

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 1**

-----🙞🙜🕮🙞🙜-----



**BÁO CÁO ĐỀ TÀI**

**“CÀO DỮ LIỆU PAGE FACEBOOK ĐỂ THEO DÕI TÌNH TRẠNG PHÁT TRIỂN CỦA PAGE VÀ XU HƯỚNG TRONG THỜI GIAN GẦN ĐÂY”**

**Giảng viên hướng dẫn**: **VŨ HOÀI NAM**

**Sinh viên: Nguyễn Quang Huy**

**Đặt vấn đề**

Trong thực tế hiện nay việc cào dữ liệu Facebook là một việc vô cùng cần thiết và có ứng dụng rất rộng rãi . Nhiều công ty về truyền thông ở Việt Nam đã làm nhiều năm nay đã thực hiện cào dữ liệu để thu thập thông tin và đo mức độ ảnh hưởng cho các chương trình marketing / quảng cáo. Một số công ty khác thì bán giải pháp trích xuất thông tin từ Facebook với đa dạng các nội dung có thể lấy được. Hoặc đơn giản nhất với sinh viên hoặc một số shop bán hàng nhỏ họ thực hiện cào dữ liệu page facebook để theo dõi tình trạng phát triển của page, xu hướng phát trong thời gian gần đây. Trong một số cuộc thi về truyền thông cần tìm tất cả các bài viết Facebook để tìm xem bài viết nào nhiều tương tác nhất để giao giải thưởng , hoặc trong 1 bài viết / livestream ta cần cào tất cả các bình luận để xem xem ai là người bình luận sớm nhất và đúng nhất ,….Việc cào dữ liệu FaceBook là một ý tưởng rất hay và cần thiết để nghiên cứu . Chính vì vậy trong đề tài 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 dược hoàn thiện hơn .

Em xin chân thành cảm ơn thầy!

Hà Nội, Tháng 9 năm 2022

Huy

Nguyễn Quang Huy

Contents

[Phần 1 : Cào dữ liệu Facebook về và lưu dữ liệu: 4](#_Toc114038499)

[I, yêu cầu về cài đặt 4](#_Toc114038500)

[1.Thư viện dùng 4](#_Toc114038501)

[2. File cookies.txt 4](#_Toc114038502)

[II, Cào dữ liệu FB về và xem thông tin dữ liệu cào được: 5](#_Toc114038503)

[1.Cào dữ liệu về: 5](#_Toc114038504)

[2.Xem thông tin dữ liệu cào được: 7](#_Toc114038505)

[3. Lưu dữ liệu về 10](#_Toc114038506)

[Phần 2: Phân tích tình trạng phát triển của page : 12](#_Toc114038507)

[I, Phân tích sự phát triển lượt react của từng post: 12](#_Toc114038508)

[1.Bảng chung 12](#_Toc114038509)

[2.Tính tổng reaction : 13](#_Toc114038510)

[3.Phân tích 13](#_Toc114038511)

[II, Phân tích từng loại react thống kê theo từng ngày: 14](#_Toc114038512)

[1.Xem lại dữ liệu thống kê: 14](#_Toc114038513)

[2.Tinh chỉnh lại dữ liệu : 15](#_Toc114038514)

[3. biểu đồ thống kê từng loại cảm xúc theo từng ngày 15](#_Toc114038515)

[4.Bức tranh tổng hợp : 20](#_Toc114038516)

[5.Sự chênh lệch nhau tương đối của các loại cảm xúc với số lượt share 20](#_Toc114038517)

[Phần 3 : Phân tích bài viết của page và comment của người dùng 21](#_Toc114038518)

[I, Phân tích bài viết của page 21](#_Toc114038519)

[1. Tìm các bài viết chứa 1 từ khóa nào đó 21](#_Toc114038520)

[2. Tần suất các từ được sử dụng nhiều nhất qua thống kê: 21](#_Toc114038521)

[3. Sử dụng thuật toán để dò ra keyword 21](#_Toc114038522)

[II, Phân tích comment của người dùng 22](#_Toc114038523)

[1.lấy thông tin các comment\_full của tất cả 100 bài đăng 22](#_Toc114038524)

[2. Tần suất các từ được sử dụng nhiều nhất qua thống kê: 23](#_Toc114038525)

[3. Sử dụng thuật toán để dò ra keyword 24](#_Toc114038526)

# Phần 1 : Cào dữ liệu Facebook về và lưu dữ liệu:

## I, yêu cầu về cài đặt

### 1.Thư viện dùng

Thư viện cho phép cào dữ liệu thông tin các bài viết và comment trong các bài viết đó từ 1 trang trên Facebook để người dùng có thể sử dụng phân tích.

