

BÁO CÁO LAB1

TÍNH TOÁN ĐA PHƯƠNG TIỆN

DỮ LIỆU DẠNG TEXT - CRAWLER

Giảng viên hướng dẫn: Đỗ Văn Tiến

NHÓM

Trần Duy Tân

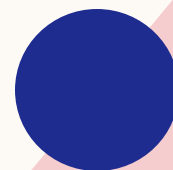
21522576

Nguyễn Duy Thái

21522581

Lê Nhật Minh

21522339



I. NỘI DUNG

- Viết crawler thu thập bài báo khoa học của một tên tác giả bất kỳ trên 1 thư viện số như ACM, IEEE, Googleschcoolar, Spinger.
- Viết crawler thu thập các bài viết và bình luận từ 1 trang facebook ví dụ UIT Fanpage, UIT confession.
- Viết crawler thu thập các bài viết và bình luận từ 1 trang báo điện tử.
- Viết crawler thu thập ảnh là kết quả trả về từ google khi sử dụng một từ khóa.



II. THƯ VIỆN SỬ DỤNG

1. SELENIUM

- Selenium là một thư viện được sử dụng để kiểm thử tự động các ứng dụng web. Nó cung cấp một giao diện lập trình ứng dụng (API) để điều khiển các trình duyệt web, cho phép người dùng kiểm thử các chức năng và tương tác với các trang web một cách tự động.
- Ưu điểm:
 - ☐ Kiểm thử tự động các ứng dụng web.
 - ☐ Tự động hóa các tác vụ trên trình duyệt web.
 - ☐ Cung cấp API cho nhiều ngôn ngữ lập trình, bao gồm Python.
 - ☐ Có thể tạo các kịch bản trình duyệt web.
 - ☐ Giúp tiết kiệm thời gian và công sức so với kiểm thử thủ công.

2. BEAUTIFULSOUP

- BeautifulSoup là một thư viện mã nguồn mở của Python được sử dụng để phân tích cú pháp HTML và XML. Nó cho phép người dùng dễ dàng truy xuất và phân tích các thông tin trong các tài liệu HTML hoặc XML, từ đó lấy được thông tin cần thiết một cách dễ dàng.
- Với BeautifulSoup, người dùng có thể truy xuất các phần tử của HTML hoặc XML, lấy nội dung của các thẻ, tìm kiếm các phần tử theo tên hoặc lớp, thậm chí là tìm kiếm các phần tử theo nội dung của chúng.

3. REQUESTS

- Requests là một thư viện HTTP của Python, được sử dụng để thực hiện các yêu cầu HTTP từ các ứng dụng Python. Nó cho phép người dùng tạo và gửi các yêu cầu HTTP đến các máy chủ web và nhận lại các phản hồi từ chúng.
- Requests được thiết kế để có tính dễ sử dụng cao, với các phương thức đơn giản và gần giống với các yêu cầu HTTP cơ bản. Nó cũng cho phép người dùng tùy chỉnh các yêu cầu và phản hồi HTTP bằng cách sử dụng các tham số và tùy chọn khác nhau.


3. PYODBC


- Pyodbc là một module python mã nguồn mở giúp truy cập sở dữ liệu ODBC- giao diện lập trình ứng dụng tiêu chuẩn truy cập hệ thống quản lý cơ sở dữ liệu.
- Cách cài đặt pyodbc: `pip install pyodbc`



CRAWLER THU THẬP BÀI BÁO KHOA HỌC

TRUY CẬP VÀO GOOGLE SCHOLAR ĐỂ TÌM PROFILE TÁC GIẢ





Ngoc Hoang Luong

[University of Information Technology \(UIT-HCM\)](#)
Verified email at uit.edu.vn

[Evolutionary Computation](#) [Machine Learning](#) [Artificial Intelligence](#) [Optimization](#)

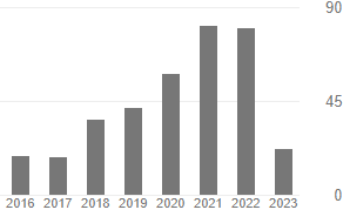
[FOLLOW](#)

[GET MY OWN PROFILE](#)

TITLE	CITED BY	YEAR
Expanding from discrete cartesian to permutation gene-pool optimal mixing evolutionary algorithms PAN Bosman, NH Luong, D Thierens Proceedings of the Genetic and Evolutionary Computation Conference 2016, 637-644	33	2016
Application and benchmarking of multi-objective evolutionary algorithms on high-dose-rate brachytherapy planning for prostate cancer treatment NH Luong, T Alderliesten, A Bel, Y Niatsetski, PAN Bosman Swarm and Evolutionary Computation	32	2017
Evaluation of bi-objective treatment planning for high-dose-rate prostate brachytherapy—A retrospective observer study SC Maree, NH Luong, ES Kooreman, N van Wieringen, A Bel, KA Hinnen, ... Brachytherapy 18 (3), 396-403	28	2019
Multi-objective gene-pool optimal mixing evolutionary algorithms NH Luong, H La Poutré, PAN Bosman Proceedings of the 2014 Annual Conference on Genetic and Evolutionary ...	28	2014
Elitist archiving for multi-objective evolutionary algorithms: To adapt or not to adapt HN Luong, PAN Bosman Parallel Problem Solving from Nature-PPSN XII: 12th International Conference ...	28	2012
Multi-objective optimization techniques and applications in electric power systems	27	2012

Cited by [VIEW ALL](#)

	All	Since 2018
Citations	404	319
h-index	12	11
i10-index	15	11



Public access [VIEW ALL](#)

