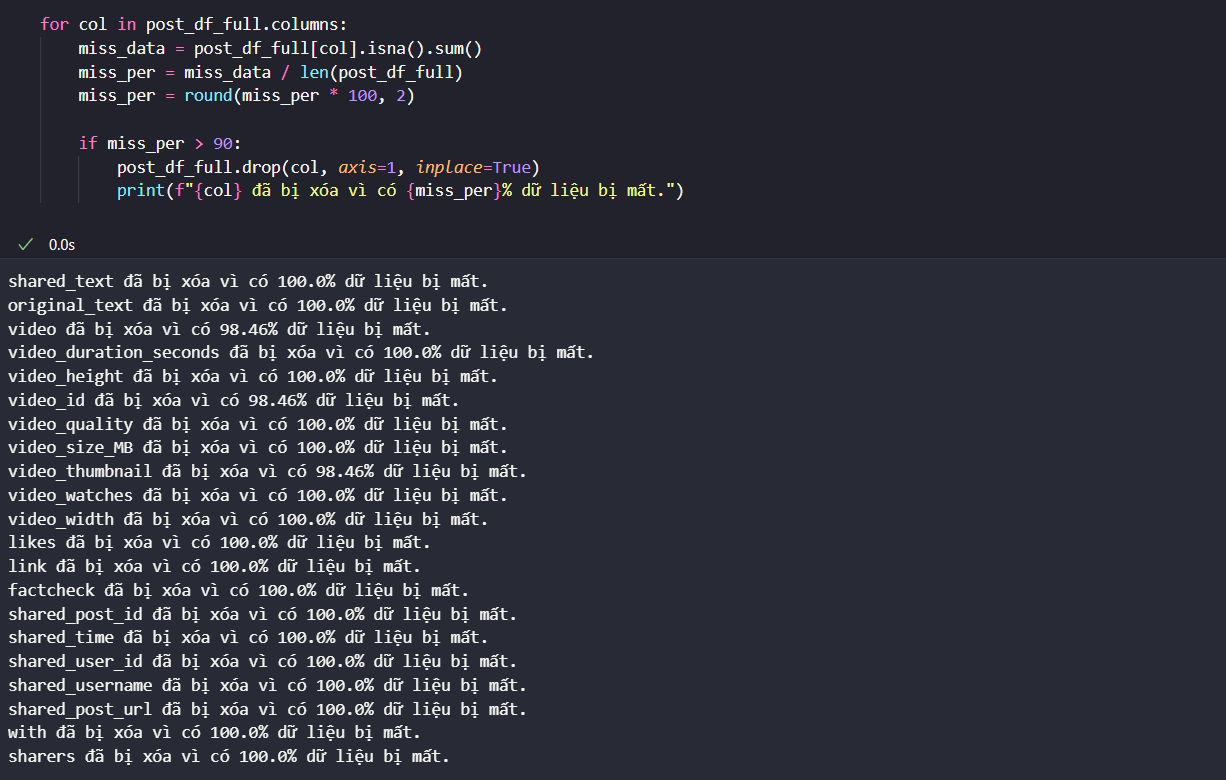
-Xác định dữ liệu bị mất:



-Hiển thị dữ liệu bị mất dưới dạng biểu đồ:

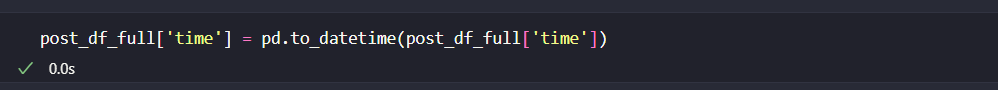


-Xóa những dự liệu không đủ để sử dụng:

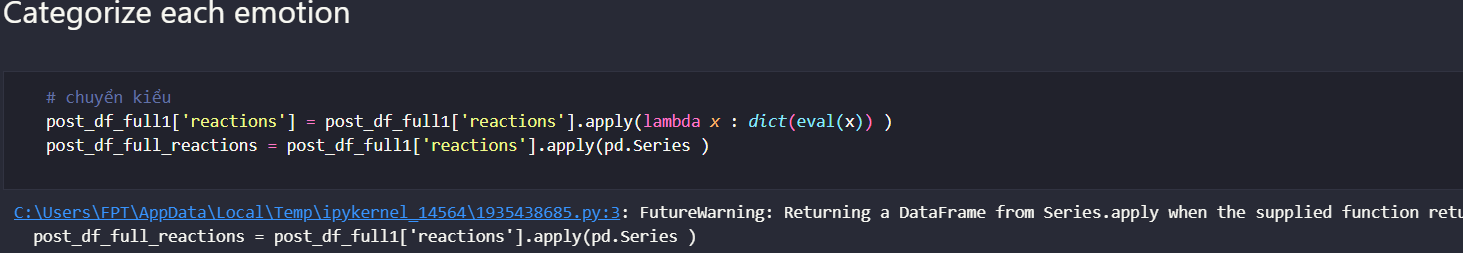


#### 1.2,Định dạng dữ liệu

-Dữ liệu ngày đang ở dạng chuỗi cần phải chuyển sang dạng datetime :

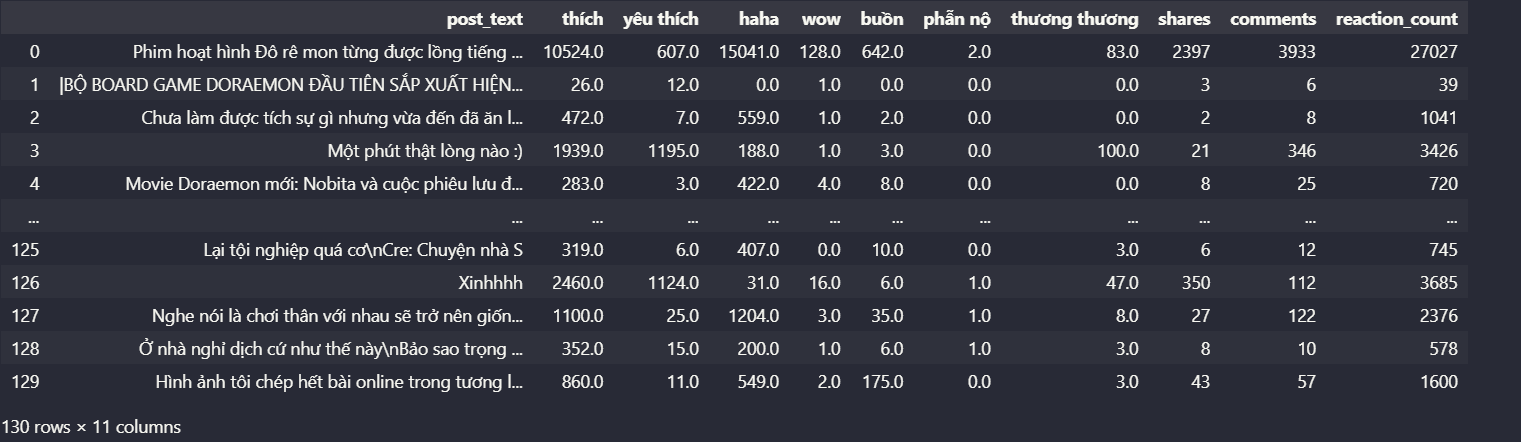


-Tách cột reactions thành những cột cảm xúc riêng biệt để dễ dàng thao tác:





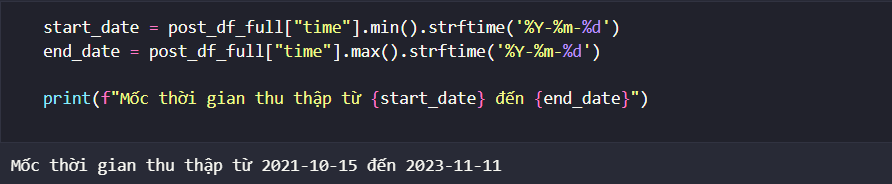
-Dữ liệu đã được tách riêng,các cột cảm xúc được phân chia rõ ràng:



### 2.Thống kê và trực quan hóa

#### 2.1, Số lần tương tác, cái nhìn khái quát

##### 2.1.1, Thời gian thu thập



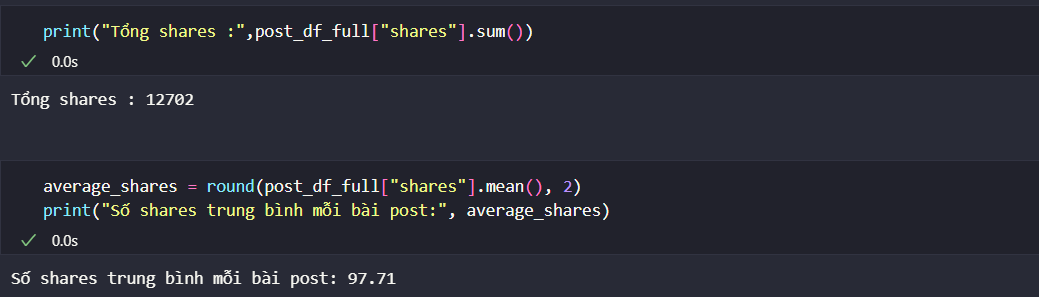
-Do Page đăng bài với số lượng bài viết không quá lớn, nên các số liệu thu thập được đến tận năm 2021

