CRAWL DỮ LIỆU TỪ FACEBOOK

[Phần A : các công cụ được sử dụng 3](#_Toc150548263)

[Phần 1: vấn đề 3](#_Toc150548264)

[Phần 2: công cụ 3](#_Toc150548265)

[Phần B : thực thi 7](#_Toc150548266)

[Phần 1 : Crawl dữ liệu 7](#_Toc150548267)

[1.Page dùng để crawl 7](#_Toc150548268)

[2.Phương pháp 8](#_Toc150548269)

[3.Dữ liệu crawl được 8](#_Toc150548270)

[Phần 2 : Phân tích , trực quan hóa dữ liệu 10](#_Toc150548271)

[1.Tiền xử lý dữ liệu 10](#_Toc150548272)

[1.1,Xóa bỏ những dòng dữ liệu bị lỗi 10](#_Toc150548273)

[1.2,Xóa bỏ những cột bị mất nhiều dữ liệu và định dạng 10](#_Toc150548274)

[1.2.1,Loại bỏ dữ liệu 10](#_Toc150548275)

[1.2.2,Định dạng dữ liệu 12](#_Toc150548276)

[2.Thống kê và trực quan hóa 14](#_Toc150548277)

[2.1, Số lần tương tác, cái nhìn khái quát 14](#_Toc150548278)

[2.1.1, Thời gian thu thập 14](#_Toc150548279)

[2.1.2, Một số dữ liệu đáng chú ý 14](#_Toc150548280)

[2.1.3, Cái nhìn khái quát 15](#_Toc150548281)

[2.2, Reactions hàng ngày 15](#_Toc150548282)

[2.3, Bức tranh tổng quan 16](#_Toc150548283)

[2.4, Max,min của likes,shares và comments 18](#_Toc150548284)

[2.4.1, Reactions 18](#_Toc150548285)

[2.4.2, Shares 18](#_Toc150548286)

[2.4.3, Comments 19](#_Toc150548287)

[2.5, Regular posting timestamps 20](#_Toc150548288)

[2.6, Tương tác theo giờ 21](#_Toc150548289)

[3.Phân tích từ khóa trong bài viết 22](#_Toc150548290)

[3.1, Số lượng xuất hiện của một vài từ khóa bất kỳ 22](#_Toc150548291)

[3.2, Tần suất các từ được sử dụng nhiều nhất 22](#_Toc150548292)

[3.3,Tìm keyword 23](#_Toc150548293)

[Phần 3 : Phân tích, đánh giá comments 24](#_Toc150548294)

[1.Tổng số comments 24](#_Toc150548295)

[2.Lọc comments 24](#_Toc150548296)

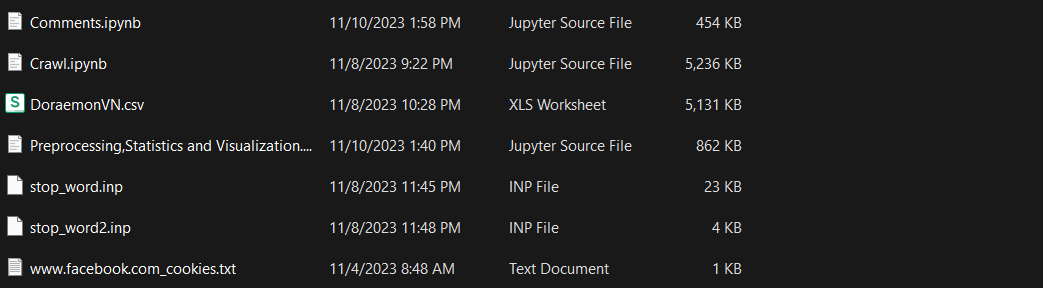
[3.Wordclound comments 25](#_Toc150548297)

# 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 cô.

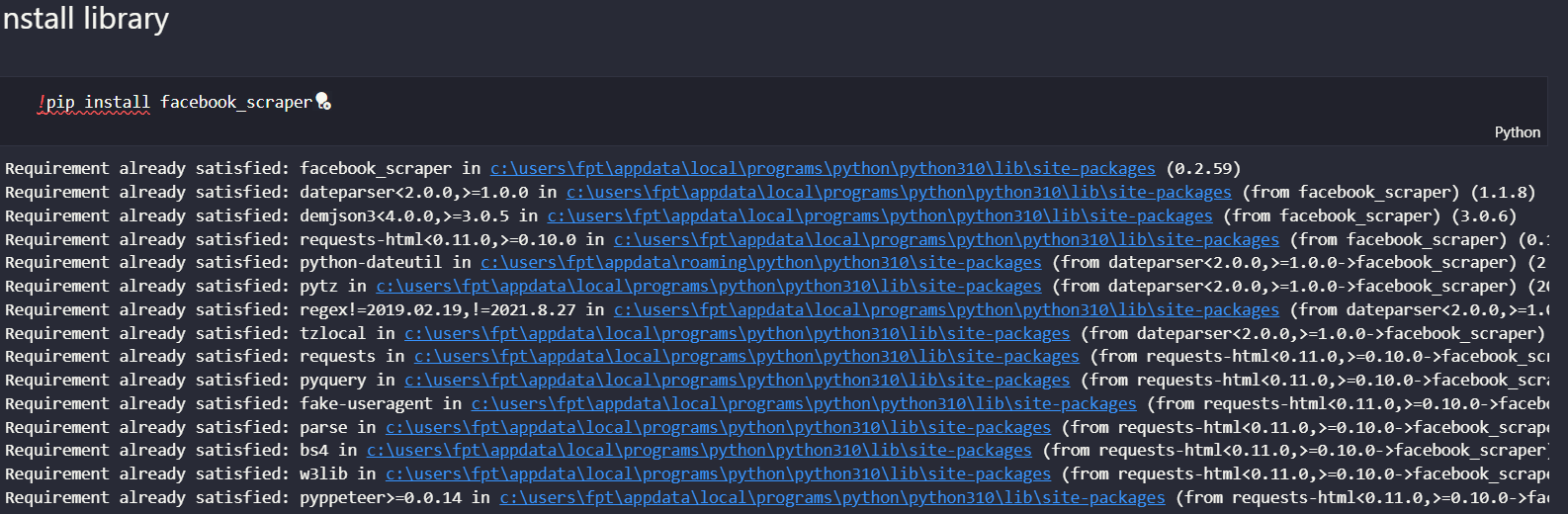
-Project bao gồm các file dữ liệu:



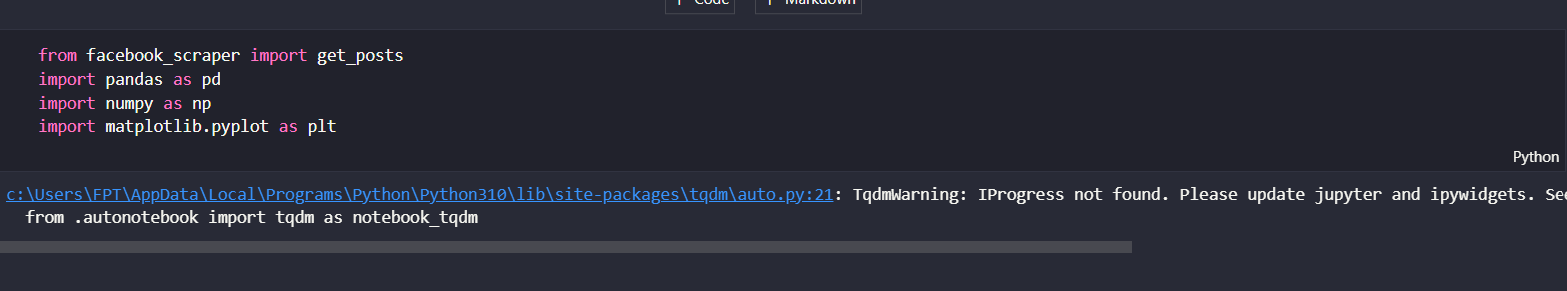
## Phần 2: công cụ

1.Crawl

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

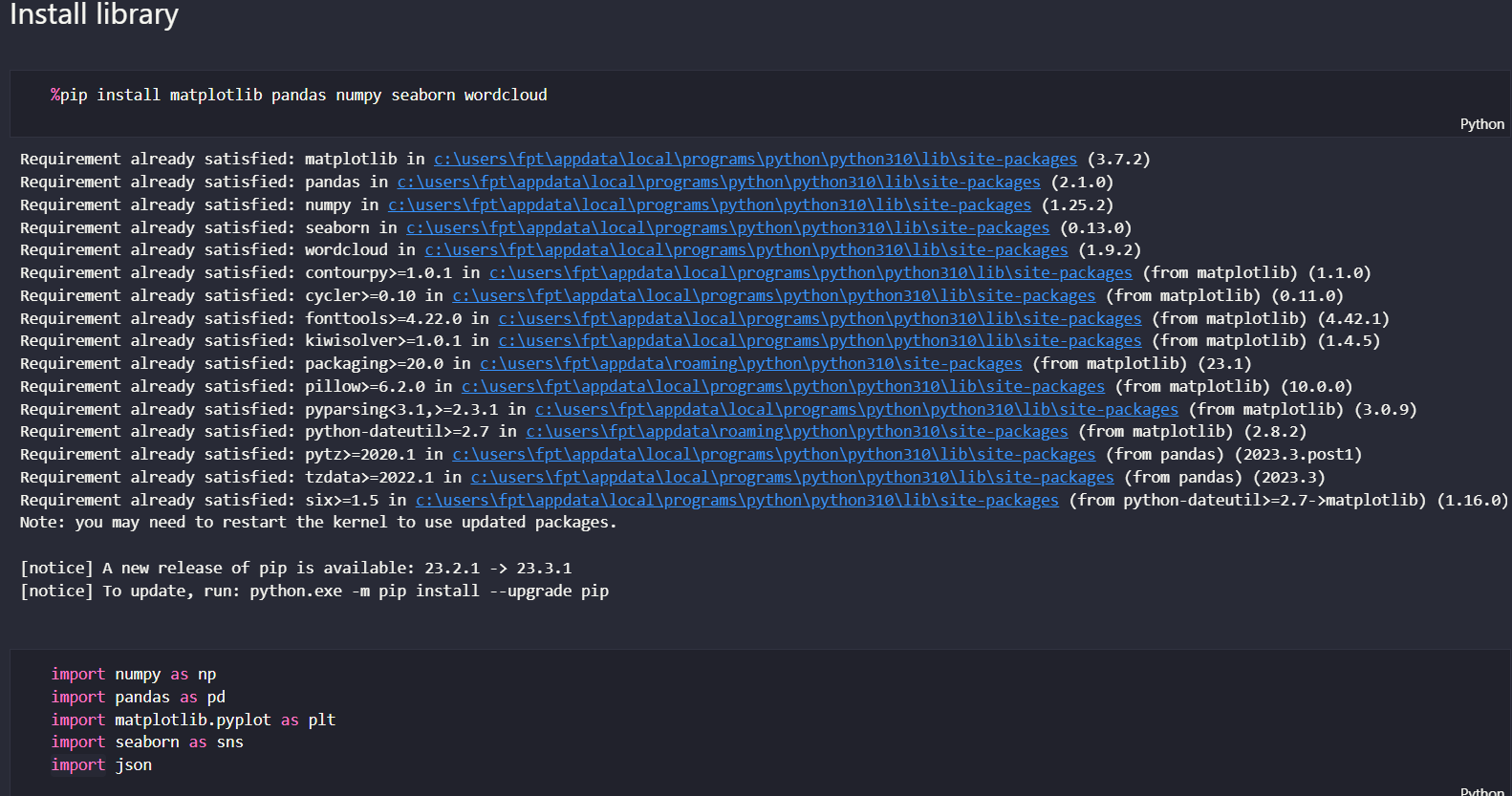


-Import các thư viện cần thiết :



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.