Public access	Public access
1 article	7 articles
not available	available

Based on funding mandates

**DUYỆT QUA TỪNG BÀI BÁO
VÀ LẤY CÁC THÔNG TIN
TITLE, AUTHOR,
CITED BY VÀ YEAR**

TỔNG HỢP CÁC THÔNG TIN VÀO FILE .CSV

	C1 ÷ C2	÷ C3	÷ C4	÷ C5 ÷
1	ID	Title	Authors	Cited by Year
2	0	Expanding from discrete cartesian to permutation gene-pool optimal...	PAN Bosman, NH Luong, D Thierens	33 2016
3	1	Application and benchmarking of multi-objective evolutionary algor...	NH Luong, T Alderliesten, A Bel, Y Niatsetski, PAN Bosman	32 2017
4	2	Evaluation of bi-objective treatment planning for high-dose-rate p...	SC Maree, NH Luong, ES Kooreman, N van Wieringen, A Bel, KA Hinnen...	28 2019
5	3	Multi-objective gene-pool optimal mixing evolutionary algorithms	NH Luong, H La Poutré, PAN Bosman	28 2014
6	4	Elitist archiving for multi-objective evolutionary algorithms: To ...	HN Luong, PAN Bosman	28 2012
7	5	Multi-objective optimization techniques and applications in electr...	MOW Grond, NH Luong, J Morren, JG Sloatweg	27 2012
8	6	Fine-tuning bert for sentiment analysis of vietnamese reviews	QT Nguyen, TL Nguyen, NH Luong, QH Ngo	26 2020
9	7	Multi-objective gene-pool optimal mixing evolutionary algorithm wi...	NH Luong, H La Poutré, PAN Bosman	24 2018
10	8	The multi-objective real-valued gene-pool optimal mixing evolution...	A Bouter, NH Luong, C Witteveen, T Alderliesten, PAN Bosman	23 2017
11	9	Practice-oriented optimization of distribution network planning us...	MOW Grond, HN Luong, J Morren, PAN Bosman, HJG Sloatweg, ...	20 2014
12	10	Entropy-based efficiency enhancement techniques for evolutionary a...	HN Luong, HTT Nguyen, CW Ahn	19 2012

KHÓ KHĂN

Không biết trước số lần cần ấn nút “Show more” để load hết tất cả bài báo.

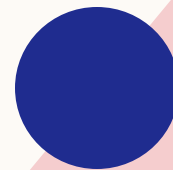
Articles 1–20  SHOW MORE

Articles 1–47  SHOW MORE

GIẢI PHÁP

Sử dụng vòng lặp while, bắt Exception trả về khi nút “Show more” bị disabled.

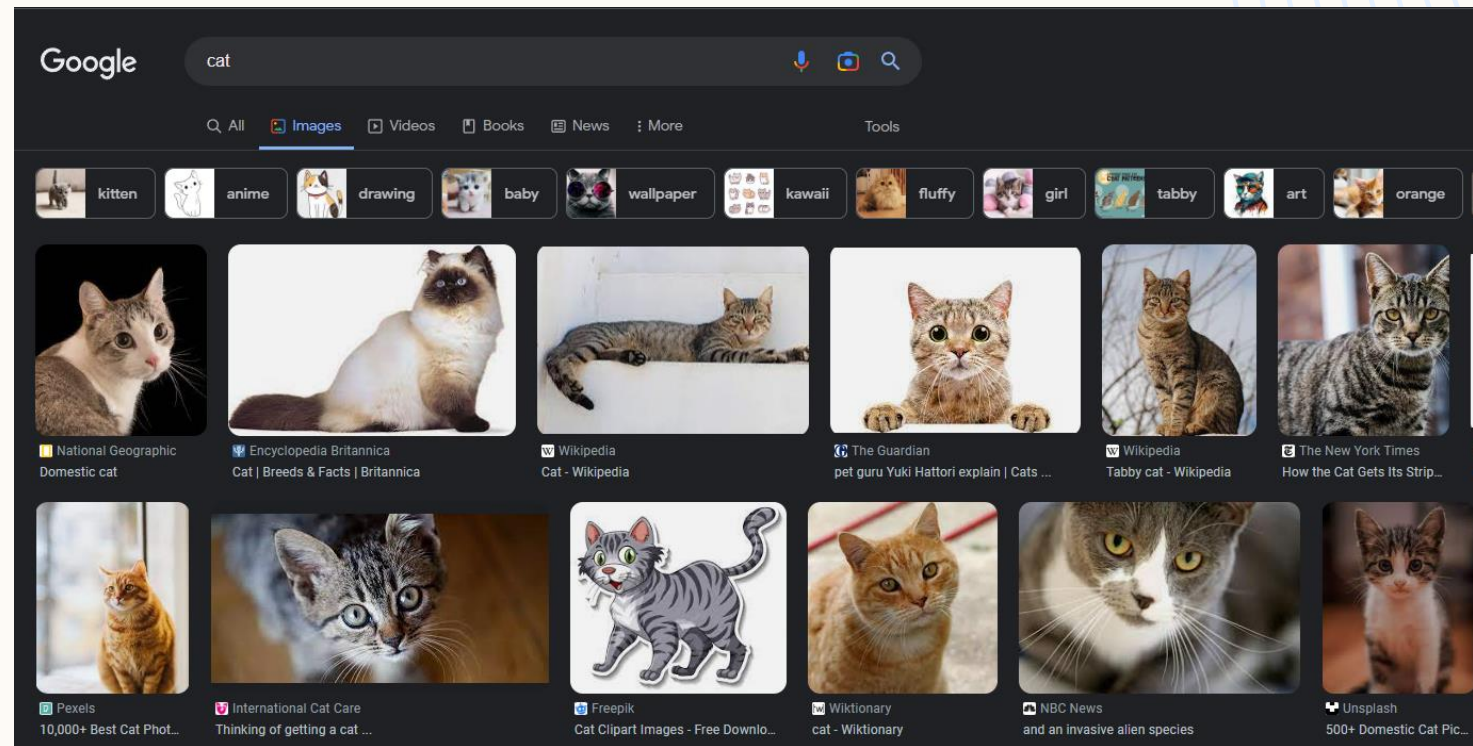
CRAWLER THU THẬP ẢNH TỪ GOOGLE IMAGES



NHẬP VÀO TỪ KHÓA CẦN TÌM KIẾM VÀ THỰC HIỆN QUERY TRÊN WEBSITE

```
C:\Users\PC\anaconda3\envs\Crawler\python.exe C:\Users\PC\PycharmProjects\Crawler\image.py  
Please type the keyword...cat
```