%pip install facebook\_scraper

Text

Description automatically generated

### 2. File cookies.txt

Bước 1 : Thêm tiện ích get cookies:

<https://chrome.google.com/webstore/detail/get-cookiestxt/bgaddhkoddajcdgocldbbfleckgcbcid>

Graphical user interface, text, application, Teams

Description automatically generated

Bước 2 : mở facebook và download cookies về :

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Bước 3 : Rồi upload file lên drive

Graphical user interface, text, application

Description automatically generated

## II, Cào dữ liệu FB về và xem thông tin dữ liệu cào được:

### 1.Cào dữ liệu về:

Page dùng để cào post

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Thực hiện cào dữ liệu cuộn trên 10 trang đầu của page:

Graphical user interface, text, application

Description automatically generated

Kết quả được 100 post :

Graphical user interface, text, application

Description automatically generated

### 2.Xem thông tin dữ liệu cào được:

#### Thông tin của 1 post :

Text

Description automatically generated

Còn nhiều text phía sau :

#### Các trường thông tin của 1 post:

Text

Description automatically generated

#### Thông tin các comment trong 1 post :

Graphical user interface, text, application

Description automatically generated

#### Thống kế chung về tất cả các post :

Table

Description automatically generated

#### Hiển thị sơ qua 10 post đầu tiên:

Graphical user interface

Description automatically generated with medium confidence

### 3. Lưu dữ liệu về

#### Lưu dạng excel :

Chuyển từng post vừa có từ list thành 1 dataframe:

Graphical user interface, text, application, email

Description automatically generated

Lưu vào file xlsx:

Graphical user interface, text, application

Description automatically generated

#### Lưu dạng npy ( để giữu nguyên thuộc tính trong list)

Graphical user interface, text, application

Description automatically generated

Kết quả cuối cùng ta được 1 file npy và 1 file xlsx:

Graphical user interface, application

Description automatically generated

# Phần 2: Phân tích tình trạng phát triển của page :

## I, Phân tích sự phát triển lượt react của từng post:

### 1.Bảng chung

Table

Description automatically generated with low confidence

### 2.Tính tổng reaction :

Table

Description automatically generated

### 3.Phân tích

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

## II, Phân tích từng loại react thống kê theo từng ngày:

### 1.Xem lại dữ liệu thống kê:

Application

Description automatically generated with low confidence

### 2.Tinh chỉnh lại dữ liệu :

Graphical user interface, application, Word

Description automatically generated

A picture containing table

Description automatically generated

### 3. biểu đồ thống kê từng loại cảm xúc theo từng ngày

#### Thích

Chart

Description automatically generated

#### Yêu thích

Chart, histogram

Description automatically generated

#### Haha

A picture containing chart

Description automatically generated

#### Wow

Chart, histogram

Description automatically generated

#### Buồn

Chart, histogram

Description automatically generated

#### Phẫn nộ:

Chart

Description automatically generated with medium confidence

#### Thương thương

Chart, histogram

Description automatically generated

#### Số lượt share:

Chart, histogram

Description automatically generated

#### Số lượt comment:

Chart, histogram

Description automatically generated

### 4.Bức tranh tổng hợp :

Chart, histogram

Description automatically generated

### 5.Sự chênh lệch nhau tương đối của các loại cảm xúc với số lượt share

Graphical user interface, text, application, Word

Description automatically generated

# Phần 3 : Phân tích bài viết của page và comment của người dùng

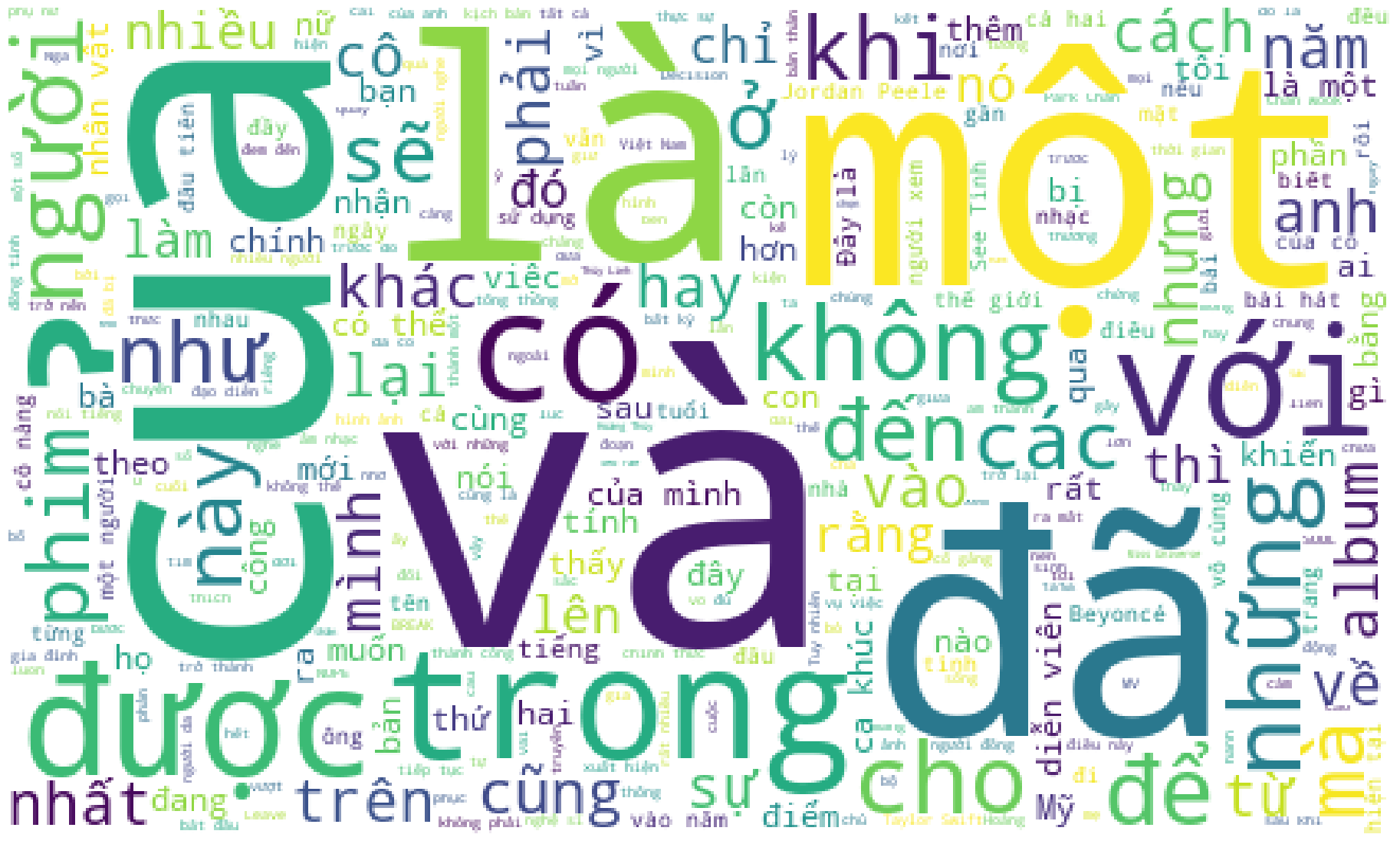
## I, Phân tích bài viết của page

### 1. Tìm các bài viết chứa 1 từ khóa nào đó

Application

Description automatically generated with medium confidence

### 2. Tần suất các từ được sử dụng nhiều nhất qua thống kê:



### 3. Sử dụng thuật toán để dò ra keyword

Graphical user interface, text, application, email

Description automatically generated

## II, Phân tích comment của người dùng

### 1.lấy thông tin các comment\_full của tất cả 100 bài đăng

Graphical user interface, text, application

Description automatically generated

Thông tin tất cả các comment đưa vào 1 list

Text

Description automatically generated

Tổng số từ và tổng số câu sau khi đã chuẩn hóa stopword:

Text

Description automatically generated

### 2. Tần suất các từ được sử dụng nhiều nhất qua thống kê:

Text

Description automatically generated

A close-up of a dollar bill

Description automatically generated with low confidence

### 3. Sử dụng thuật toán để dò ra keyword

Ta tạo 2 mảng stopword trước khi thực hiện tìm keyword:

A picture containing scatter chart

Description automatically generated

Text

Description automatically generated

* Kết quả:

Graphical user interface, text

Description automatically generated