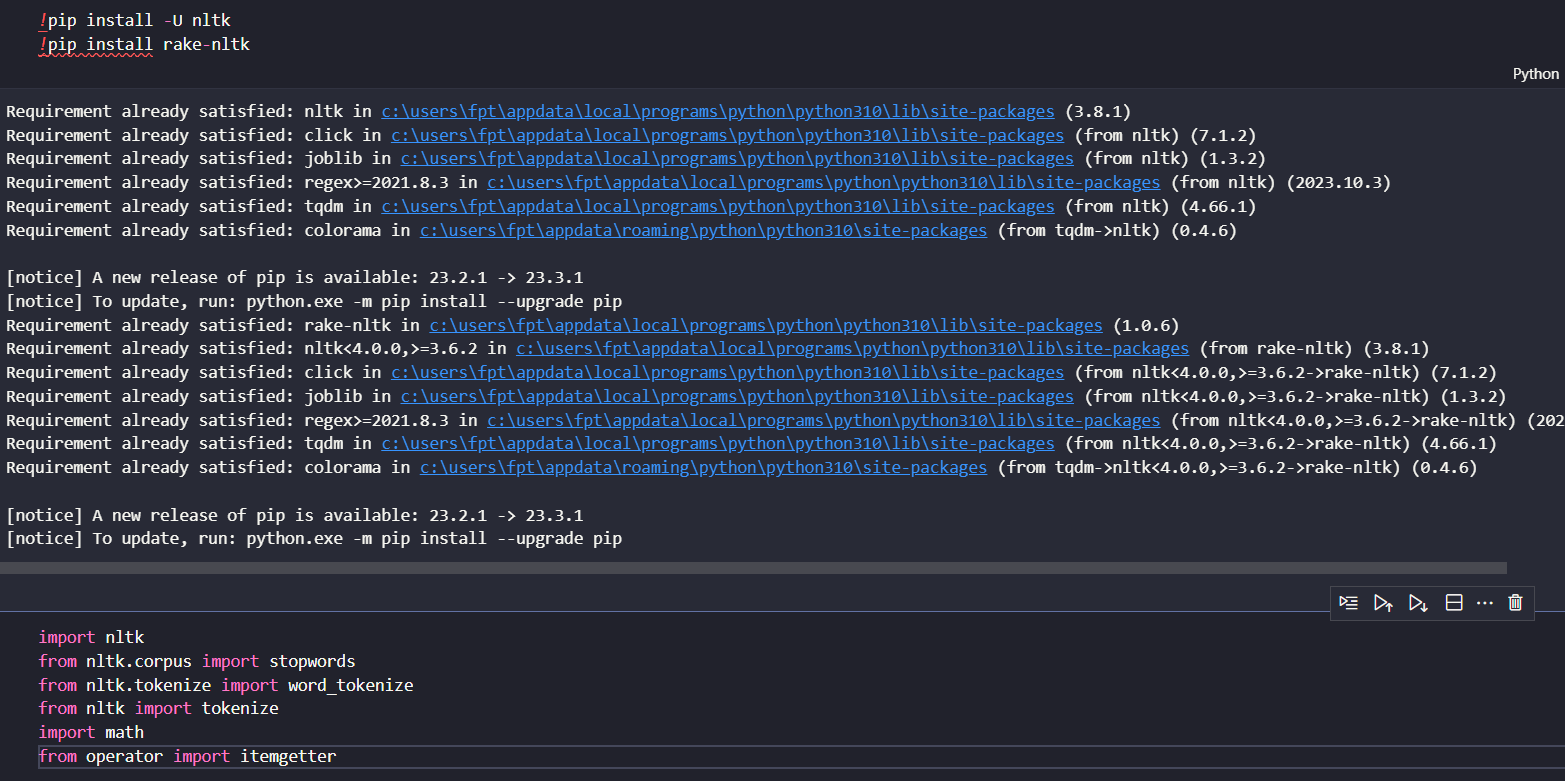
+Seaborn:

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.

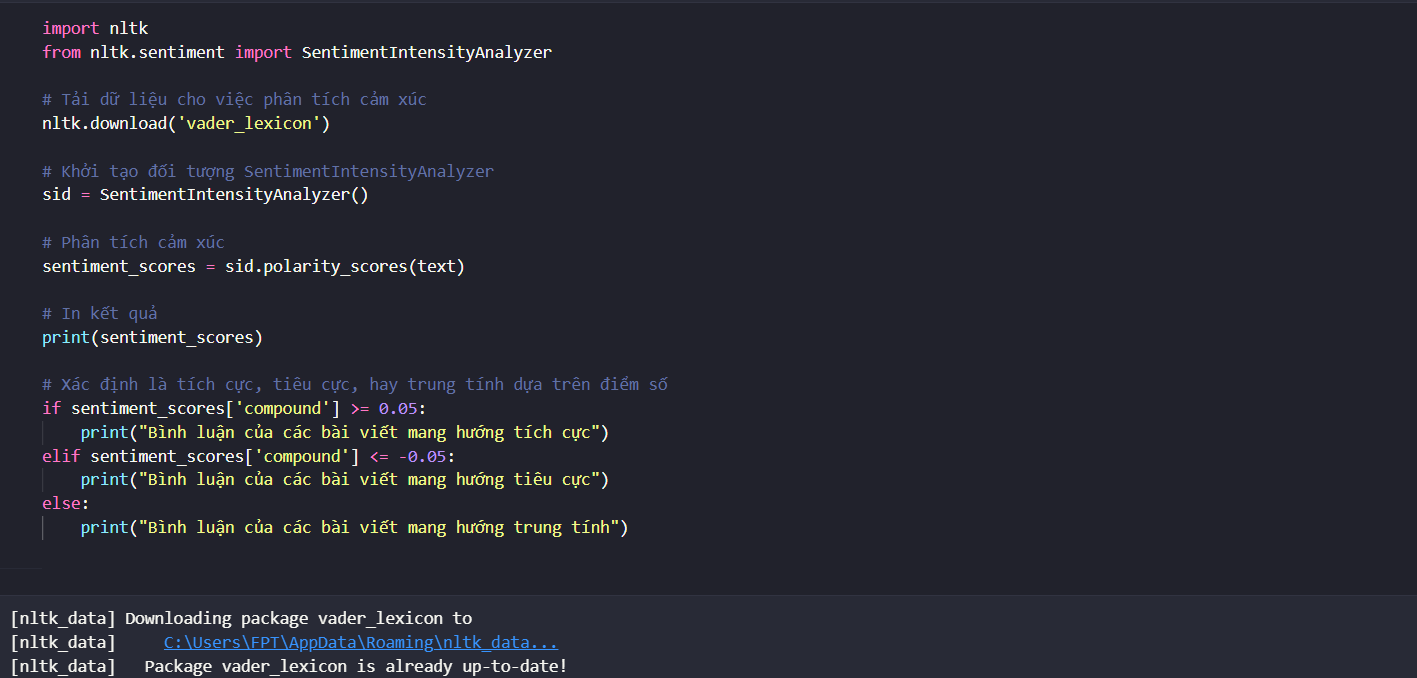
+WordCloud:  
Tác dụng: WordCloud là một thư viện trực quan hóa dữ liệu, tạo ra hình ảnh từ các từ xuất hiện trong văn bản.  
Ứng dụng: Hiển thị các từ quan trọng trong văn bản, đặc biệt thường được sử dụng trong phân tích ý kiến và trực quan hóa từ khóa.

-Sử dụng stopword để tìm keyword:



3.Analysis

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

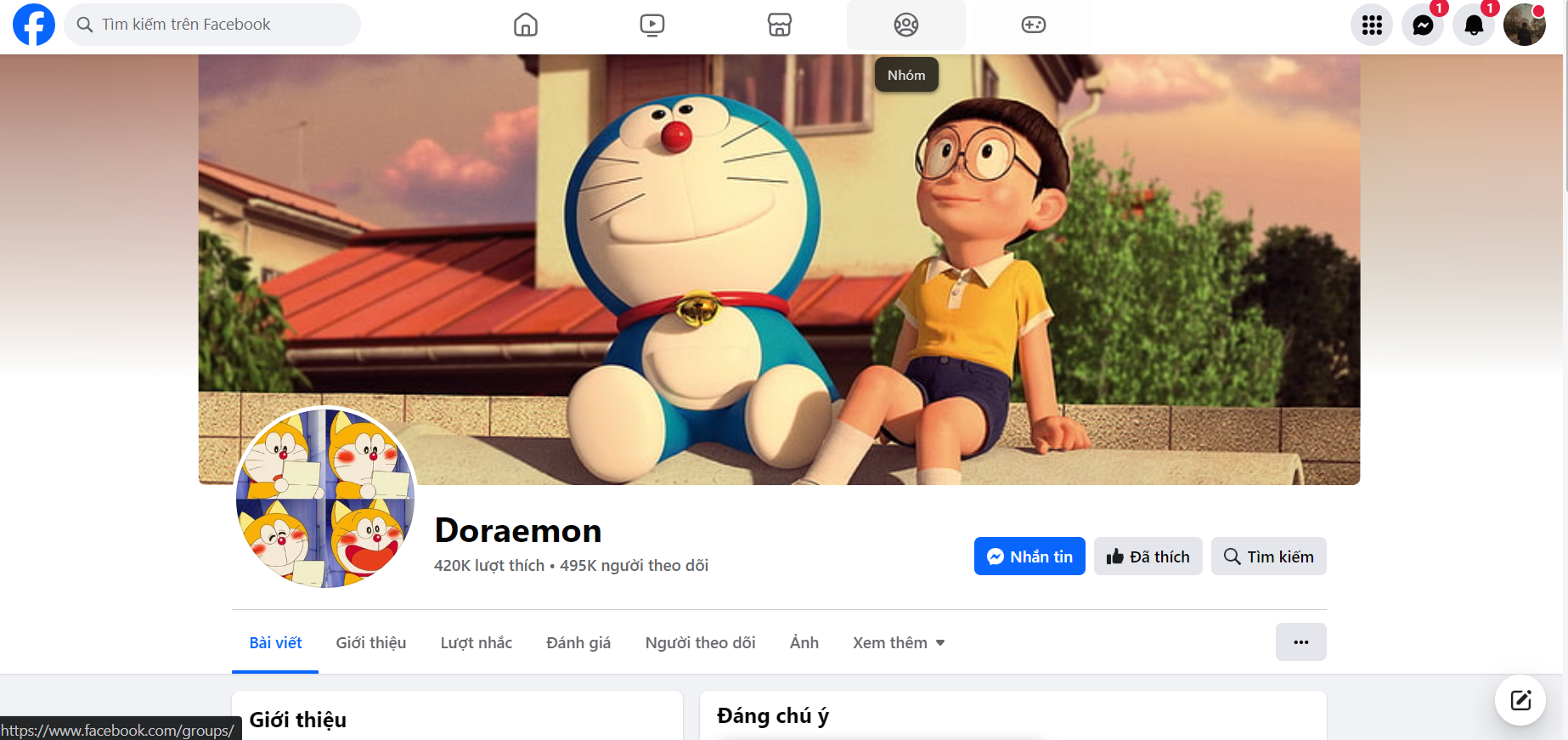


# Phần B : thực thi

## Phần 1 : Crawl dữ liệu

### 1.Page dùng để crawl

-Trang fanpage được chọn là DoraemonVietNam :