DUYỆT QUA TỪNG ẢNH VÀ DOWNLOAD VỀ



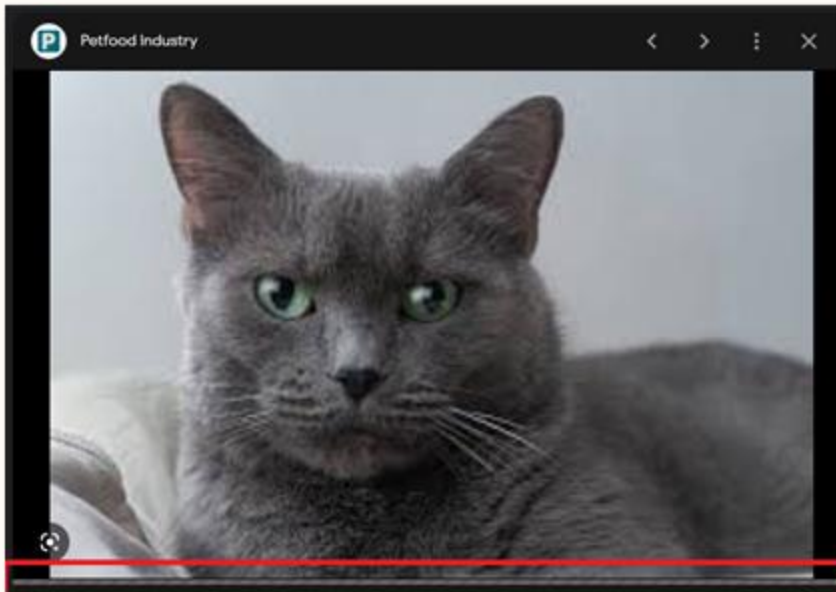
Downloaded element 1 out of 49 total. URL: <https://i.natgeofe.com/n/548467d8-c5f1-4>
Downloaded element 2 out of 49 total. URL: <https://cdn.britannica.com/39/7139-050-A>
Couldn't download an image 3, continuing downloading the next one
Downloaded element 4 out of 49 total. URL: <https://i.quim.co.uk/img/media/26392d053>
Couldn't download an image 5, continuing downloading the next one
Downloaded element 6 out of 49 total. URL: <https://static01.nyt.com/images/2021/09/>
Downloaded element 7 out of 49 total. URL: <https://res.cloudinary.com/dk-find-out/i>
Timeout! Will download a lower resolution image and move onto the next one
Couldn't download an image 8, continuing downloading the next one
Downloaded element 9 out of 49 total. URL: <https://images.pexels.com/photos/1170986>

DUYỆT QUA TỪNG ẢNH VÀ DOWNLOAD VỀ



KHÓ KHĂN

Một số ảnh có độ phân giải cao, thời gian để load ảnh lâu.



GIẢI PHÁP

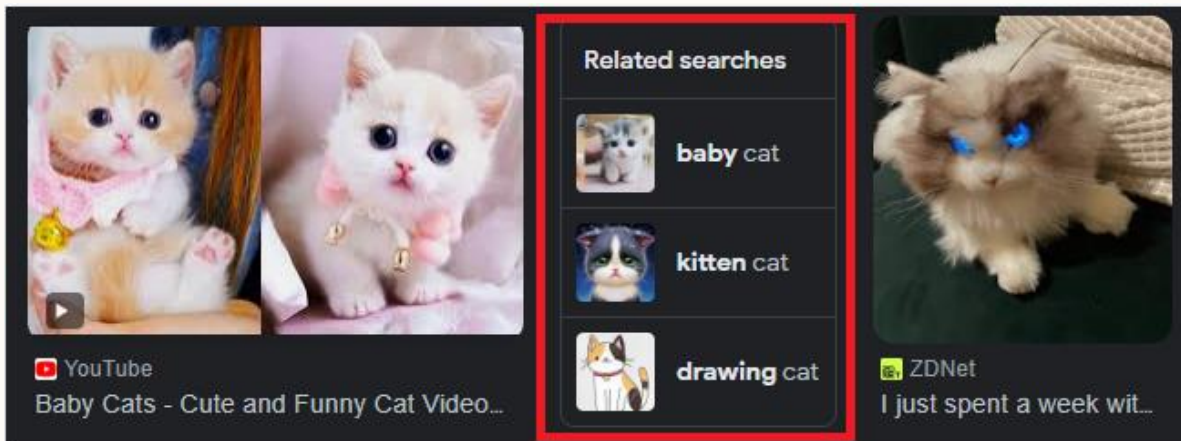
Sử dụng một bộ đếm thời gian, khi timeout sẽ bỏ qua ảnh và tải về phiên bản có độ phân giải thấp hơn.

KHÓ KHĂN

Gặp những thẻ gợi ý nội dung liên quan của Google. Những thẻ này không thể lấy về hình ảnh.

GIẢI PHÁP

Đọc attribute của thẻ và bỏ qua những thẻ này.