##### 2.1.2, Một số dữ liệu đáng chú ý

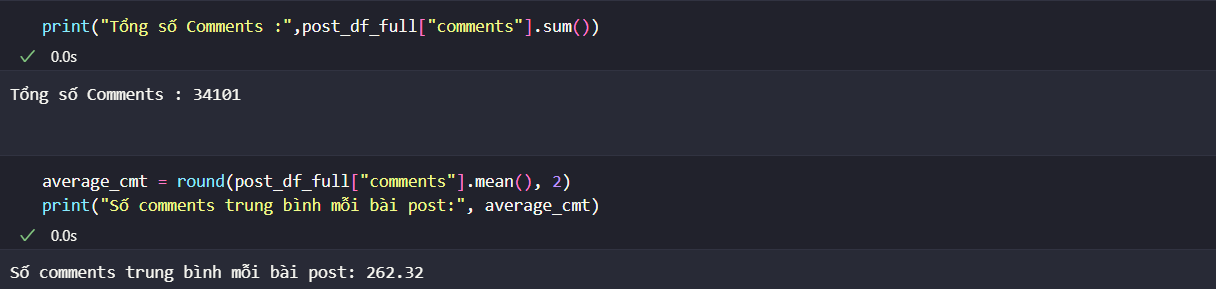
-Về lượt reactions,số lượng bày tỏ cảm xúc gần 900k lượt, một con số mơ ước với mỗi FanPage:



-Về lượt shares cũng khá cao gần 13k lượt:

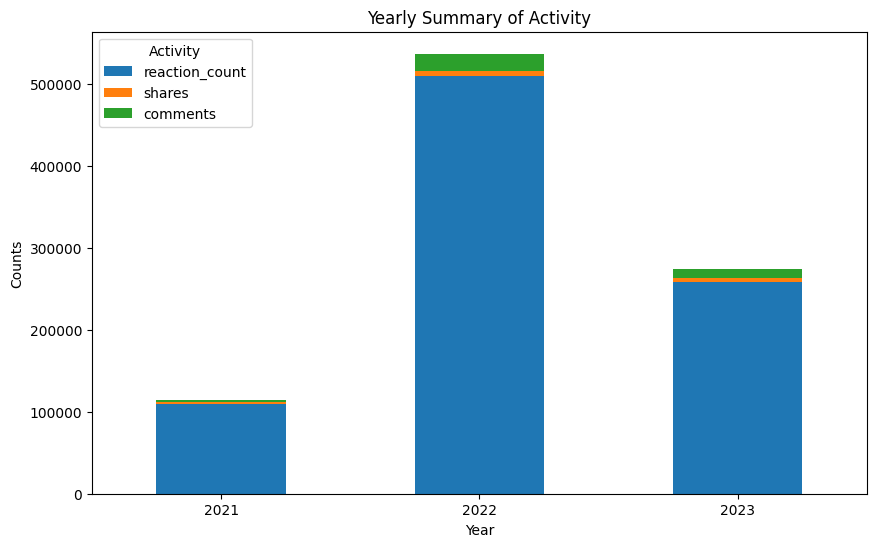


-Về lượt comments:



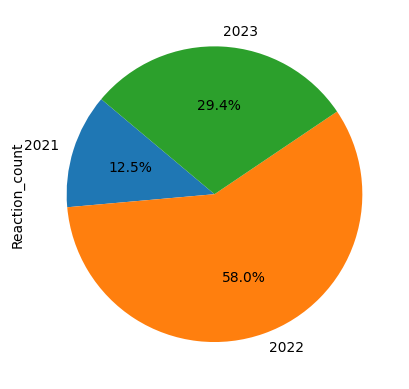
##### 2.1.3, Cái nhìn tổng thể

-Do sự chênh lệch về thời gian thu thập nên biểu đồ của năm 2022 vượt trội hơn cả:



#### 2.2, Reactions hàng ngày

-Số lượng tương tác được tính theo năm :

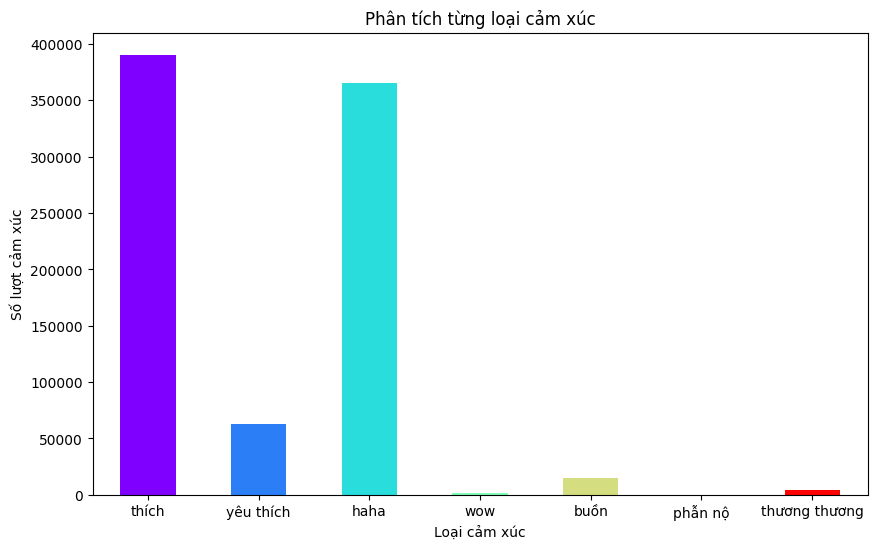


-Số lượng tương tác được tính theo tháng,lượt tương tác vào những tháng mùa Thu là nổi trội hơn cả:



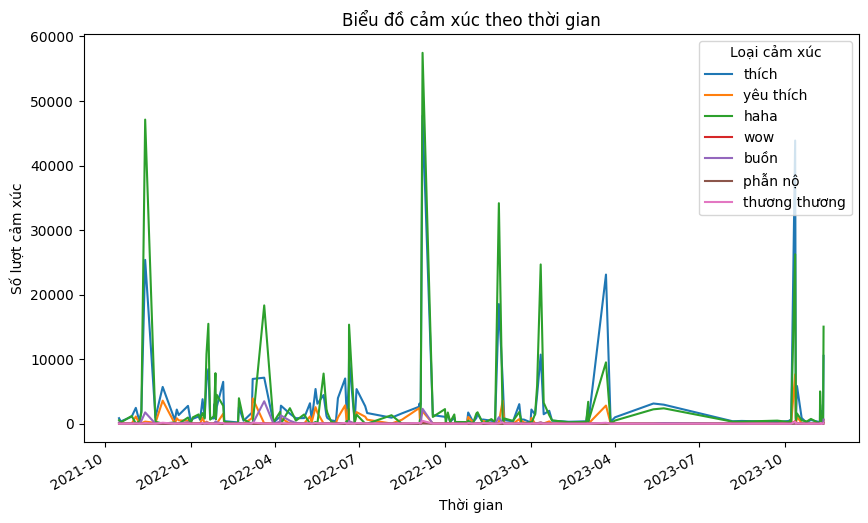
#### 2.3, Bức tranh tổng quan

-Biểu đồ thống kê cảm xúc,những lượt cảm xúc “Wow” hay là “Phẫn nộ” khá là kén người dùng:



-Chúc ta thấy được nội dung của những bài đăng mang thiên hướng vui vẻ nên nhận được số lượng “Haha” rất lớn

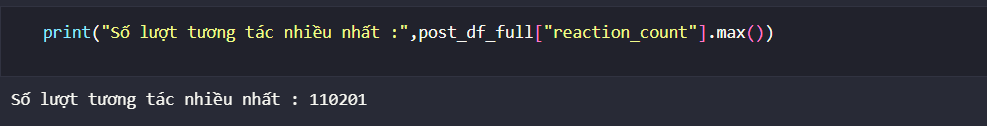
-Biểu đồ cảm xúc theo thời gian:



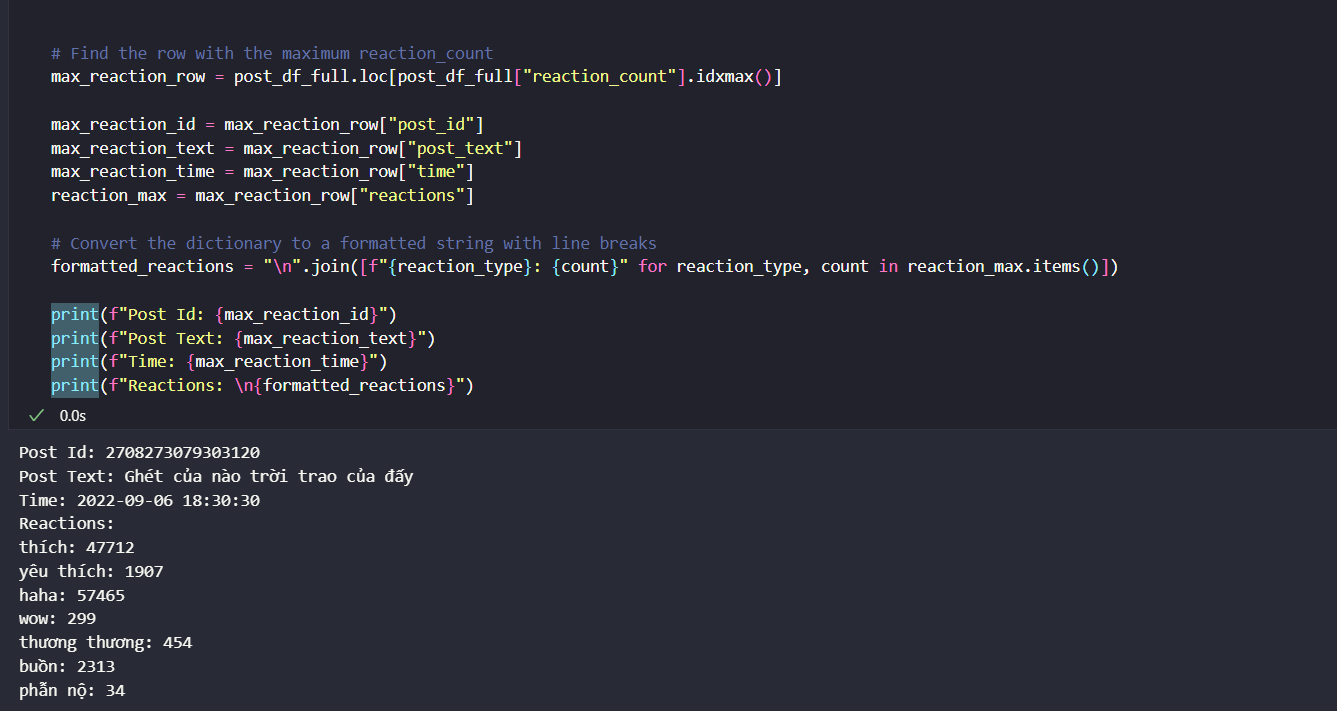
#### 2.4, Max,min của likes,shares và comments

##### 2.4.1, Reactions

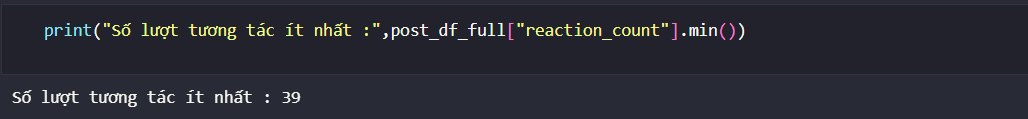
-Chênh lệch rất lớn giữa số bài viết có tương tác nhiều nhất, và bài viết nhận được ít sự quan tâm nhất:

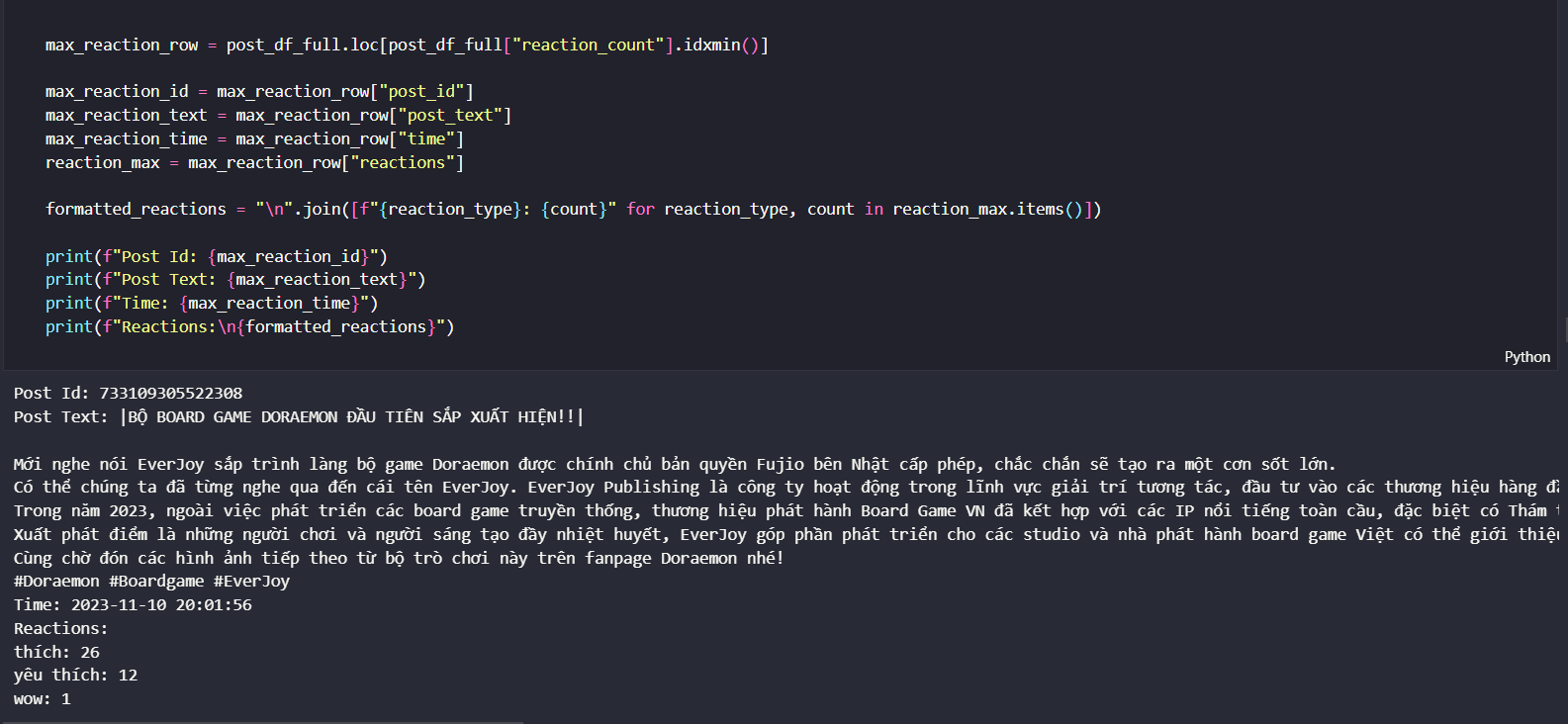


+Thông tin về bài viết có số lượng reactions lớn nhất:



+Bài viết có số lượt reactions ít nhất:



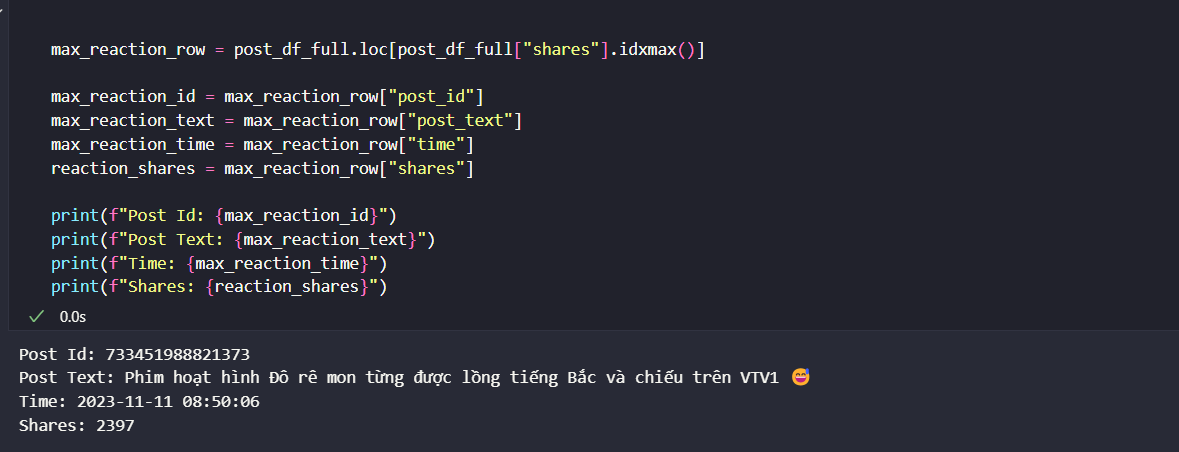


-Những bài viết khác nhận được sự quan tâm lớn:

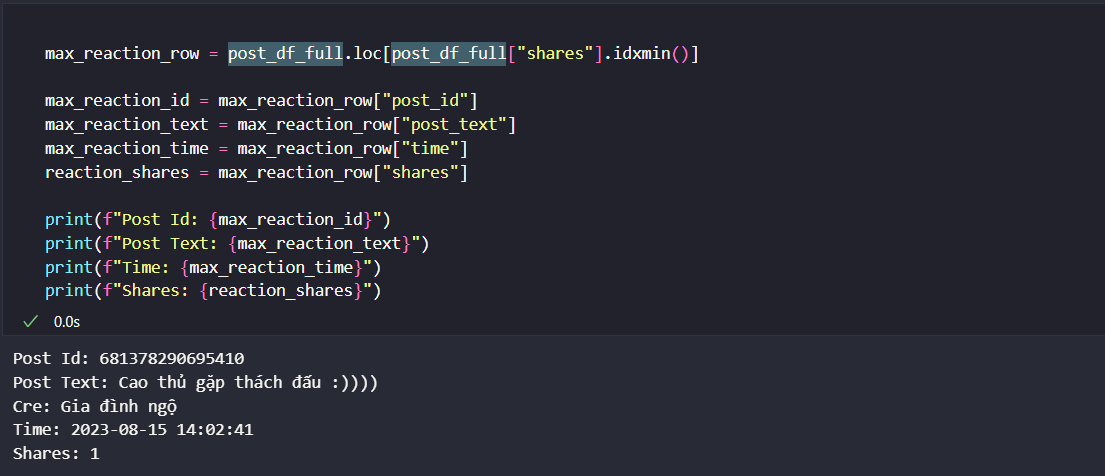


##### 2.4.2, Shares

-Bài đăng có nhiều lượt shares nhất:

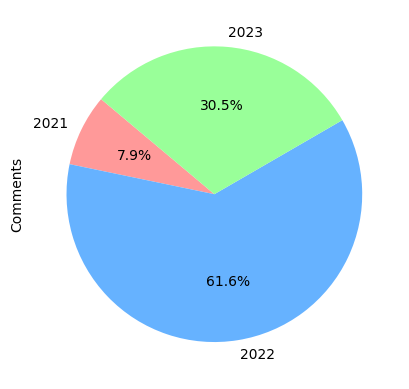


-Bài đăng có ít lượt shares nhất:

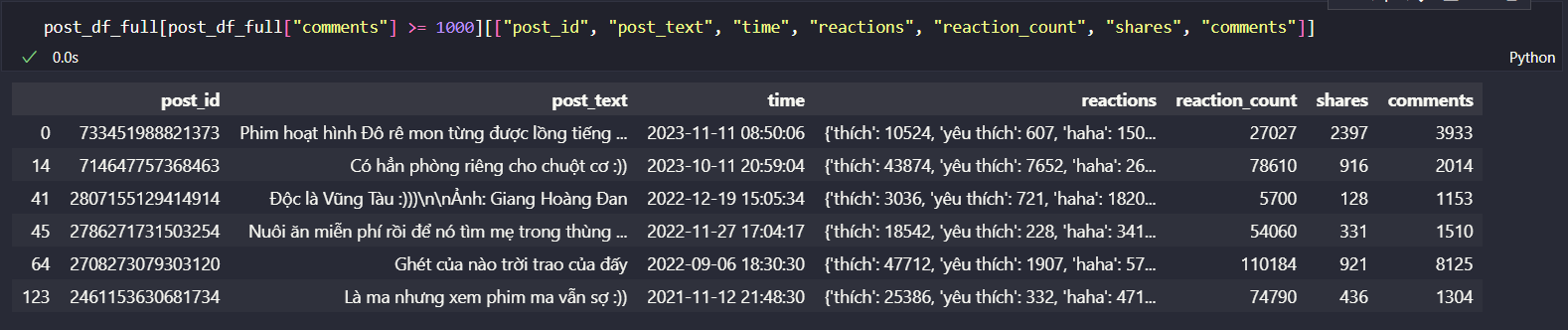


##### 2.4.3, Comments

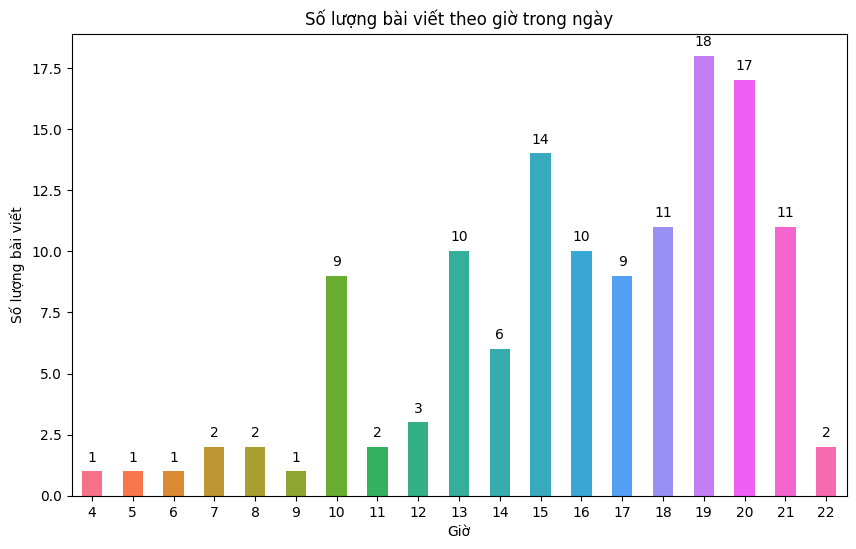
-Do thời gian thu thập chênh lệch nhau, nên số comments cũng chênh lệch khá lớn, biểu đồ chỉ là sự thống kê của số comments đã được thu thập:



-Những bài viết nhận được nhiều lượt reactions , thì cũng nhận được nhiều phản hồi bằng comments cũng như là lượt chia sẻ rất cao:

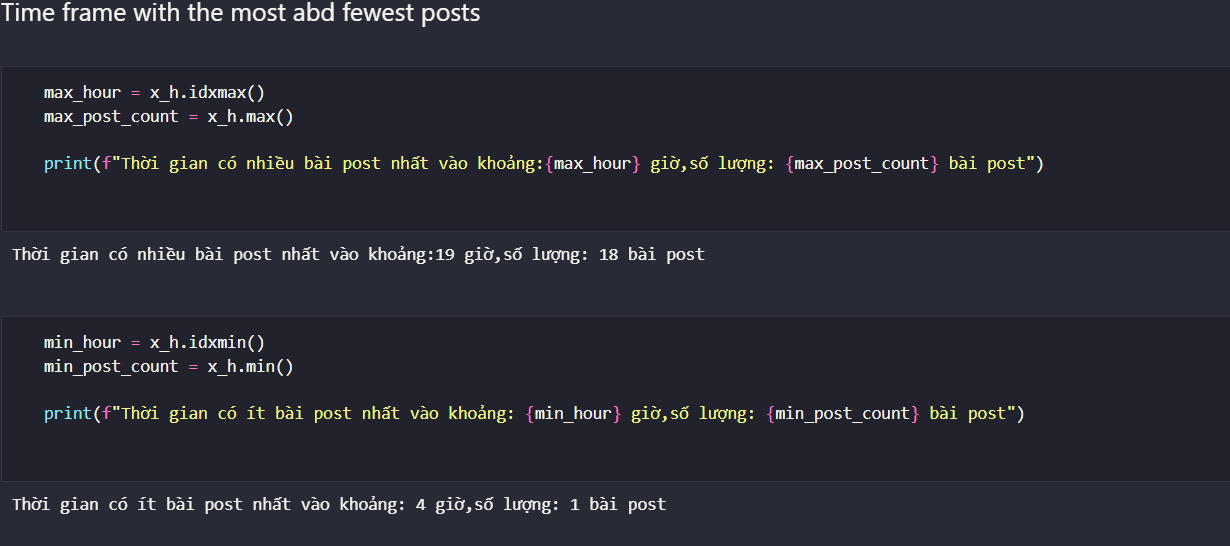


#### 2.5, Regular posting timestamps

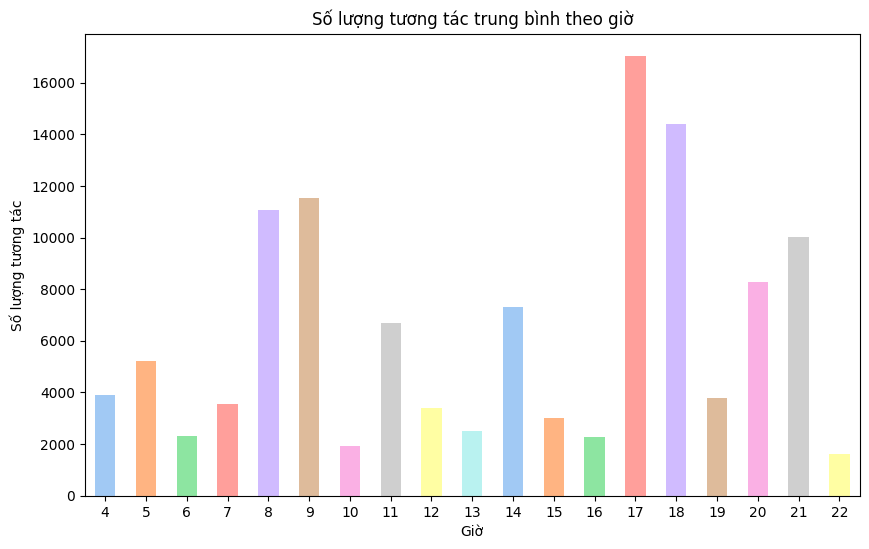


-Những giờ nghỉ ngơi, số bài viết được đăng rất nhiều

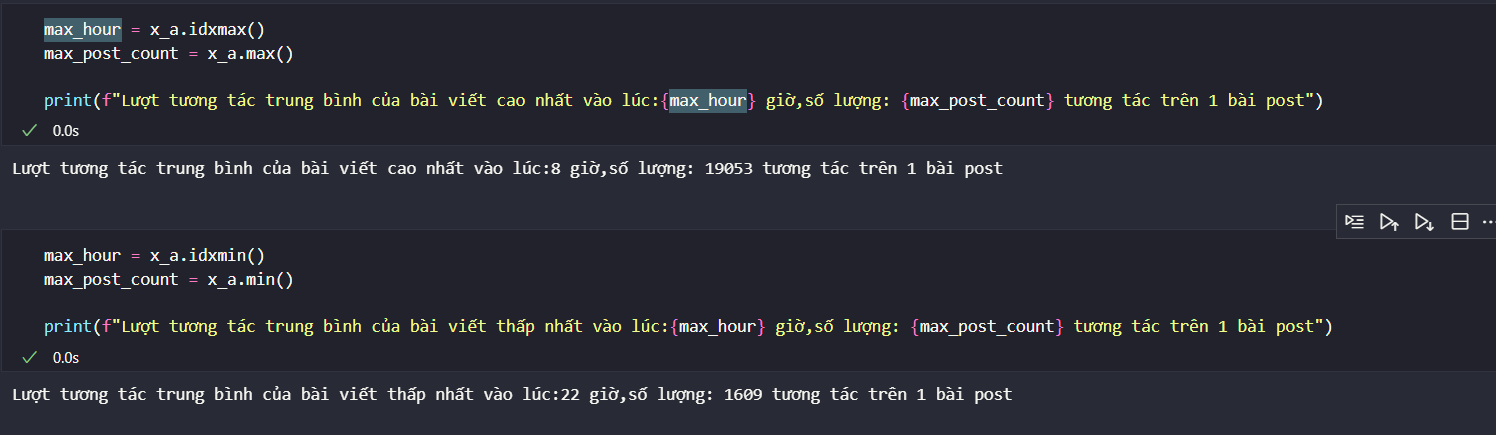
-Những khung giờ đêm khuya và rạng sáng thì không xuất hiện những bài đăng nào



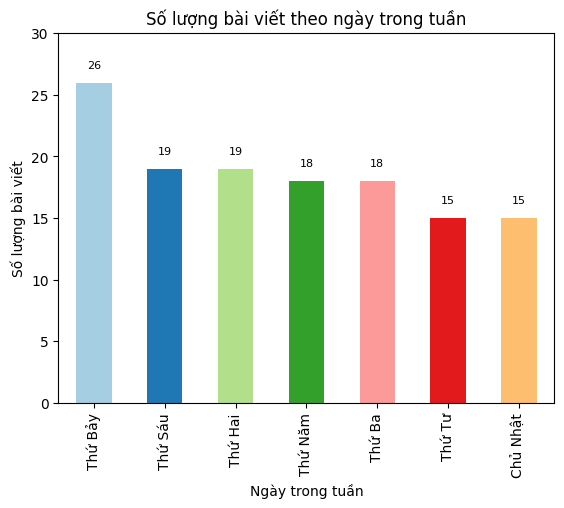
#### 2.6, Tương tác theo giờ



-Vào những khung giờ nghỉ ngơi lượng tương tác khá là ít, và đổ dồn vào những khung giờ còn lại trong ngày



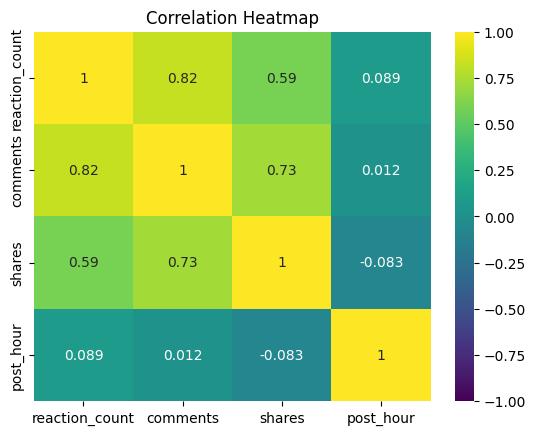
#### 2.7,Thống kê trong tuần



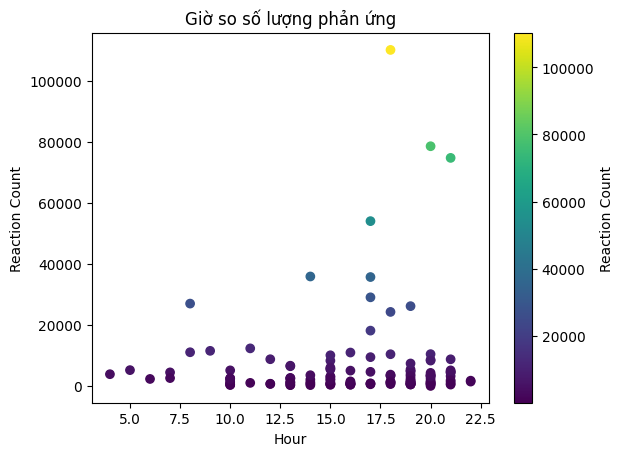
-Không quá ngạc nhiên khi thứ 7 chính là ngày mà số bài viết được đăng lên nhiều nhất,thứ 7 là thứ mà hầu hệt mọi người đều nghỉ ngơi thư giãn sau 1 tuần làm việc mệt mỏi

#### 2.8,Tương quan thời gian đăng vài viết và lượt tương tác

-Bản đồ nhiệt về sự tương quan giữa các trường dữ liệu:



-Thời gian đăng bài so với số lượng reactions



### 3.Phân tích từ khóa trong bài viết

#### 3.1, Số lượng xuất hiện của một vài từ khóa bất kỳ

-Bộ truyện xoay quanh cuộc sống của nhân vật Nobita và bạn than là Doraemon nên chúng ta xem số lần xuất hiện của hai từ khóa này trong các bài đăng của Fanpage:



#### 3.2, Tần suất các từ được sử dụng nhiều nhất

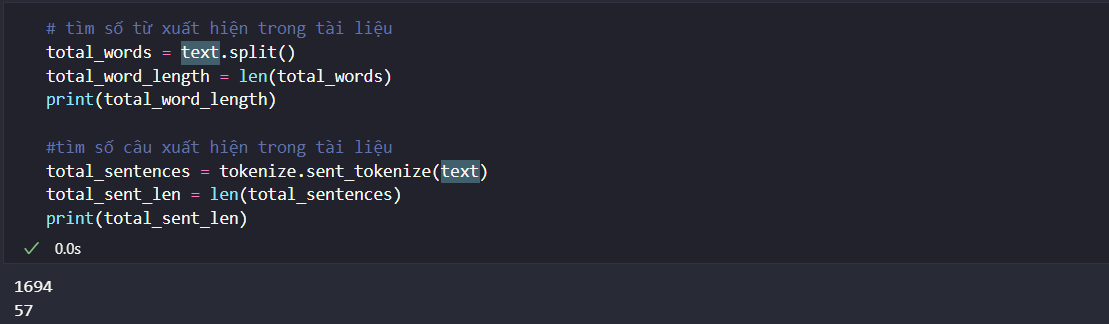


-Từ khóa Doraemon là 1 trong từ khóa xuất hiện gần như là nhiều nhất trong các bài post không quá ngạc nhiên vì Page này được lập ra là dành cho chú mèo máy cũng như bộ truyện cùng tên