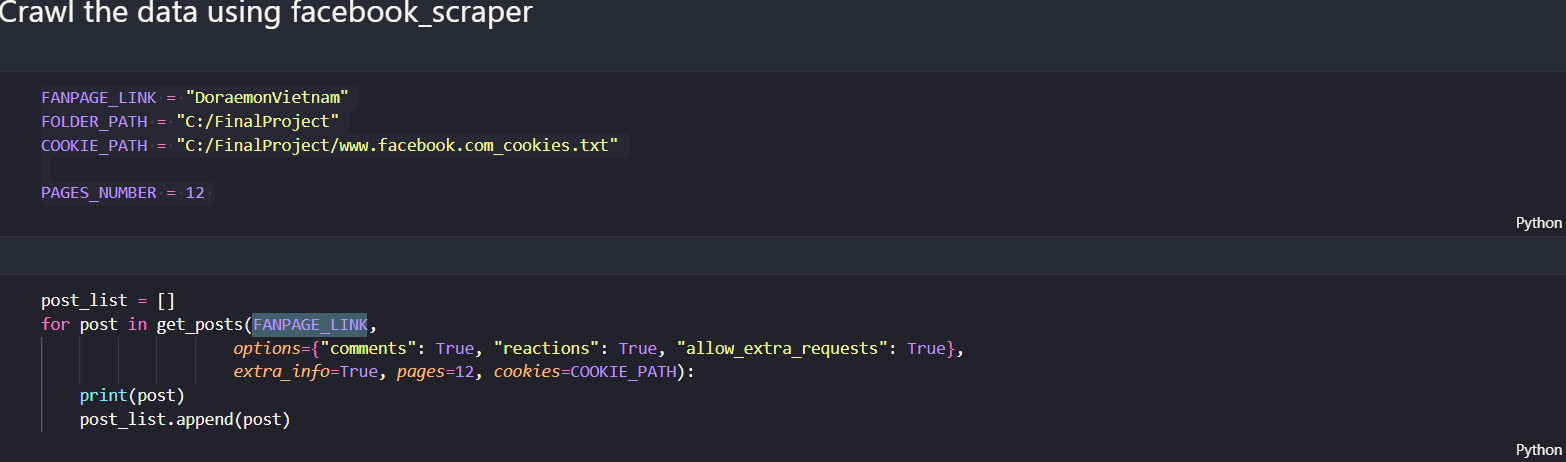


### 2.Phương pháp

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

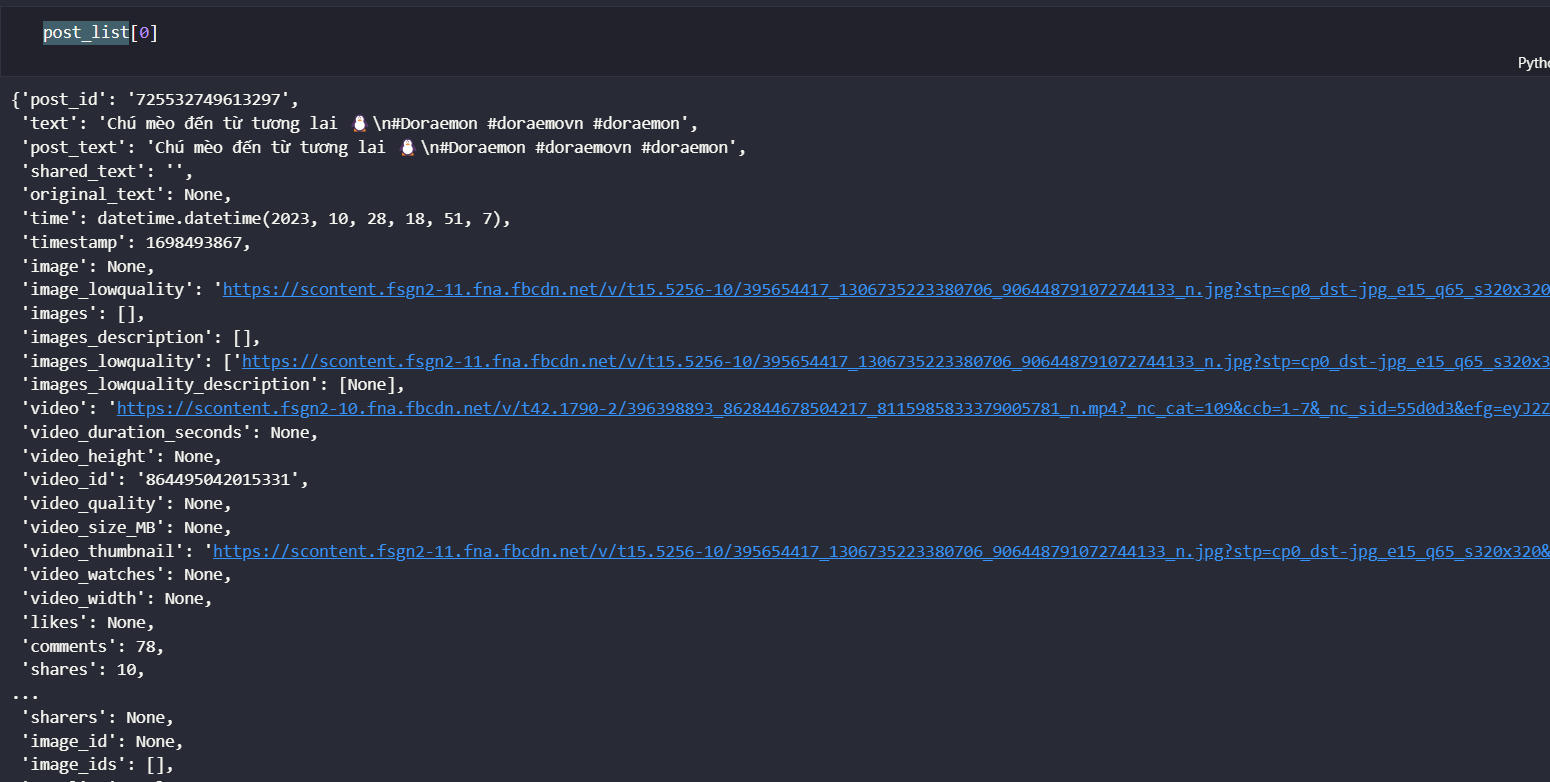
-COOKIE\_PATH: Đường dẫn đến tệp cookie mà chúng tôi sẽ sử dụng để xác thực với Facebook. Tệp cookie này có thể lấy được bằng cách đăng nhập vào Facebook và sao chép cookie từ trình duyệt.

-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.

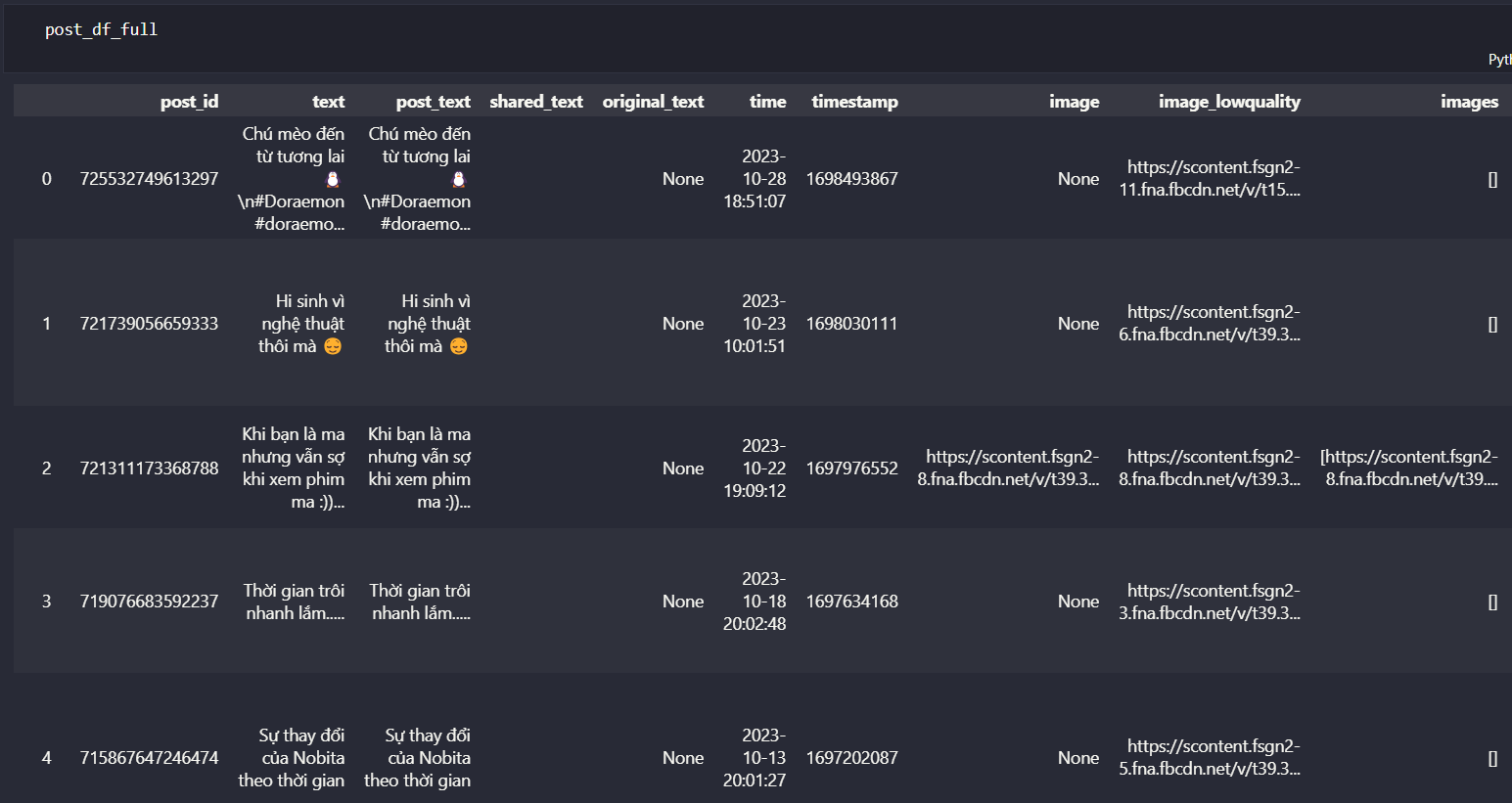


### 3.Dữ liệu crawl được

-Thông tin của 1 post:



-Thông tin crawl được:

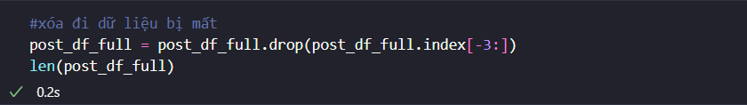


## Phần 2 : Phân tích , trực quan hóa dữ liệu

### 1.Tiền xử lý dữ liệu

#### 1.1,Xóa bỏ những dòng dữ liệu bị lỗi

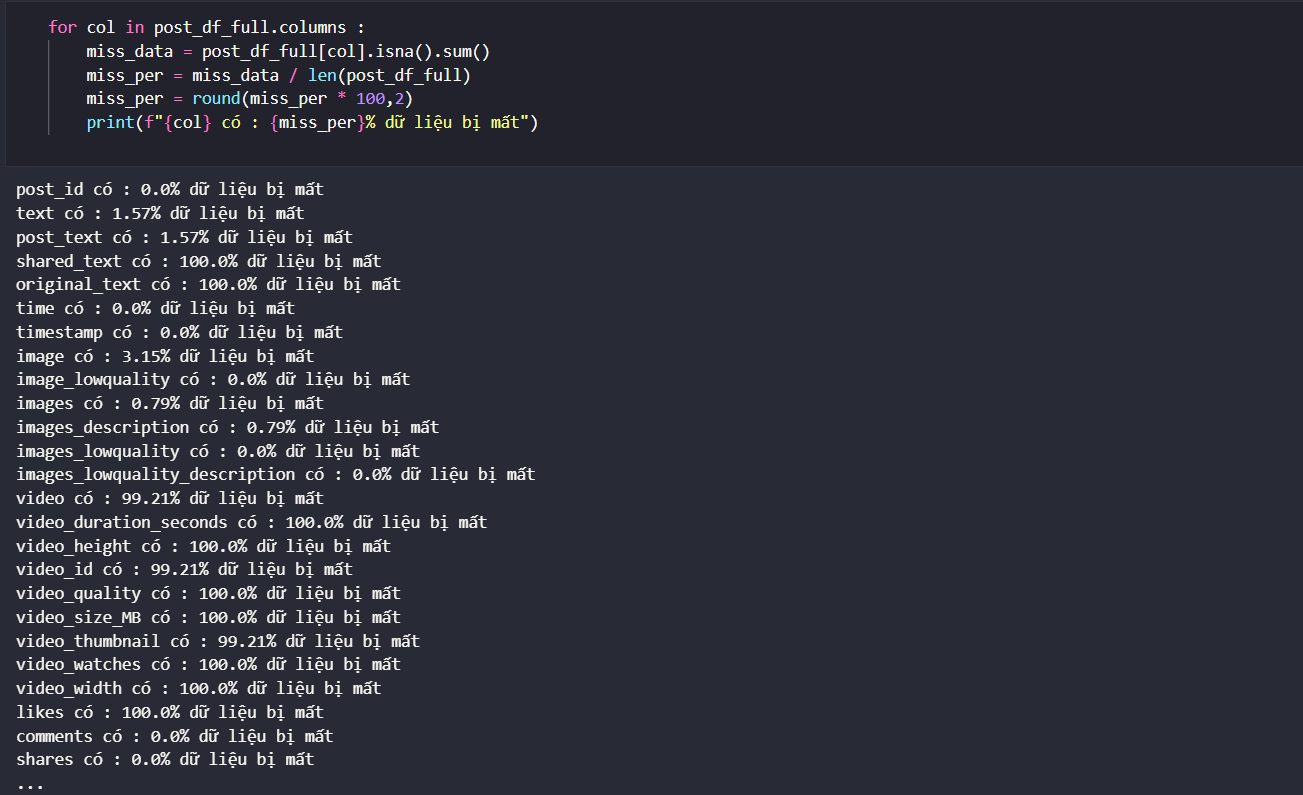
-Những dòng cuối dữ không có dữ liệu đã được loại bỏ:



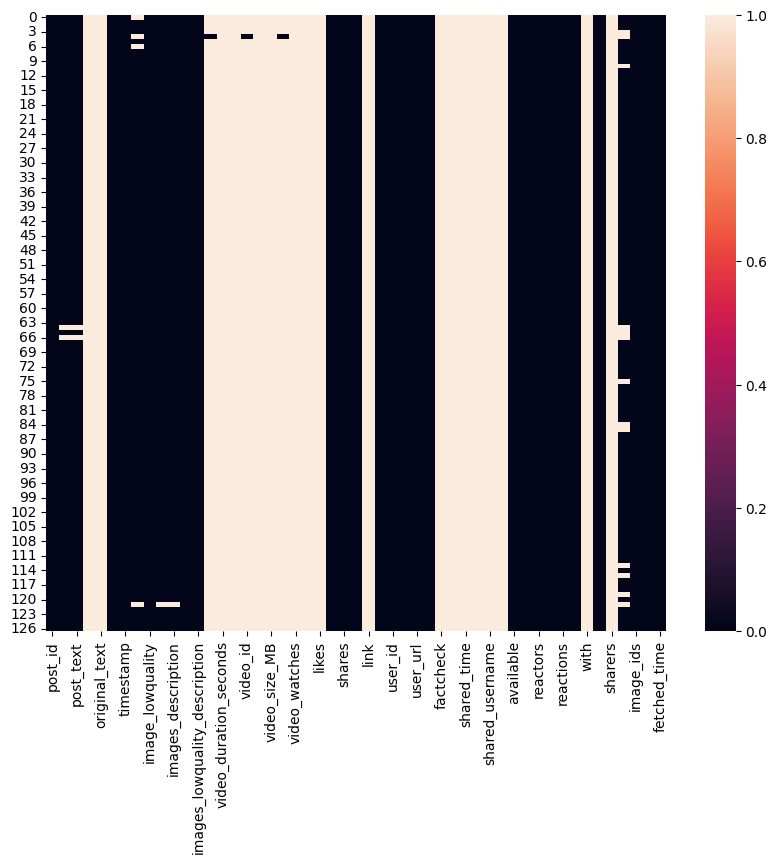
#### 1.2,Xóa bỏ những cột bị mất nhiều dữ liệu và định dạng

##### 1.2.1,Loại bỏ dữ liệu

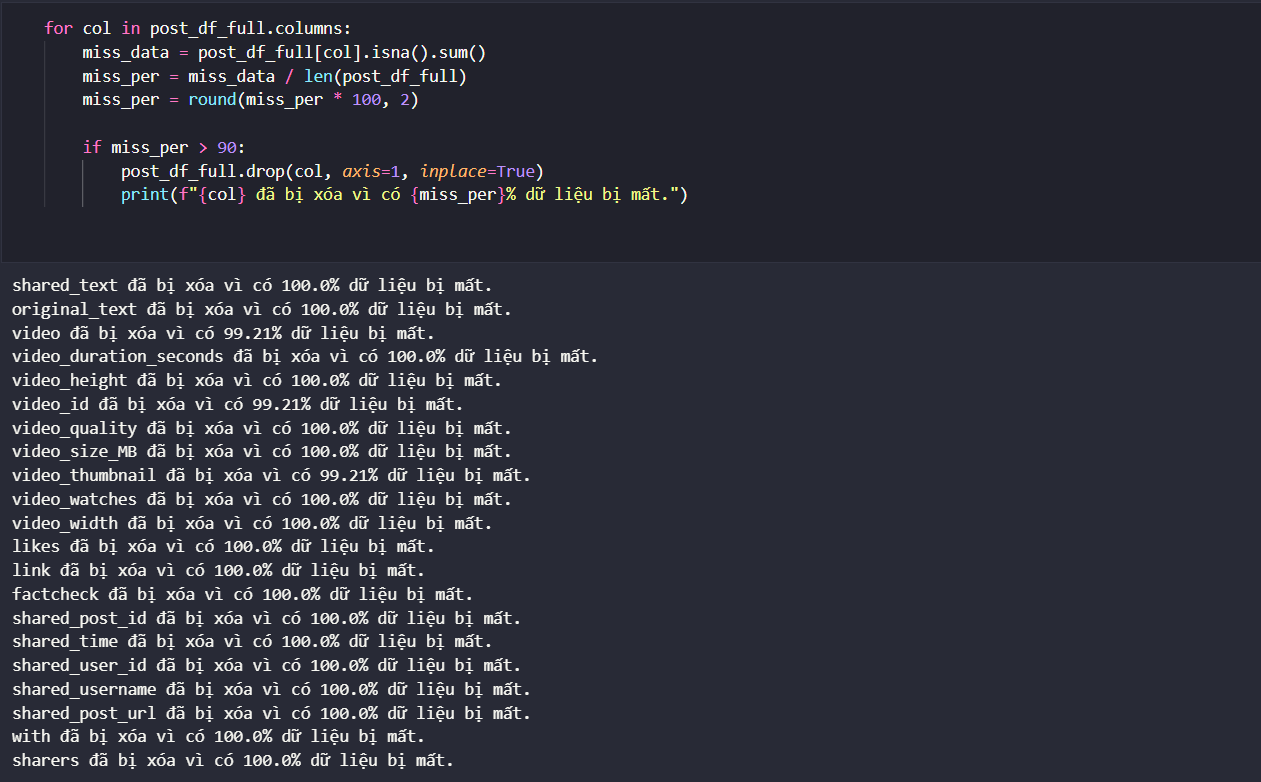
-Xác định dữ liệu bị mất:



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

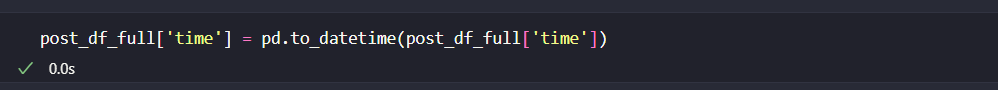


-Xóa những dự liệu không đủ để sử dụng:

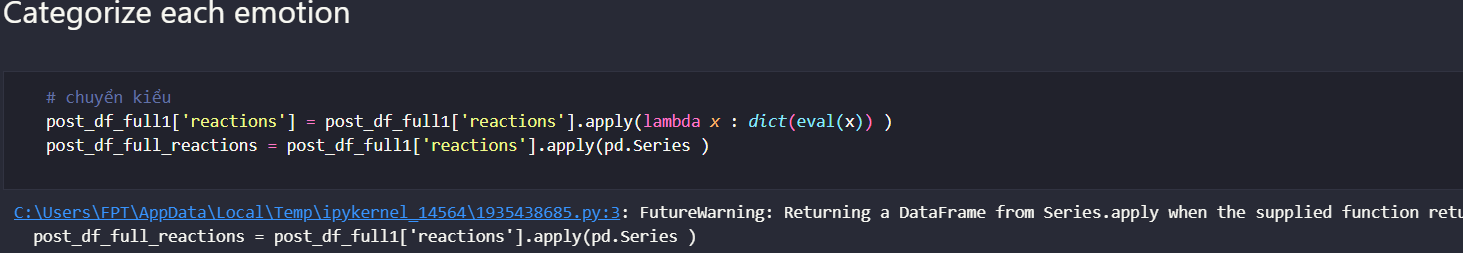


##### 1.2.2,Định dạng dữ liệu

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

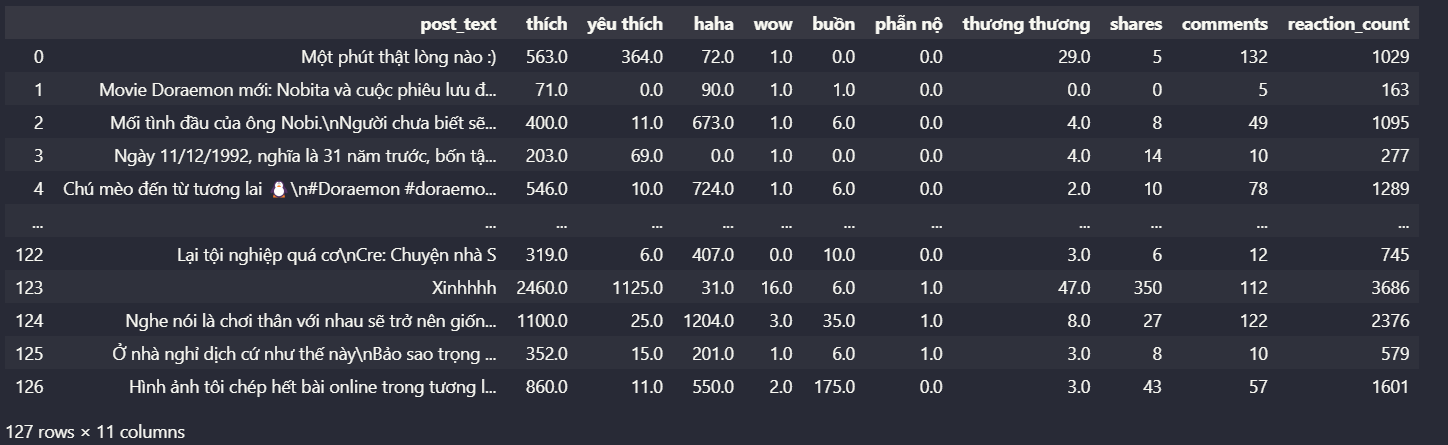


-Tách cột reactions thành những cột cảm xúc riêng biệt để dễ dàng thao tác:





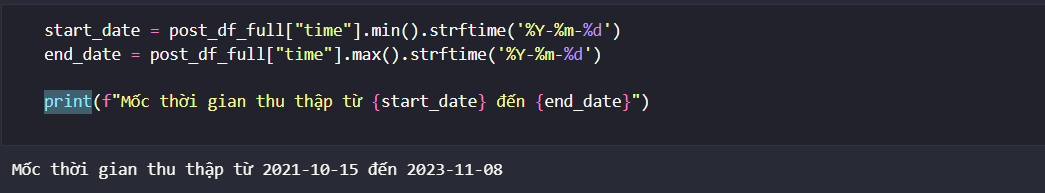
-Dữ liệu đã được tách riêng:



### 2.Thống kê và trực quan hóa

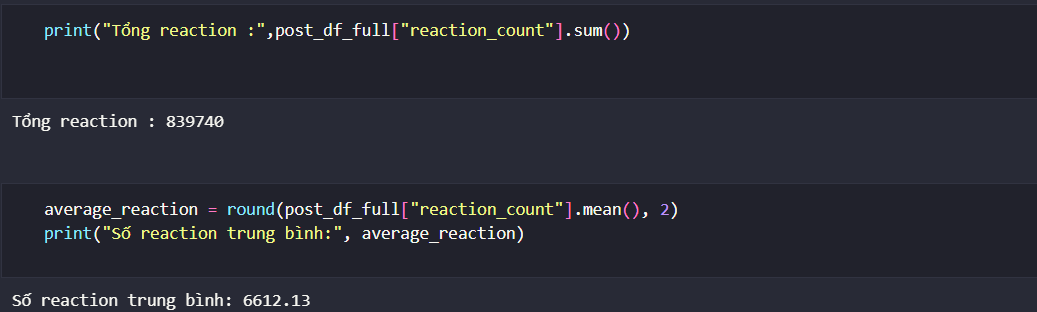
#### 2.1, Số lần tương tác, cái nhìn khái quát

##### 2.1.1, Thời gian thu thập

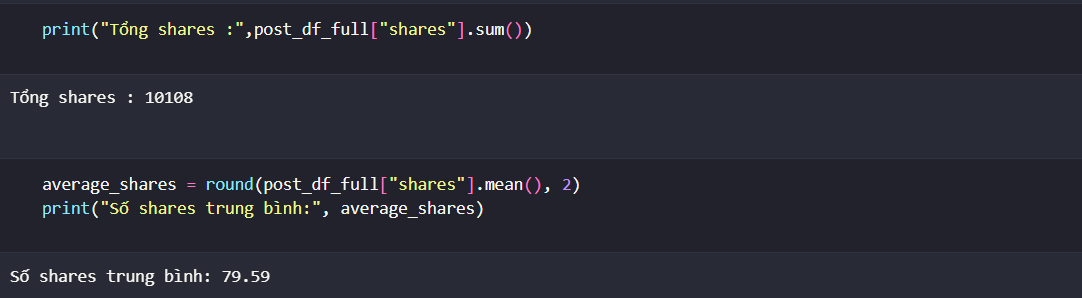


##### 2.1.2, Một số dữ liệu đáng chú ý

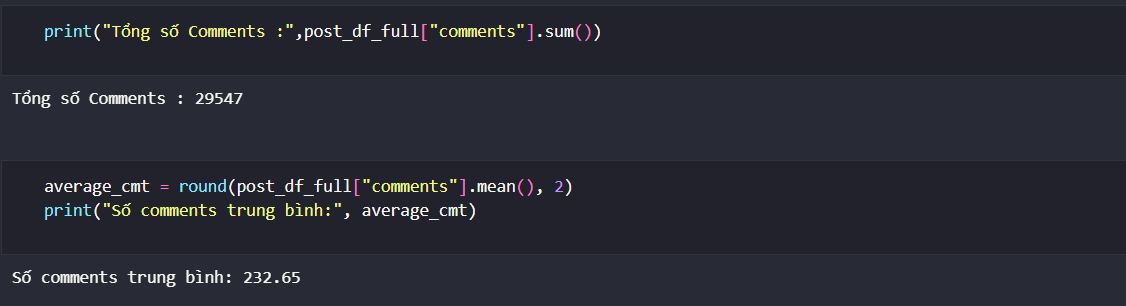
-Về lượt reactions:



-Về lượt shares:

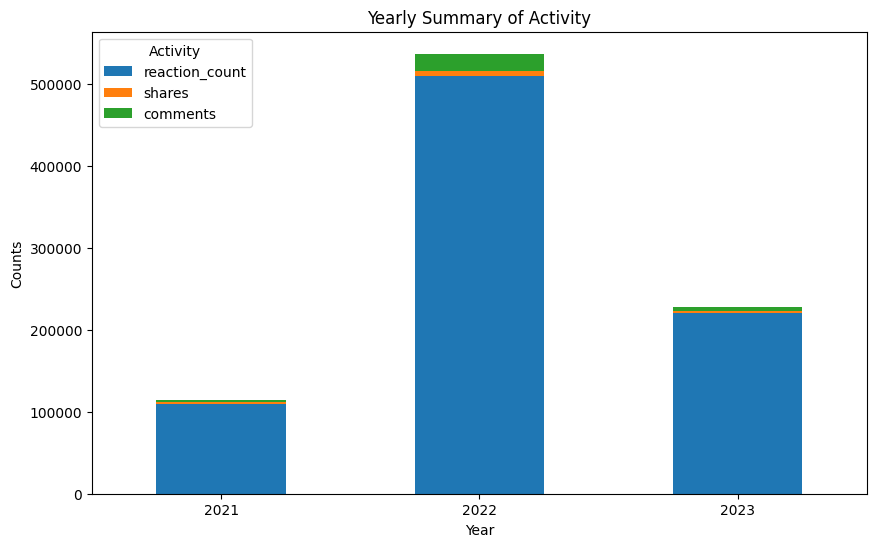


-Về lượt comments:



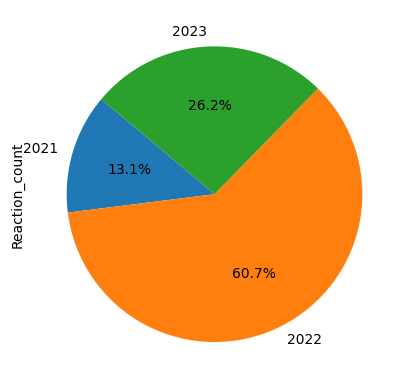
##### 2.1.3, Cái nhìn khái quát

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

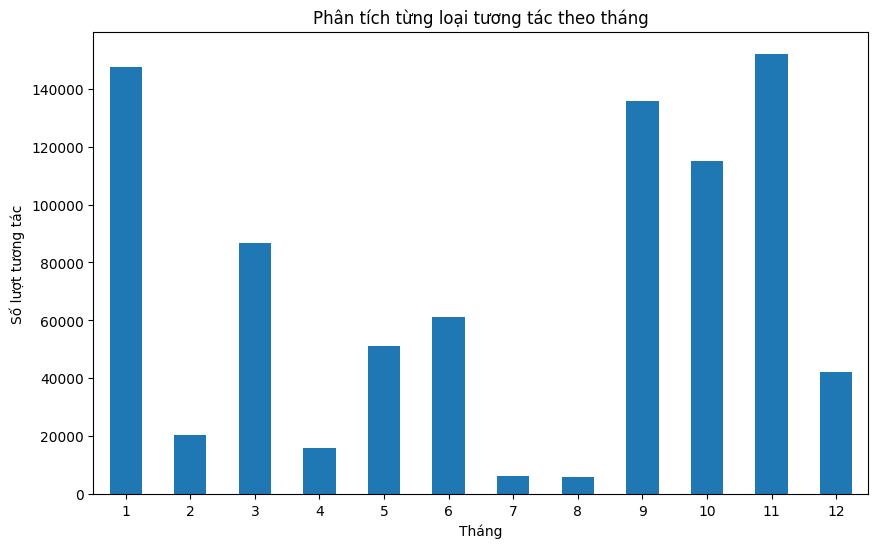


#### 2.2, Reactions hàng ngày

-Số lượng tương tác được tính theo năm :

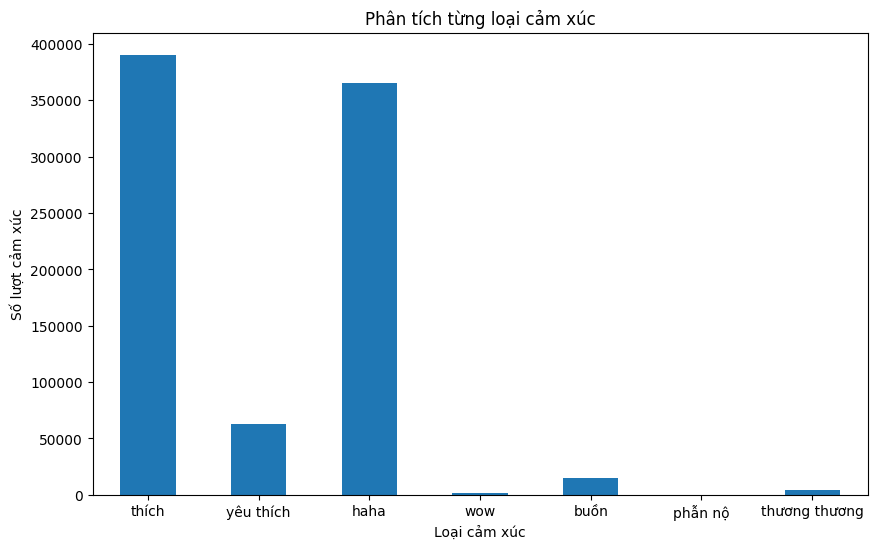


-Số lượng tương tác được tính theo tháng:



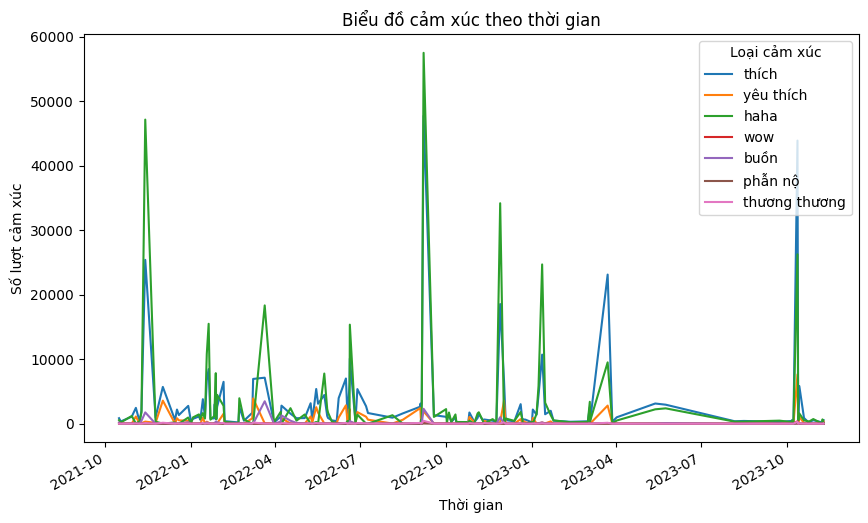
#### 2.3, Bức tranh tổng quan

-Biểu đồ thống kê cảm xúc:



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

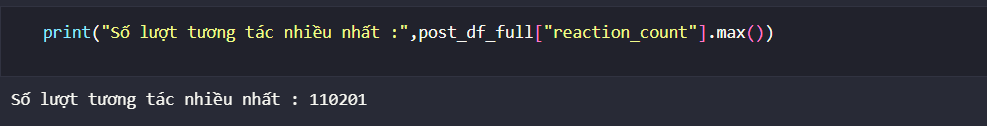
-Biểu đồ cảm xúc theo thời gian:

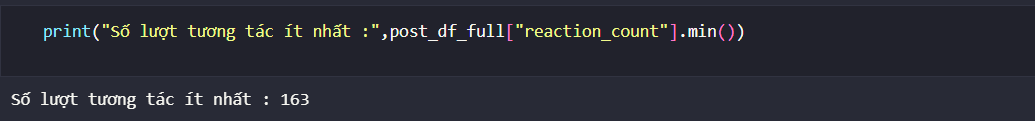


#### 2.4, Max,min của likes,shares và comments

##### 2.4.1, Reactions

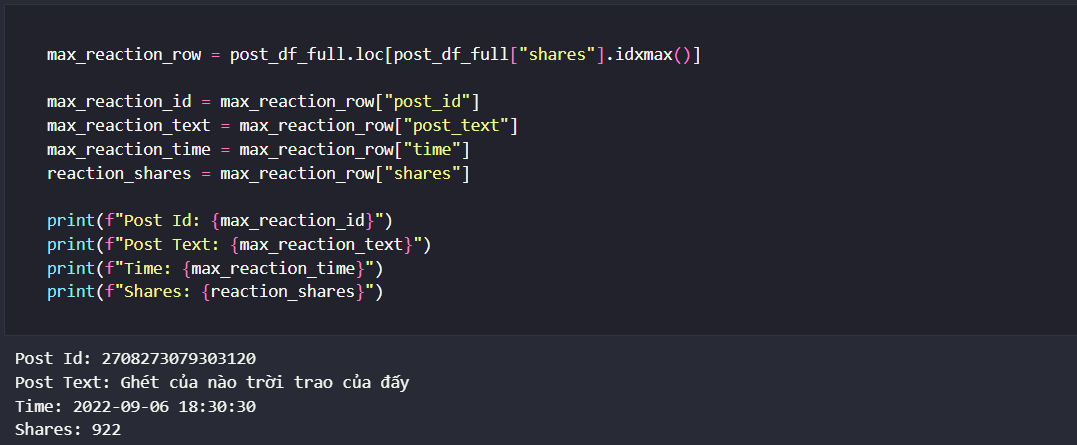
-Chênh lệch rất lớn giữa số bài viết có tương tác nhiều nhất, và bài viết nhận được ít sự quan tâm nhất:



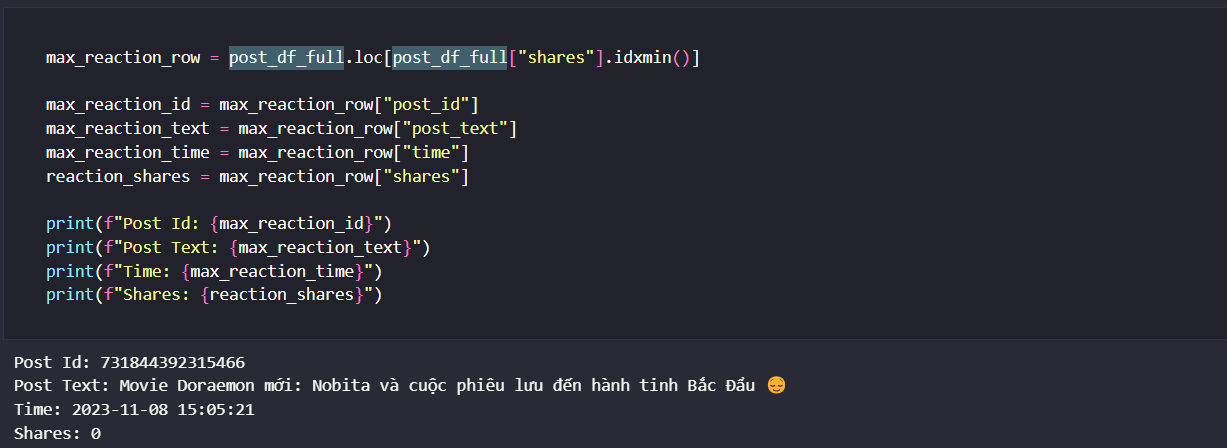


##### 2.4.2, Shares

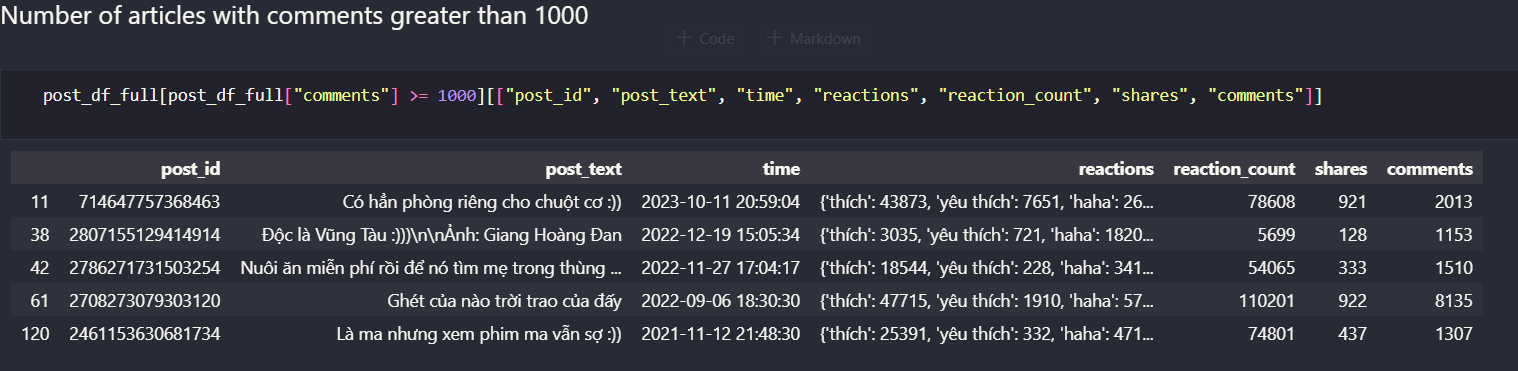
-Bài đăng có nhiều lượt shares nhất:



-Bài đăng có ít lượt shares nhất:

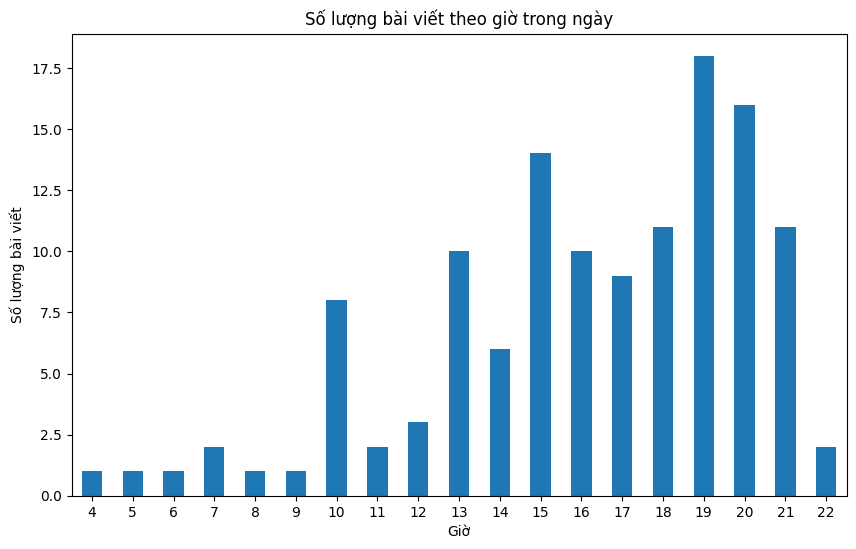


##### 2.4.3, Comments

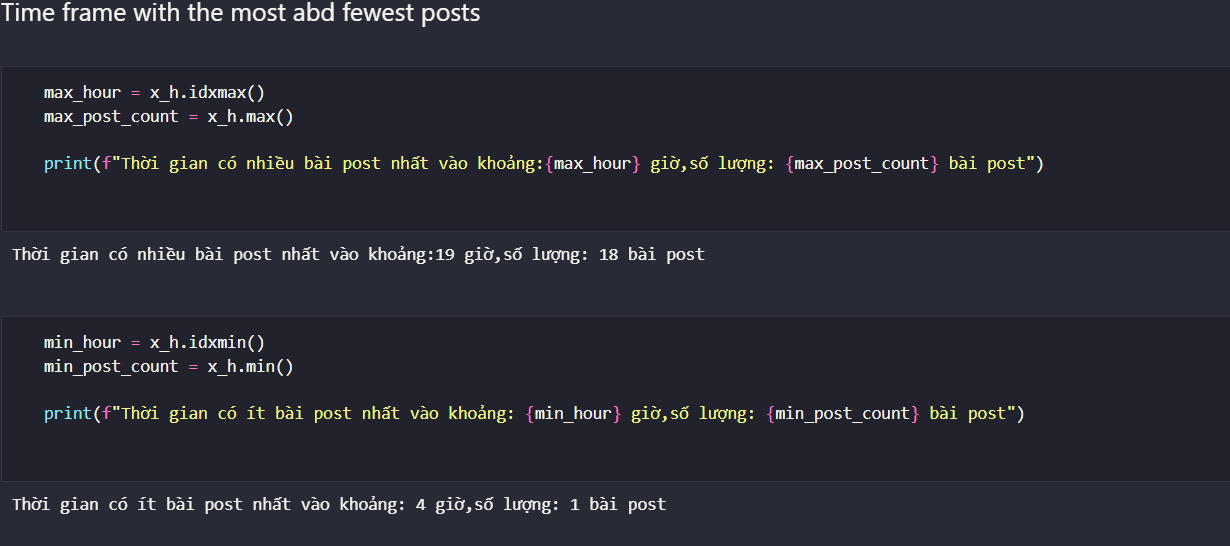


-Những bài viết nhận được nhiều lượt reactions , thì cũng nhận được nhiều phản hồi bằng comments cũng như là lượt chia sẻ rất cao