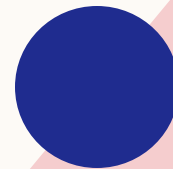


CRAWLER

THU THẬP BÀI VIẾT VÀ

BÌNH LUẬN TRÊN 1

TRANG BÁO ĐIỆN TỬ



TRUY CẬP VÀO VNEXPRESS.NET

→ ↻ 🏠 🔒 https://vnexpress.net/goc-nhin/binh-luan-nhieu
A^{aa} ☆ ⚙️ |

📖 **Mới nhất** | Thời sự **Góc nhìn** Thế giới Video Podcasts Kinh doanh Khoa học Giải trí Thể thao Pháp luật Giáo dục Sức khỏe Đời sống Du lịch Số hóa Xe Ý

Rút bảo hiểm một lần

Lúc mới ra trường, tôi gửi tiền vào một quỹ lớn, nhưng phải 35 năm nữa, tôi mới biết khoản đầu tư của mình hiệu quả ra sao.

Chính trị & chính sách 280

Nguyễn Khắc Giang

Đẩy thế khó cho dân

Đề xuất sở hữu chung cư có thời hạn chẳng khác nào 'đẩy cái khó cho dân'.

Kinh doanh & quản trị 228

Phan Tất Đức

Cứu bất động sản

Sự sụp đổ của ngành bất động sản, nếu không may xảy ra, sẽ khiến cả nền kinh tế phải gánh hậu quả.

Kinh doanh & quản trị 189

Huỳnh Thế Du

Mua bảo hiểm làm gì?

Bảo hiểm được bán kiểu 'bia kèm lạc' mà 'lạc' nhiều hơn 'bia' cả về số lượng lẫn giá trị kinh tế.

Kinh doanh & quản trị 167

Võ Nhật Vinh

**DUYỆT QUA TỪNG BÀI BÁO
VÀ LẤY CÁC THÔNG TIN
TITLE, CONTENT,
COMMENT**

TỔNG HỢP CÁC THÔNG TIN VÀO FILE .CSV

1	Title	Content	Comment
2	Con tôi tự kỷ	Hôm ấy, tôi và vợ vất vả công quay v...	Duy Khang↻Duy KhangMình gửi đến bạn sự c...
3	Mất tiền vì bảo hiểm	Sự việc chưa rõ đúng sai, Ngọc La...	David Tèo↻David TèoCách đây 2 tháng tôi ...
4	Bán mình giá rẻ	Người thợ đã nghỉ việc có trình đ...	kuteosj↻kuteosj"Thứ ít ỏi " sau 1 năm tr...
5	Rút bảo hiểm một lần	Tuy nhiên, quỹ lại không công bố ...	Manh Do↻Manh DoLương hưu đang tính lợi r...
6	Đẩy thế khó cho dân	Giải thích của cơ quan soạn thảo ...	Nguyễn Hữu Nghi↻Nguyễn Hữu NghiHoàn toàn...

KHÓ KHĂN

- Dùng BeautifulSoup không lấy được comment, có thể do bs4 không truy cập được các thông tin bị ẩn bởi javascript

GIẢI PHÁP

- Không xài BeautifulSoup
- Xài selenium

KHÓ KHĂN

- Không làm được phần đọc thêm bình luận và đọc tiếp phần mở rộng của bình luận dù đã cố gắng

ây đến với con em mình, song nêu cuộc đời đã sắp xếp như vậy không cho ai tất cả, cũng không lấy hết của ai bao giờ".
 ặc biệt nhất về trẻ tự kỷ mà mình biết được, đó là sự trong trẻo, nh cùng chúng, dấu cho thời ... [Đọc tiếp](#)