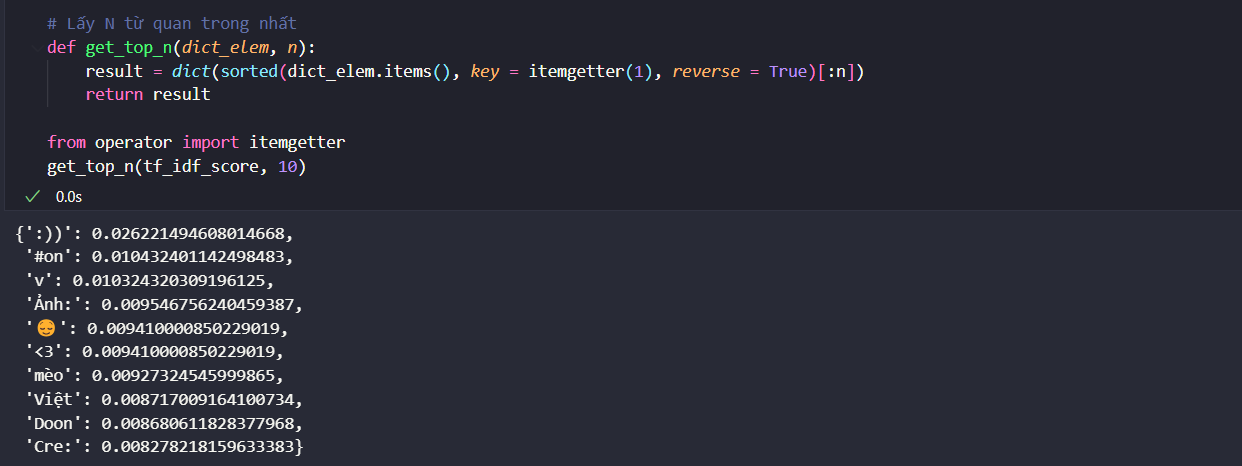
-Bên cạnh đó những từ khóa như : “Bộ”,”Hình”,”Movie”,… là những từ khóa liên quan đến bộ phim cũng xuất hiện khá nhiều

#### 3.3,Tìm mức độ quan trong của các từ trong bài post

-Số từ và câu xuất hiện trong các bài post của page,những bài post khá là ngắn sau khi chúng ta thấy được số từ và số câu mà Page đã sử dụng:

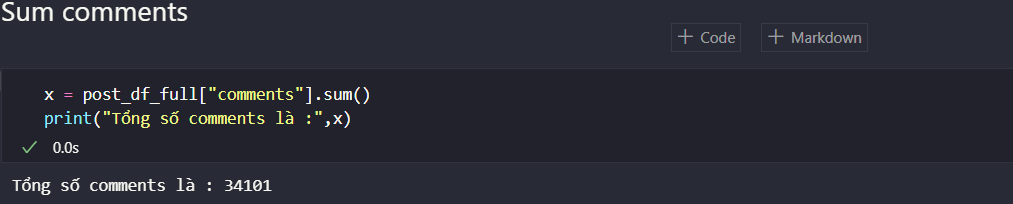


-Sử dụng file stop\_word và file stop\_word1 được lấy từ stopword Việt nam để tìm mức độ quan trọng của các từ so với bài viết:

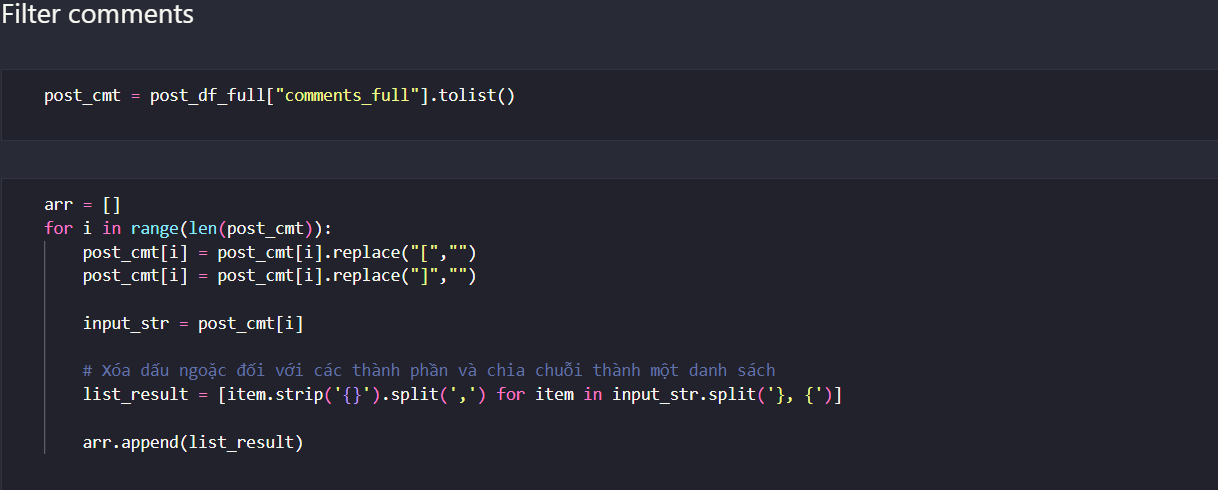


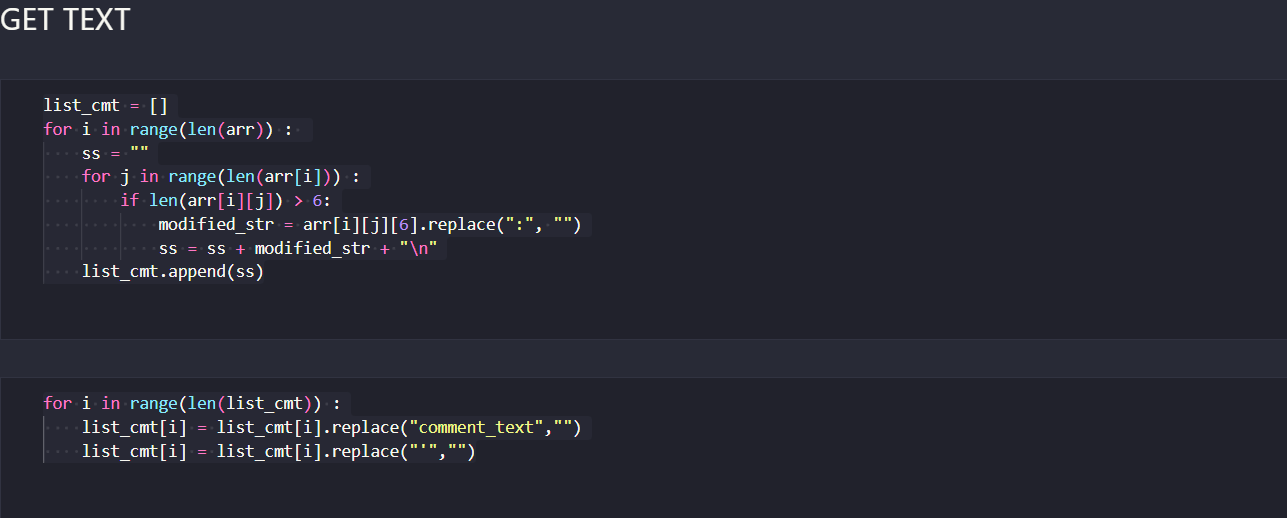
## Phần 3 : Phân tích, đánh giá comments