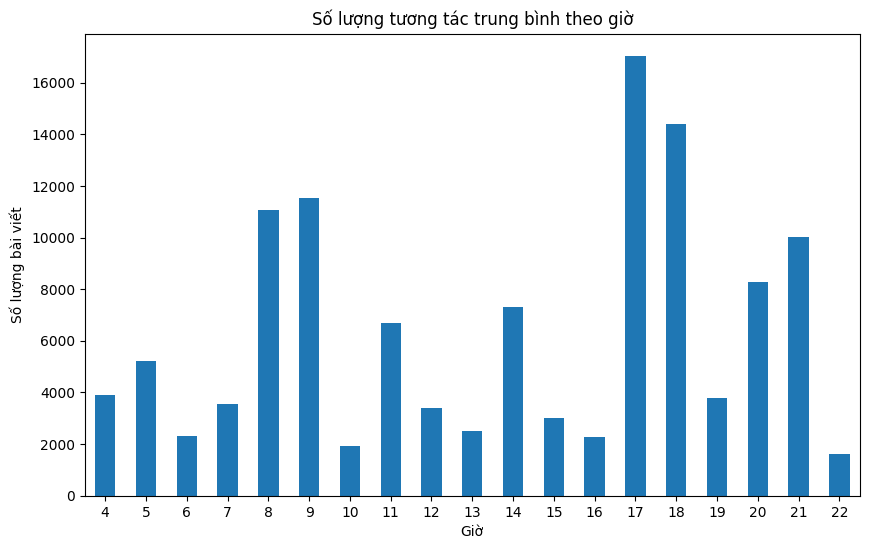
#### 2.5, Regular posting timestamps



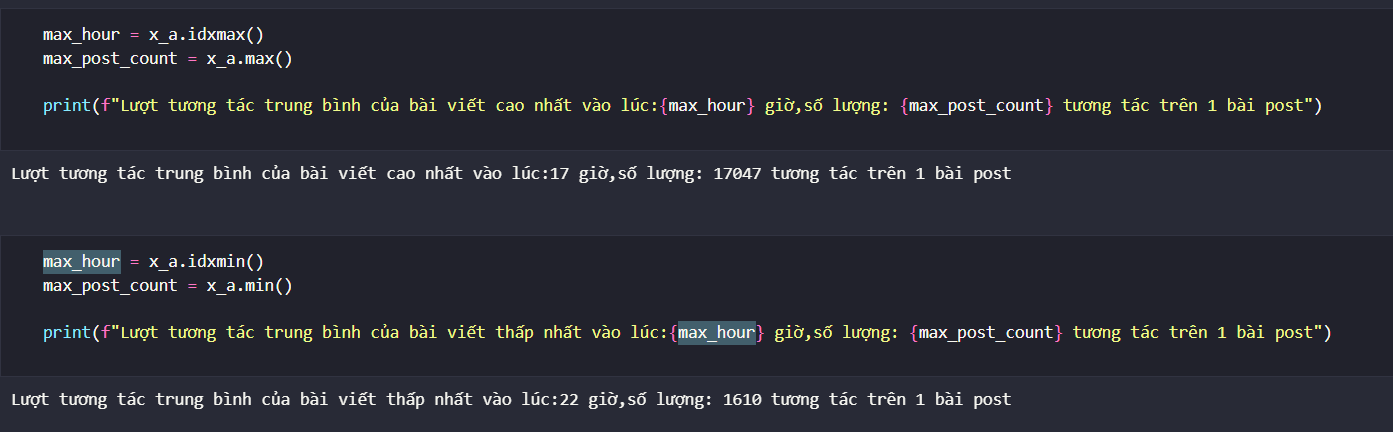
-Những giờ nghỉ ngơi, số bài viết được đăng rất nhiều