Trả lời 00:17 5/4

sự may mắn và hạnh phúc sẽ đến với cháu trong cuộc đời còn rất dài phía trước

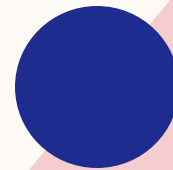
👍 Thích 🍷 222 Trả lời 02:44 5/4

Xem thêm ý kiến

GIẢI PHÁP

- Từ bỏ

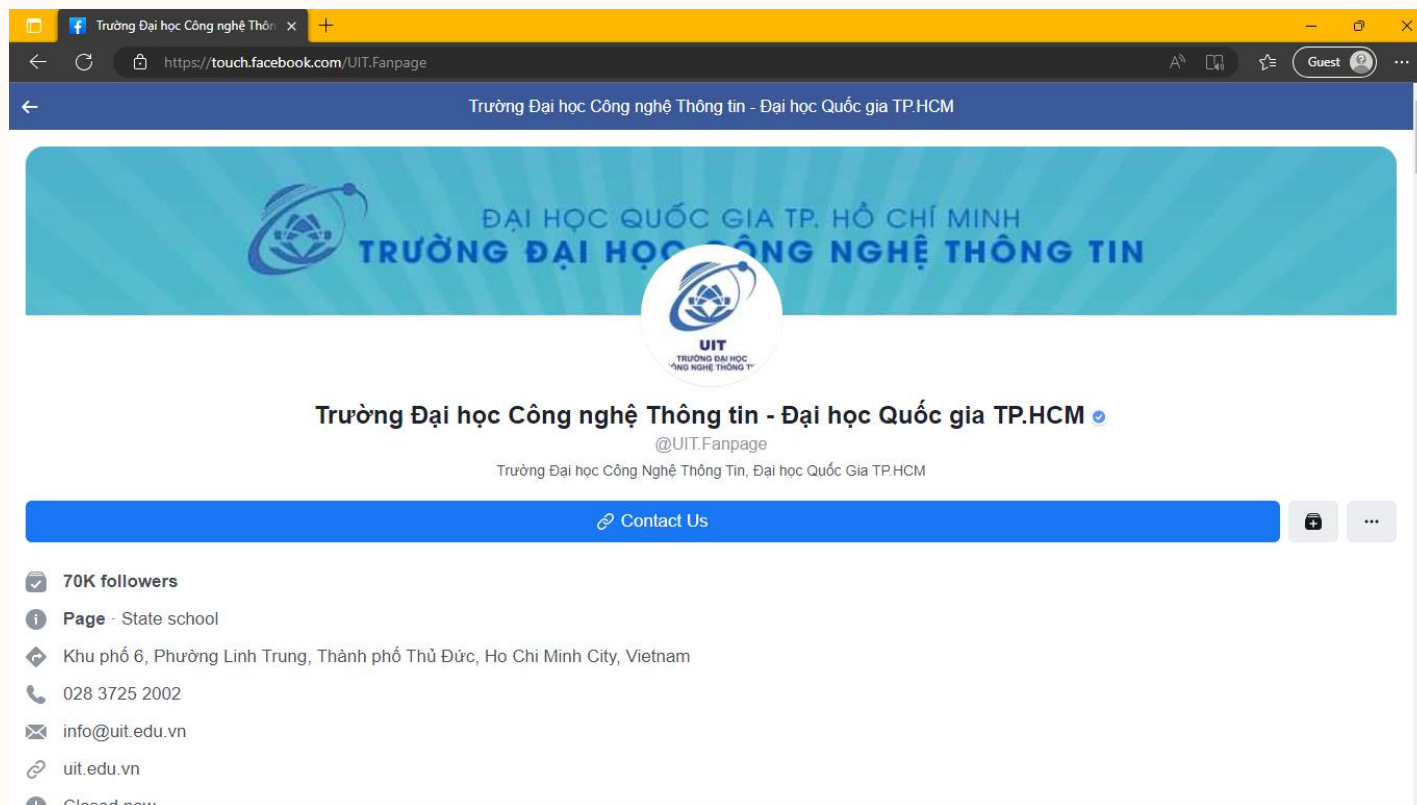
CRAWLER THU THẬP BÀI
VIẾT, BÌNH LUẬN TỪ 1
TRANG FACEBOOK



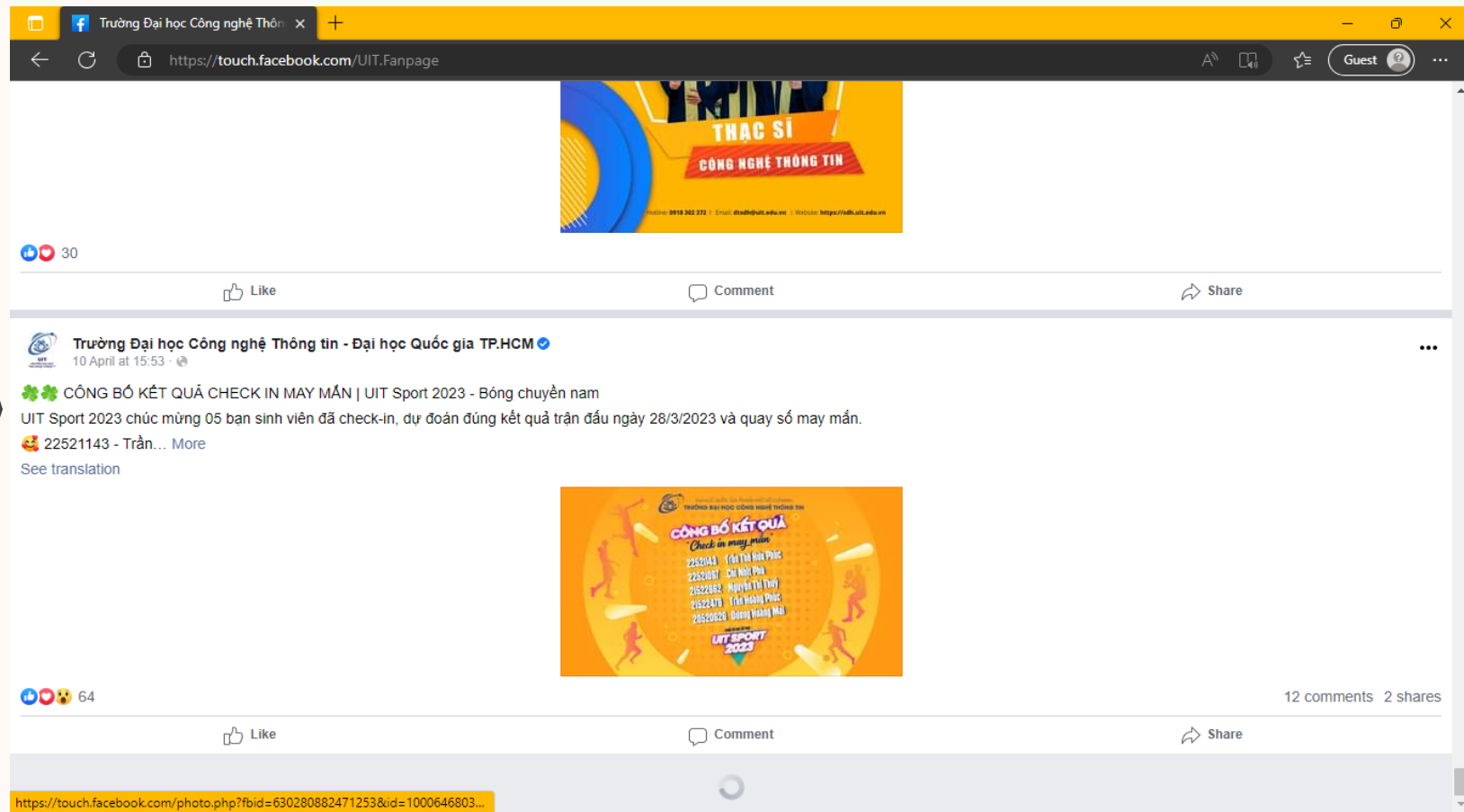
LẤY LINK FACEBOOK PAGE

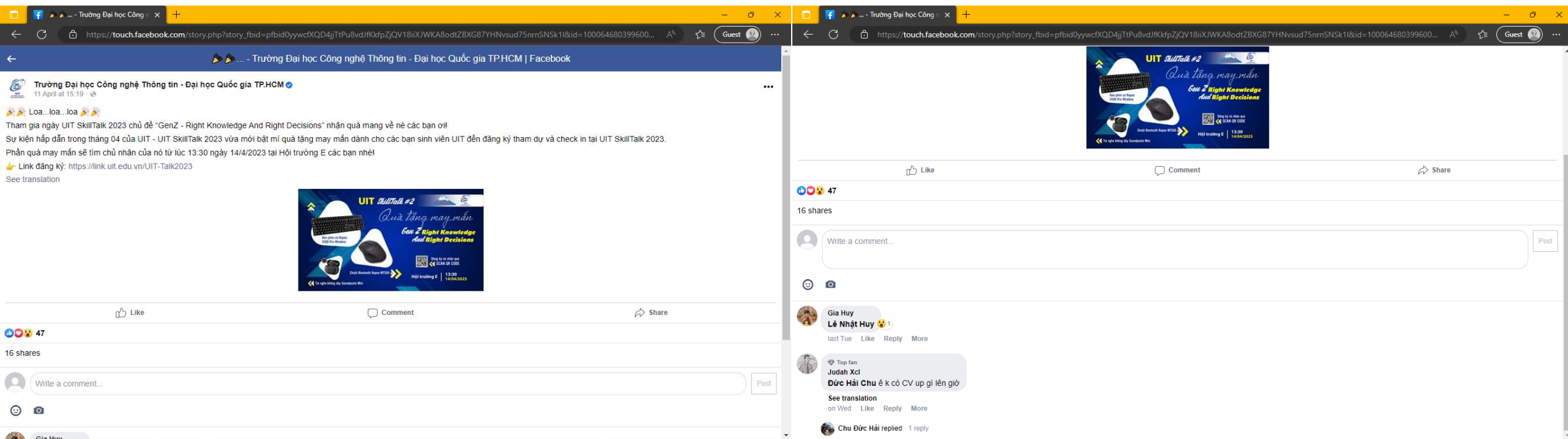


**ĐĂNG NHẬP,
TRUY CẬP
ĐƯỜNG DẪN
(TOUCH.FACEBOOK...)**



SCROLL DOWN
=> GET
ELEMS(POSTS)





TRUY CẬP VÀO TỪNG POST => GET
POST, COMMENTS

LƯU VÀO DATABASE

The screenshot displays the Microsoft SQL Server Management Studio interface. The title bar indicates the connection to 'DESKTOP-NHATMIN\SQLEXPRESS02.FB (DESKTOP-NHATMIN\DELL (52))*'. The 'Object Explorer' on the left shows the database structure, including 'Databases', 'System Databases', 'Database Snapshots', 'FB', 'FOOTBALL', 'QLSINHVIEN', 'Security', 'Server Objects', 'Replication', 'Management', and 'XEvent Profiler'. The 'Query Editor' in the center shows a SQL query:

```
select * from posts  
select * from comments  
select * from replies
```

. The 'Results' pane at the bottom displays the execution results, which are organized into two tables. The first table shows a list of posts with columns for post_id, author, and message. The second table shows a list of replies with columns for rep_id, rep_to, rep_author, and rep_message. The status bar at the bottom indicates 'Query executed successfully.' and '571 rows'.