### 1.Tổng số comments

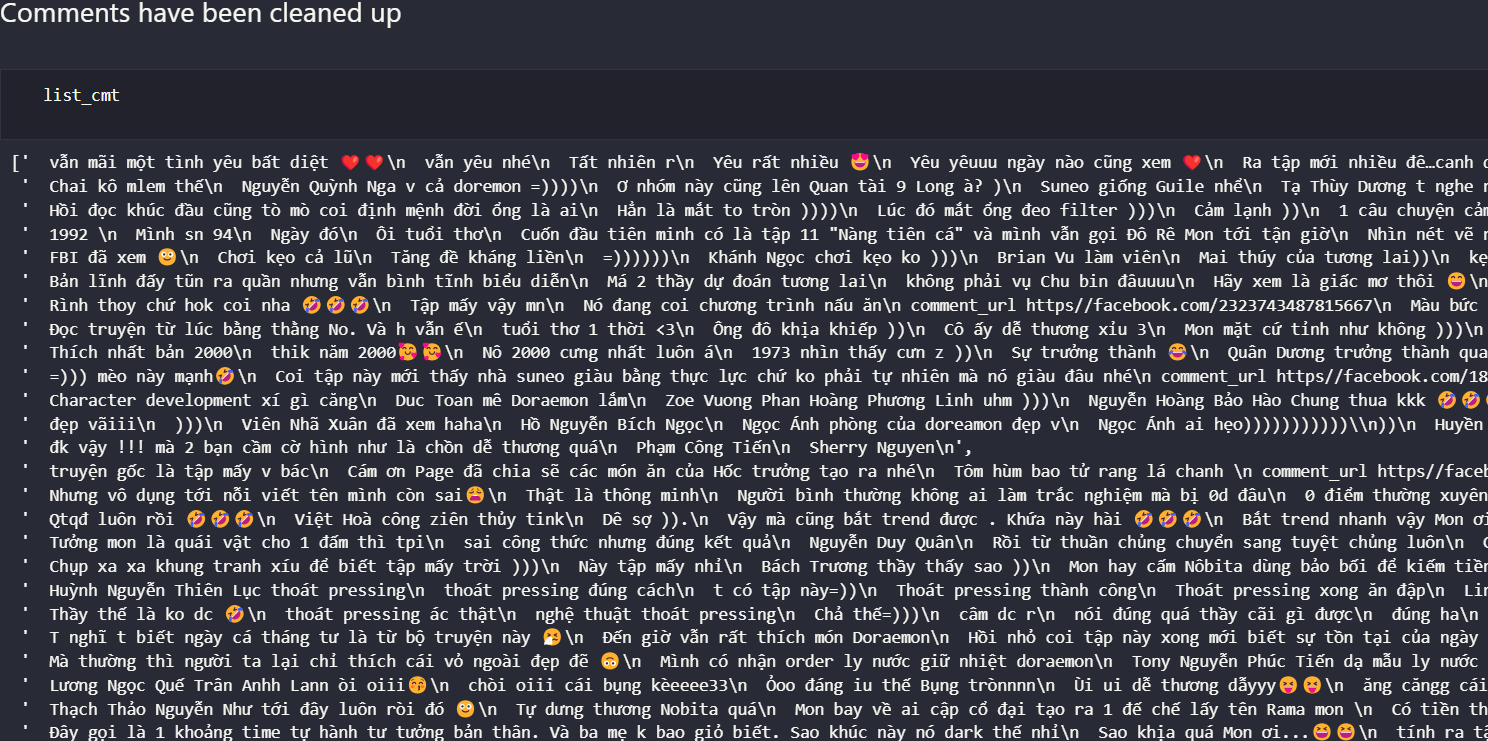


### 2.Lọc comments





-Hiển thị toàn bộ comments sau khi đã được lọc và làm sạch:



-Commens sau khi được lọc và làm sạch chúng ta thấy xuất hiện rất nhiều icon vui nhộc, thân thiện mà giới trẻ hay sử dụng bây giờ

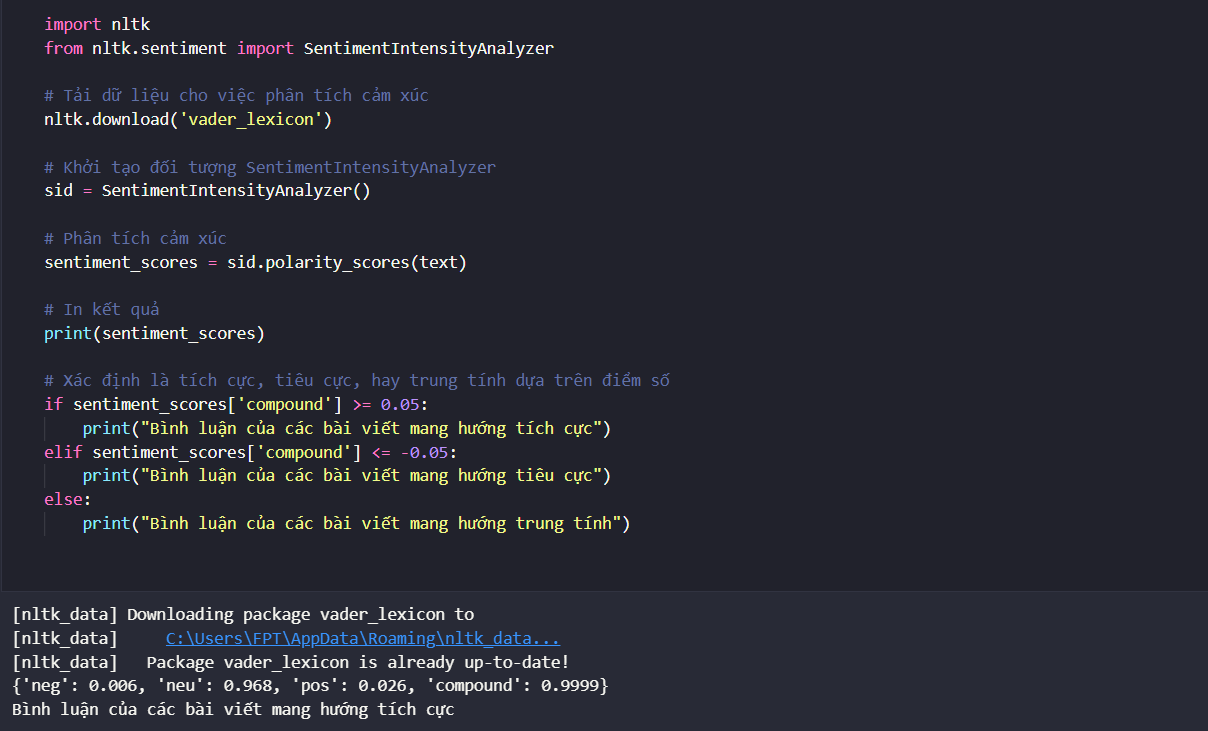
### 3.Wordclound comments

-Tần suất xuất hiện của những từ trong comments:



### 4,Phân tích độ tích cực, tiêu cực của các comments

-Những comments được đánh giá là tích cực sau khi phân tích bằng SentimentAnalyzer:



•Các chỉ số được tính toán ra là:

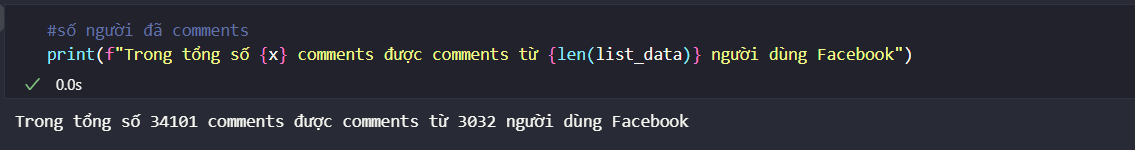
NEG:0.006

NEU:0.968

POS:0.026

### 5.Phân tích người bình luận trong các bài Post

#### 5.1, Số người dùng comments



-Từ 34101 bình luận thì có 3032 sử dụng tính năng bình luận, tính ra mỗi người đã comments hơn 10 lần trong bài viết, một con số khá tương đối.Chứng tỏ độ tương tác qua lại của người dùng khá là cao, thân thiện với nhau

#### 5.2, Phân tích

##### 5.2.1, Lấy ID người dùng

-Sử dụng những kĩ năng xử lý chuỗi đã học được để tách ID ra :



-ID sau khi đã tách ra được :