#### 2.6, Tương tác theo giờ

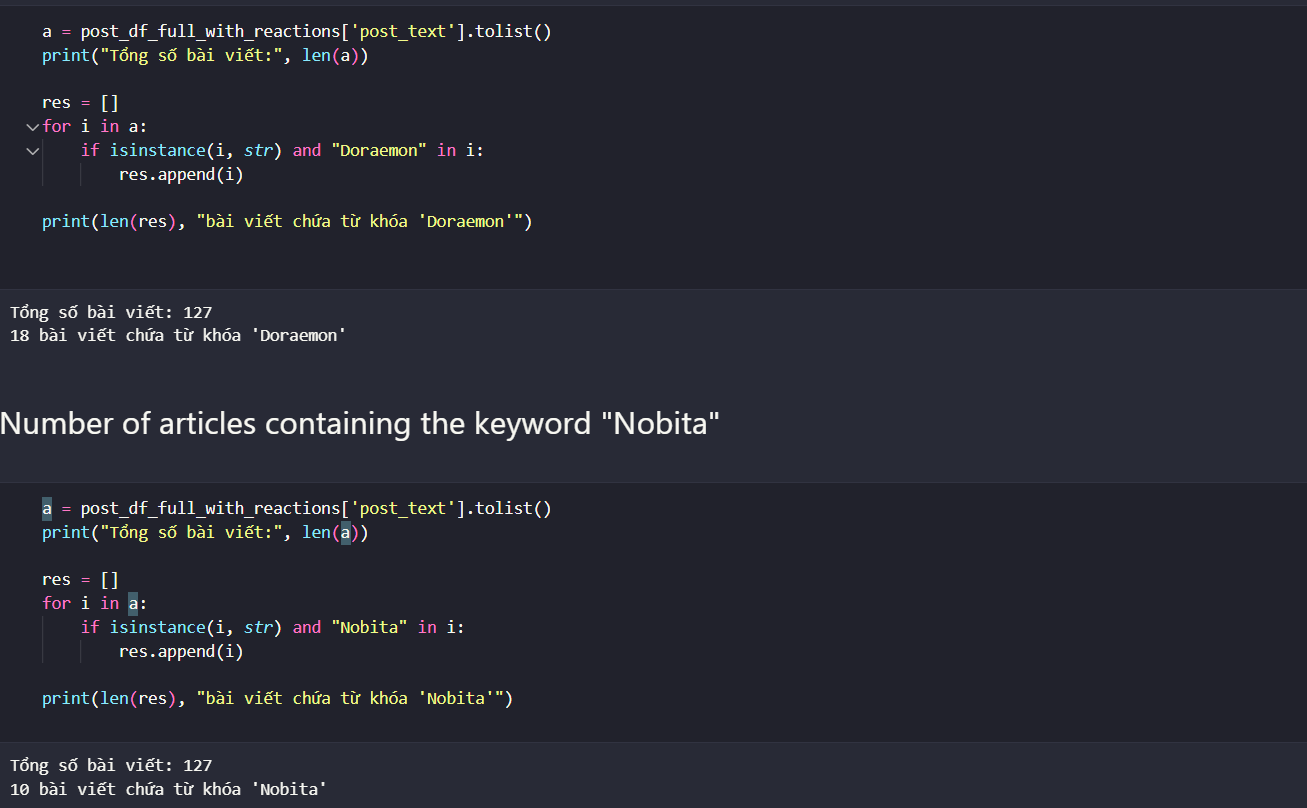


-Lượt tương tác rải khá đều vào những giờ hành chính

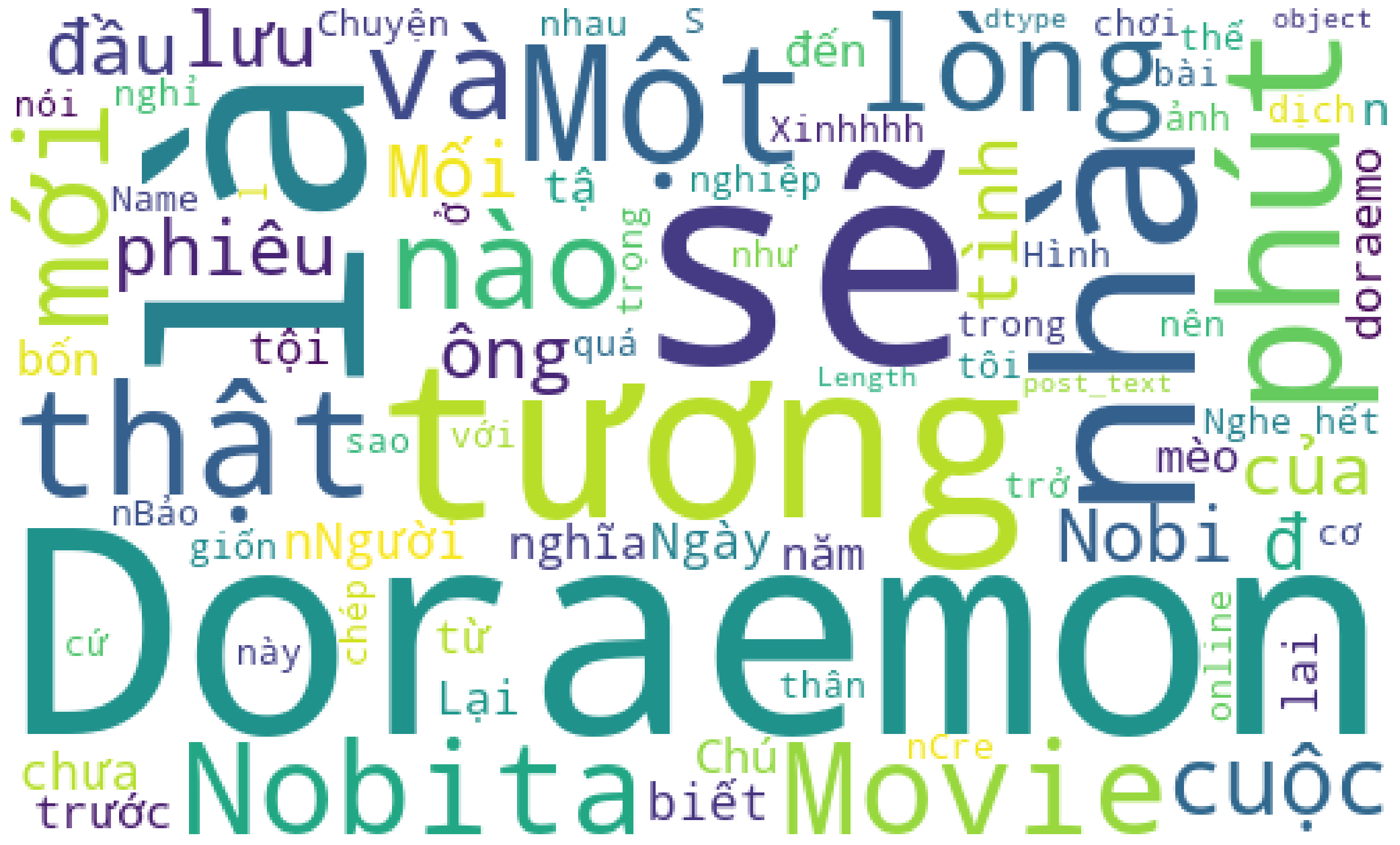


### 3.Phân tích từ khóa trong bài viết

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

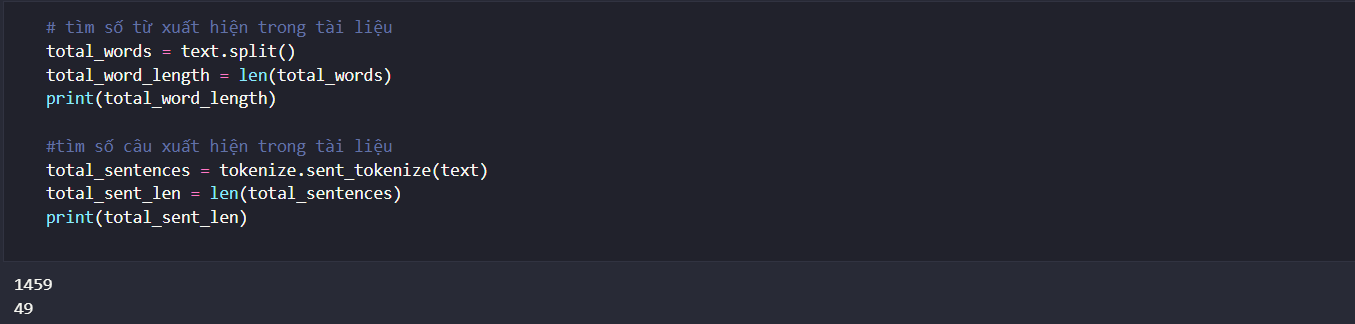


#### 3.2, Tần suất các từ được sử dụng nhiều nhất

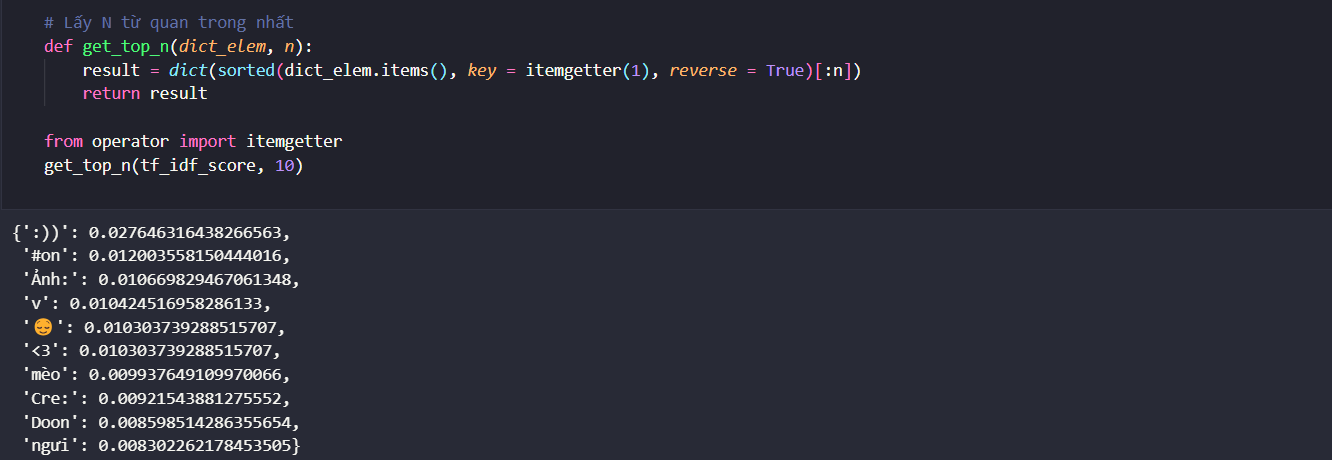


#### 3.3,Tìm keyword

-Số từ và câu xuất hiện trong các bài post của page:

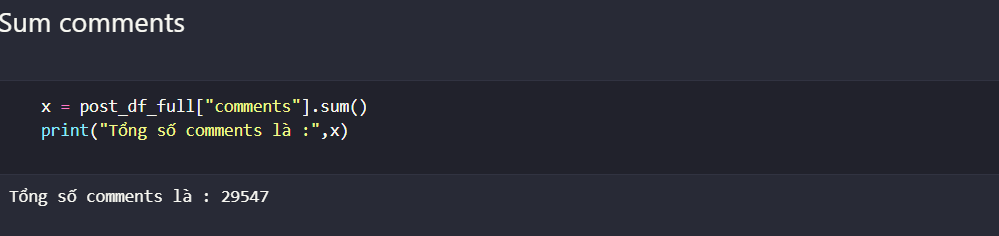


-Sử dụng file stop\_word và file stop\_word1 được lấy từ stopword Việt nam để tìm ra các keyword của bài viết:

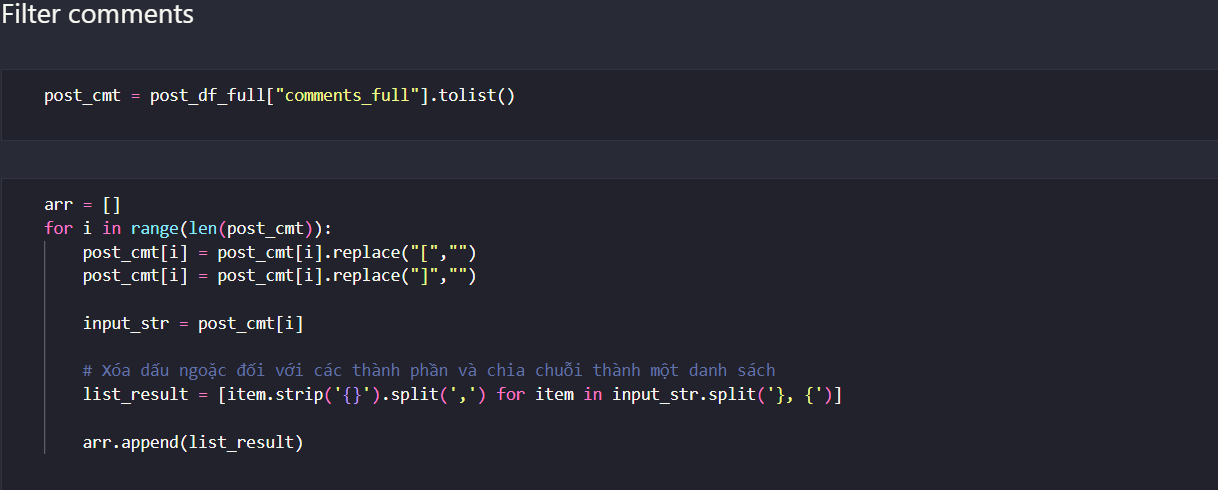


## Phần 3 : Phân tích, đánh giá comments

### 1.Tổng số comments

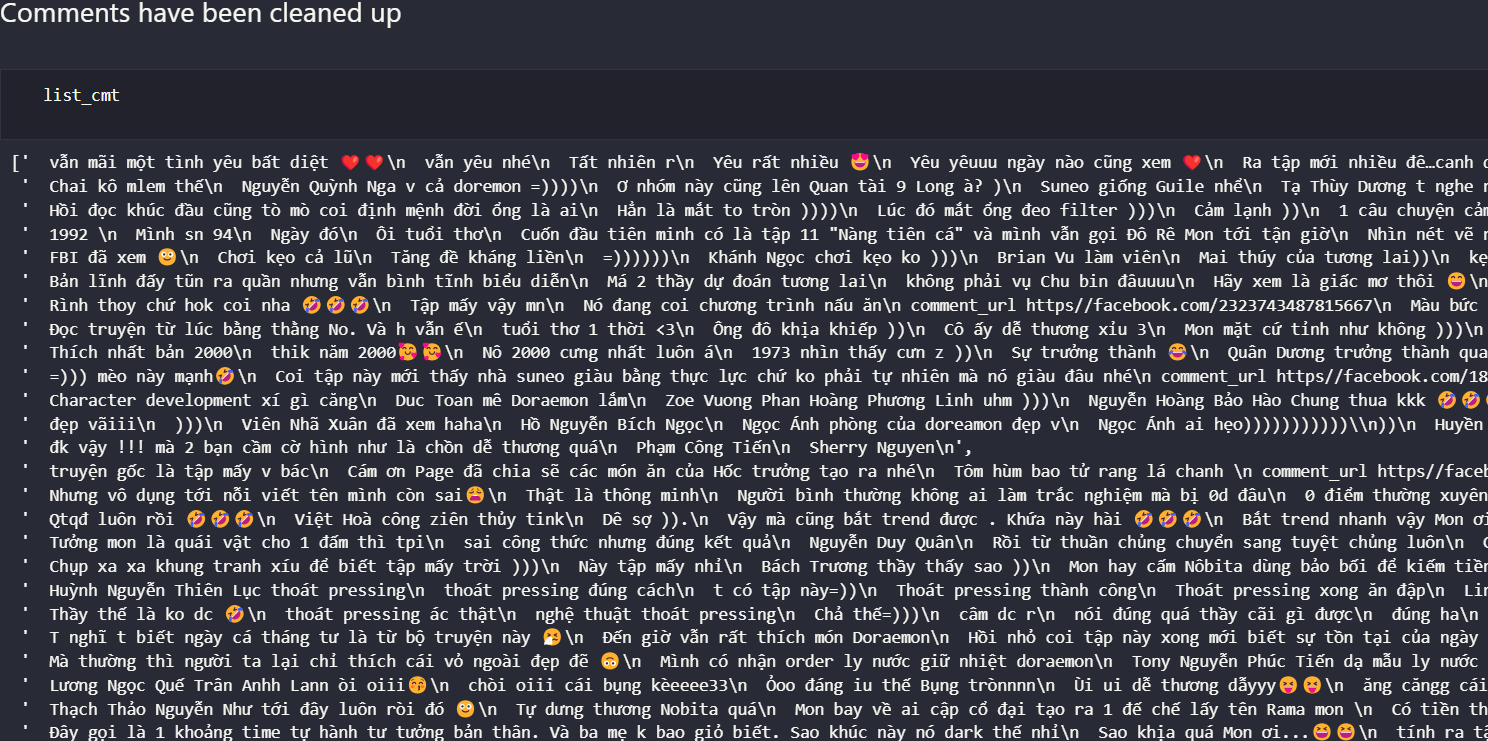


### 2.Lọc comments





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



### 3.Wordclound comments

-Tần suất xuất hiện của những từ trong comments:



3.4,Phân tích độ tích cực, tiêu cực của các comments

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