post_id	author	message
633678192131522	100064680399600	Good Morning!~~~~ Chỉ muốn nói là mình thêm cái cảm giác được r...
633697878796220	100064680399600	Mở đăng ký "UIT Olympic" CLB Taekwondo UIT cùng với Khoa C...
6659092434140592	484451281604769	Mừng lễ năm mới!!!!!!!!!!!!!! See translation
6659147000801802	484451281604769	Gợi ý chơi nội tử cực vui và lành mạnh, hạnh phúc cùng chồng iu l...
6659181940798308	484451281604769	#32346: Mình có éc với n.y cũ quá ko =))) Mình năm nay 29, chưa ...

comment_id	post_id	author	message
100556756359217	6659147000801802	Top fan Đặng Trần Hương Quỳnh	Ồ kìa :))
102249646186055	6659147000801802	Trúc Ngô	Lan Anh
1032742281463803	6659147000801802	Hồ Ngọc Anh Anh	Được nhờ
1035185590793667	6659147000801802	Thu Vũ	Trần Đô Kết Cầu Nôi từ k a ?
105303259203452	6659147000801802	Lê Thị Ong	Uớc ☺

rep_id	rep_to	rep_author	rep_message
100777203002090	607371957971665	Trang Liluu	Bảo Ngọc để nào chơi chụp hình cho coi :))
1032160124857791	1557668498045951	Ngọc Lan	Solo Yasuo học y làm gì có tiền mua túi 3 củ
1207419896576132	721176563138540	Mai Phúo...	Nguyễn Khánh Ly đội đại gia giàu ☺
1228320777889544	1879416459094820	Nguyễn ...	Lâm Đôla hahaha_tag cho vui thôi chứ tk ...
1239046023443420	1907623116239106	Nguyễn ...	Ngọc Thương áp dụng ngay
1259820321297078	3473094489685312	Thắng T...	Hà Phương hết tiền :))
1266947557574181	780583213404579	Trung Duy	Nam Trường ôn toán đi ngài
1278341799760863	259162276539056	Thắng	Công Nguyễn chuyên nểng tư gia đình ngư...