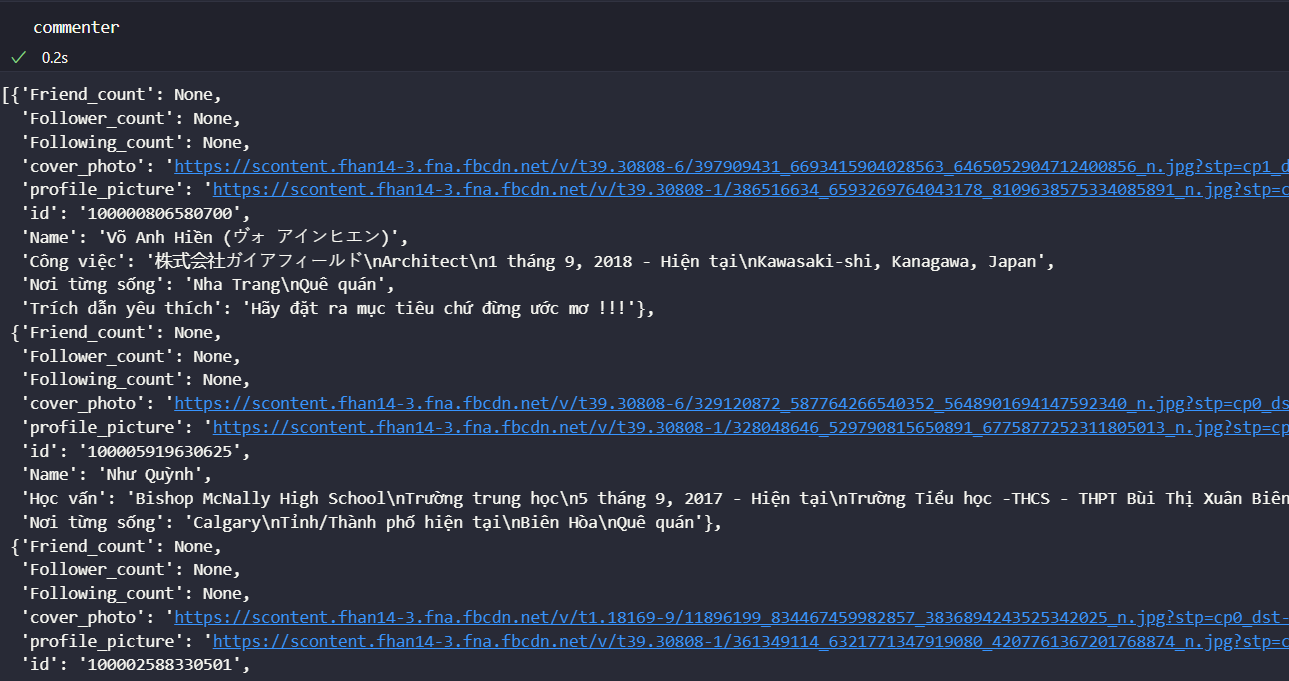


##### 5.2.2, Lấy thông tin người dùng

-Chúng ta sử dụng hàm get\_profile để lấy thông tin của người dùng, do thông tin người dùng là bảo mật, nên chúng ta chỉ lấy được một số dữ liệu hạn chế:

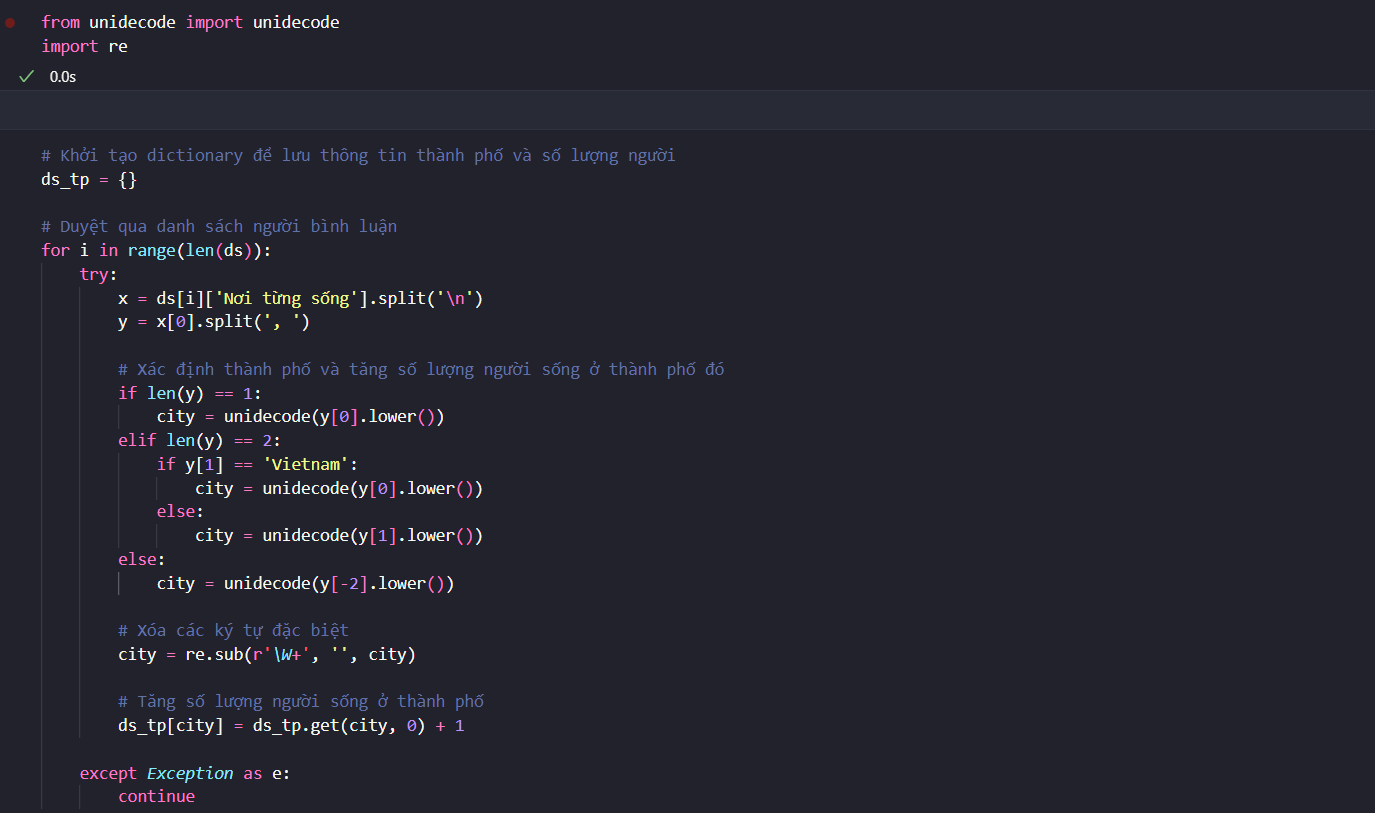


-Những thông tin thu thập được :

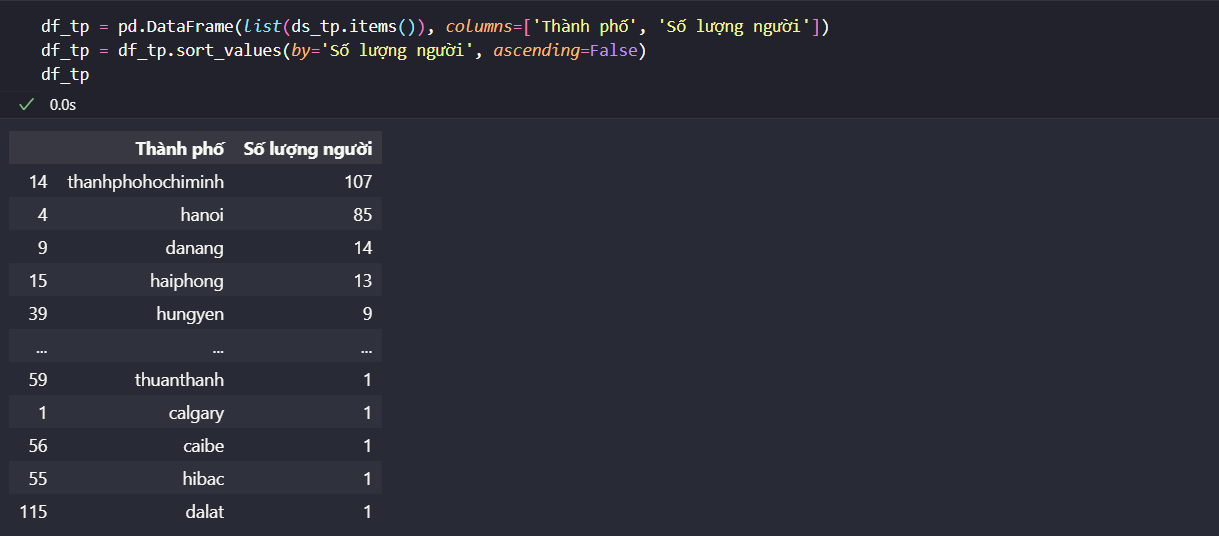


##### 5.2.3, Phân tích

-Sau khi lấy được 1 số thông tin về những người comments, chúng ta cùng qua một vài bước xử lý và lọc ra được thành phố của họ đang sinh sống :



-Thông tin về những thành phố đã được tổng hợp :



-Qua số liệu trên chúng ta thấy được Page thu hút rất nhiều những người đến từ các thành phố khác nhau, có cả những người đang làm việc và sinh sống tại nước ngoài.Nhưng số lượng đông nhất vẫn đến từ 2 thành phố lớn nhất Việt Nam đó là “Thành Phố Hồ Chí Minh”và “Hà Nội”