KHÓ KHĂN

- Drama quá nhiều => tốn thời gian
- Khó dùng API (facebook graph API access token)
- Facebook là dynamic web, cấu trúc phức tạp (www.facebook.com/...)
- Comments – replies phân cấp
- Account banned

GIẢI PHÁP

- ...
- Dùng selenium hoặc facebook scraper, ...
- Chuyển sang touch.facebook.com, hạn chế về mặt cấu trúc comment
- Chia ra hai bảng comments, replies lưu vào database
- Tạo thêm acc khác

KHÓ KHĂN

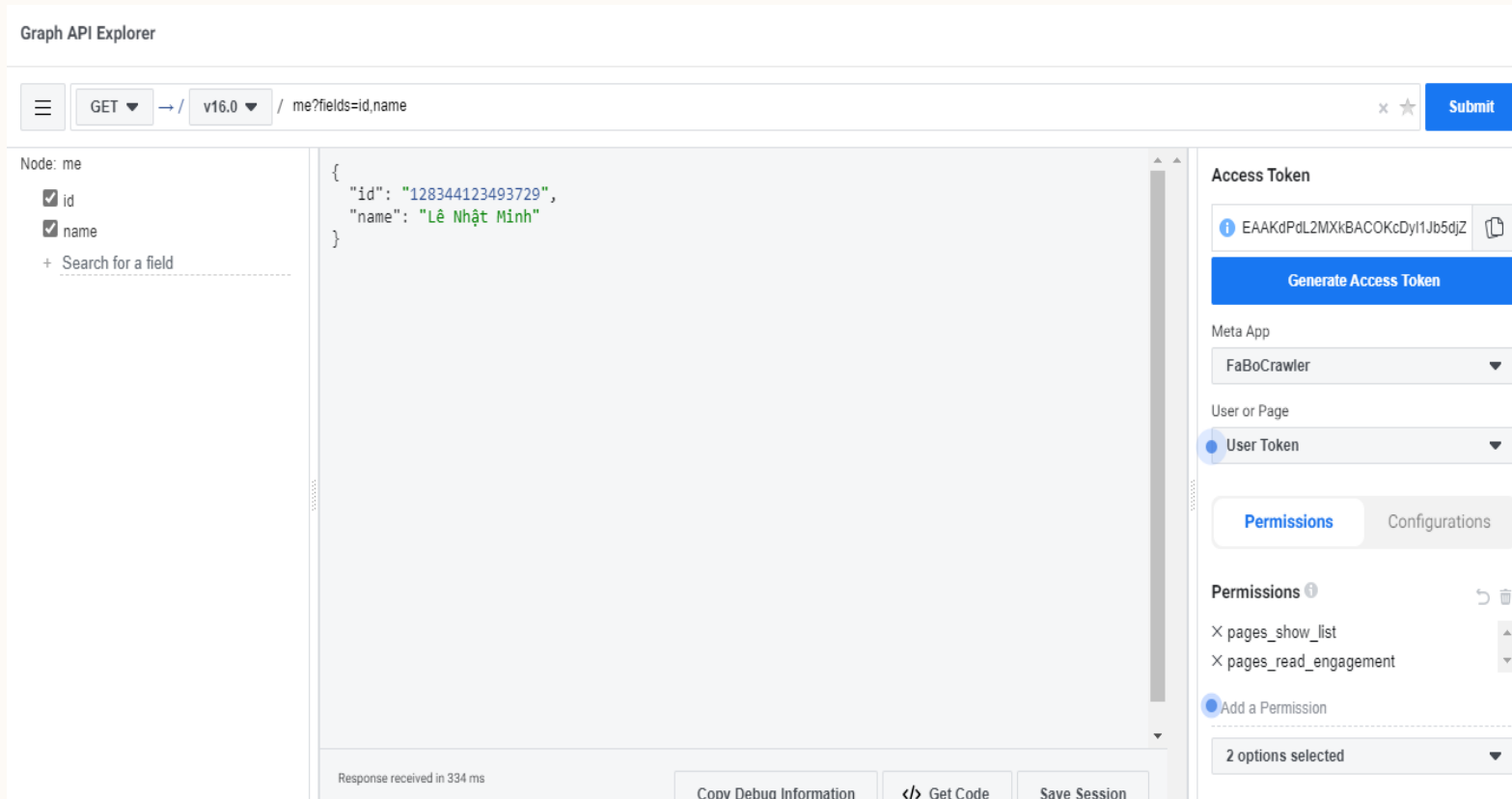
GIẢI PHÁP



...

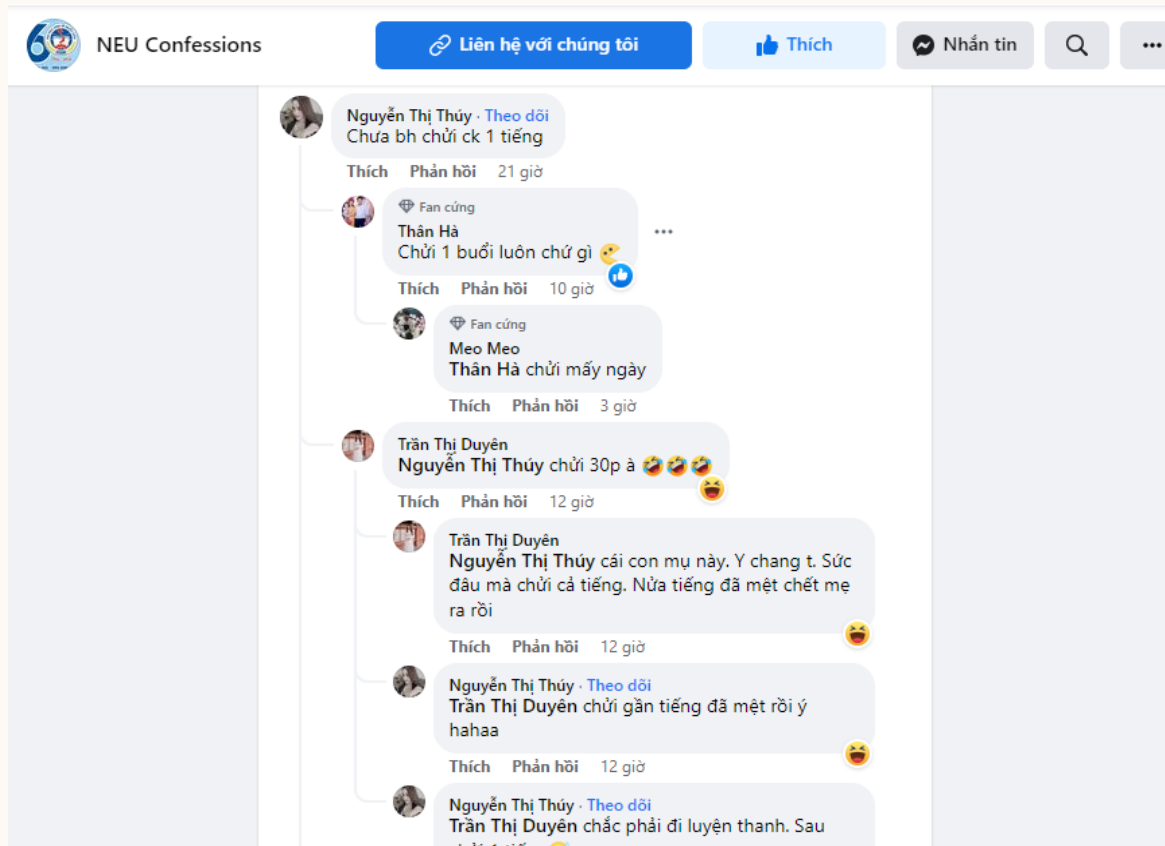
KHÓ KHĂN

GIẢI PHÁP



- Dùng selenium, beautiful soup ...

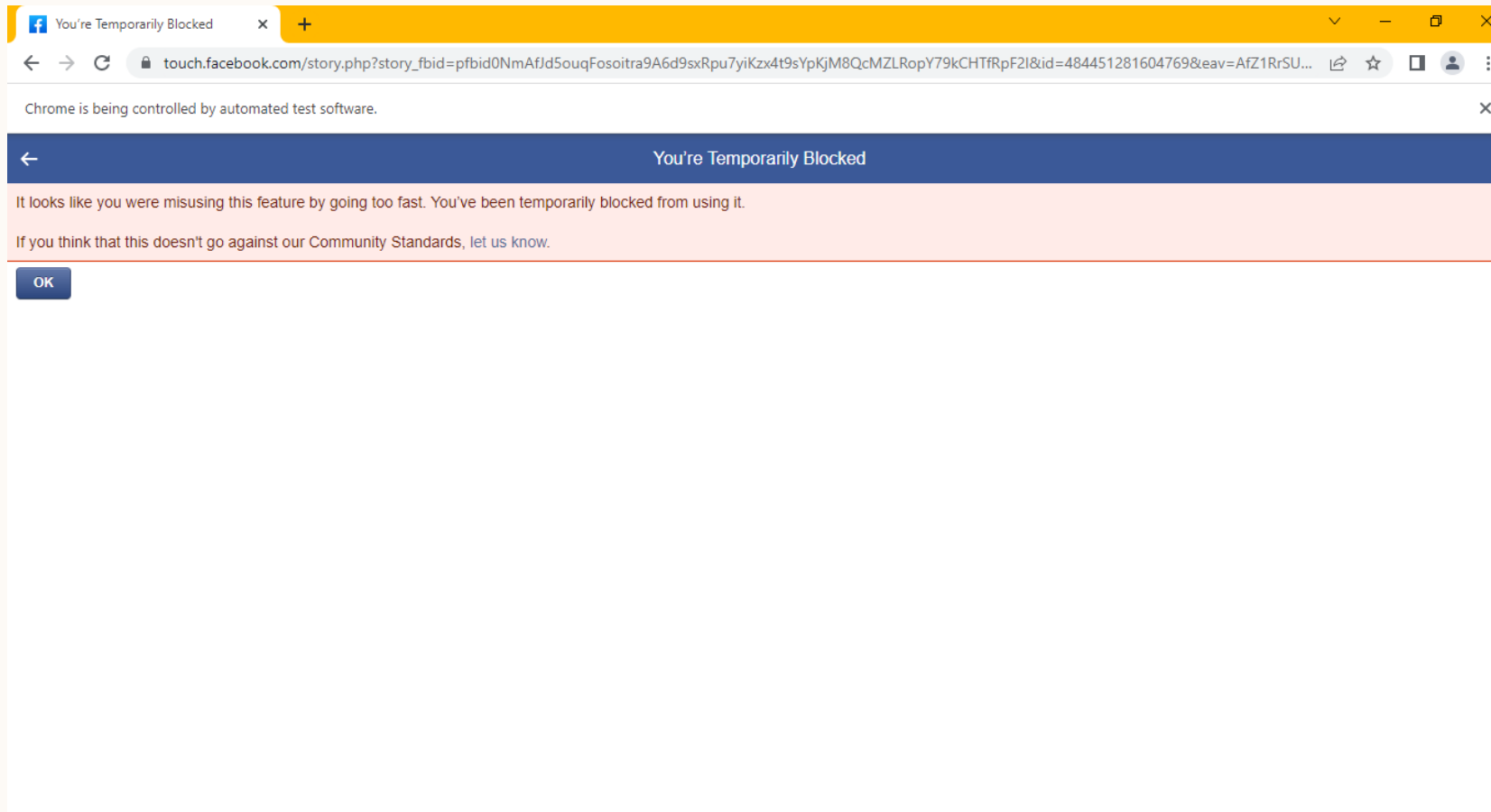
KHÓ KHĂN



GIẢI PHÁP



KHÓ KHĂN



GIẢI PHÁP

Tạo tài khoản khác

The background features a large white circle on the left and a large pink circle on the right, both partially overlapping a dark blue background. The pink circle contains several thin, white, concentric circular lines.

THANK YOU