# Mô hình tổng quát

A diagram of a software system

Description automatically generated

# Điều kiện cần

Điều kiện cần để thu dữ liệu từ các nguồn:

* Facebook: Để thu dữ liệu facebook cần 3 – 5 account đăng nhập không cần xác thực 2 bước đã tham gia các group cần lấy thông tin.
* Trang báo, tin tức và các nguồn khác: mỗi trang cần 1 scraper riêng.

# Thống kê các nguồn dữ liệu

## Dữ liệu từ Facebook

### Các trang Facebook

|  |  |  |  |
| --- | --- | --- | --- |
| Stt | Page id | Tên page | Url |
| 1 | PTITNews | **PTIT News** | **https://www.facebook.com/PTITNews** |
| 2 | PtitCfsOfficial | PTIT Confessions | https://www.facebook.com/PtitCfsOfficial |
| 3 | DRAMAPTIT | Drama PTIT | https://www.facebook.com/DRAMAPTIT |
| 4 | Antiptit | Hội anti HV công nghệ BƯU ĐIỆN | https://www.facebook.com/Antiptit |

### Các nhóm Facebook

|  |  |  |
| --- | --- | --- |
| Stt | Group id | Tên page |
| 1 | 518074720443382 | [**D23 Học Viện Công Nghệ Bưu Chính Viễn Thông Hà Nội (PTIT)**](https://www.facebook.com/groups/d23hvcnbcvt/) |
| 2 | 182737627268456 | [**D22 Học Viện Công Nghệ Bưu Chính Viễn Thông Hà Nội (PTIT)**](https://www.facebook.com/groups/182737627268456/) |
| 3 | 662206768144479 | [**D22 PTIT Hà Nội**](https://www.facebook.com/groups/d22ptithanoi/) |
| 4 | d23ptithanoi | [**D23 PTIT Hà Nội**](https://www.facebook.com/groups/d23ptithanoi/) |
| 5 | 964834897893077 | [**Tân Sinh Viên PTIT (PTIT D23)**](https://www.facebook.com/groups/964834897893077/) |
| 6 | buuchinhvienthongptit | [**2K5 Học viện Công nghệ Bưu chính Viễn thông (D21 PTIT, D22 PTIT, D23 PTIT)**](https://www.facebook.com/groups/buuchinhvienthongptit/) |
| 7 | tuyendungPtit | [**KÊNH TUYỂN DỤNG SINH VIÊN PTIT**](https://www.facebook.com/groups/tuyendungPtit/) |
| 8 | 862343463794409 | [**Tuyển dụng PTIT (Official) Miền Bắc**](https://www.facebook.com/groups/862343463794409/) |
| 9 | 667258140376081 | [**Chợ Pờ Tít (PTIT's Market) <333**](https://www.facebook.com/groups/667258140376081/) |
| 10 | 306831584046186 | [**Học bổng ptit**](https://www.facebook.com/groups/306831584046186/) |

# Hướng dẫn thu dữ liệu

## Thu dữ liệu từ Facebook

### Cài đặt thư viện facebook-scraper

Link github: [**https://github.com/kevinzg/facebook-scraper**](https://github.com/kevinzg/facebook-scraper)

Cài đặt: ***pip install facebook-scraper***

### Hướng dẫn sử dụng

#### Thu dữ liệu từ 1 trang facebook

Dùng hàm get\_post() với các tham số:

* options={"comments": True, "reactions": True, "allow\_extra\_requests": True} và extra\_info=True để lấy thông tin comments và reactions
* name\_page là page\_id
* pages là số bài muốn lấy, ví dụ: truyền 1 là 10 bài, 2 là 20 bài
* credentials là 1 tuple gồm username và password của account đăng nhập facebook

|  |
| --- |
| from facebook\_scraper import get\_posts  import pandas as pd  import datetime  post\_list = []  """  PTITNews  PtitCfsOfficial  DRAMAPTIT  Antiptit   """  name\_page = "Antiptit"  number\_page = 8  df = pd.DataFrame(columns=[])  for post in get\_posts(name\_page,                      options={"comments": True, "reactions": True, "allow\_extra\_requests": True},                      extra\_info=True, pages=number\_page, credentials=("dan00101001", "dan0020@@")):      post\_list.append(post)      df\_extend = pd.DataFrame.from\_dict(post, orient='index')      df\_extend = df\_extend.transpose()      df = pd.concat([df, df\_extend])  print(len(post\_list))  df.to\_csv("data\_" + name\_page +"\_"+str(datetime.datetime.now())[0:10]+ "\_" + str(len(post\_list))+".csv", index=False) |

#### Thu dữ liệu từ 1 group

Dùng hàm get\_post() với các tham số:

* options={"comments": True, "reactions": True, "allow\_extra\_requests": True} và extra\_info=True để lấy thông tin comments và reactions
* group là group\_id
* pages là số bài muốn lấy, ví dụ: truyền 1 là 10 bài, 2 là 20 bài
* credentials là 1 tuple gồm username và password của account đăng nhập facebook
* page\_limit là số trang đọc data ở group
* timeout là timeout gọi request lên facebook
* days\_limit là số ngày tính từ lúc gọi crawl về trước
* filename là file lưu data
* resume\_file là file lưu next page đang truy cập

|  |
| --- |
| import facebook\_scraper as fs  from facebook\_scraper import write\_posts\_to\_csv  import datetime  """  {  "518074720443382": "D23 Học Viện Công Nghệ Bưu Chính Viễn Thông Hà Nội (PTIT)",  "182737627268456": "D22 Học Viện Công Nghệ Bưu Chính Viễn Thông Hà Nội (PTIT)",  "662206768144479": "D22 PTIT Hà Nội",  "d23ptithanoi": "D23 PTIT Hà Nội",  "964834897893077": "Tân Sinh Viên PTIT (PTIT D23)",  "buuchinhvienthongptit": "2K5 Học viện Công nghệ Bưu chính Viễn thông (D21 PTIT, D22 PTIT, D23 PTIT)",  "tuyendungPtit": "KÊNH TUYỂN DỤNG SINH VIÊN PTIT",  "862343463794409": "Tuyển dụng PTIT (Official) Miền Bắc",  "667258140376081": "Chợ Pờ Tít (PTIT's Market) <333",  "306831584046186": "Học bổng ptit",  }   """  start\_time = datetime.datetime.now()  group\_id = "518074720443382"  group\_name = "D23 Học Viện Công Nghệ Bưu Chính Viễn Thông Hà Nội (PTIT)"  # Saves the first 100 pages  try:      n = 200      for i in range(1, n+1):          fs.write\_posts\_to\_csv(              group=group\_id, # The method uses get\_posts internally so you can use the same arguments and they will be passed along              page\_limit=n, # Will save the first 100 pages              credentials=("dan00101001", "dan0020@@"),              timeout=60,              options={"comments": True, "reactions": True, "allow\_extra\_requests": True},              extra\_info=True,              filename=f'./data/{group\_id}\_{i}\_{str(datetime.datetime.now())[0:10]}.csv', # # Will throw an error if the file already exists              resume\_file='next\_page\_' + group\_name + '.txt', # Will save a link to the next page in this file after fetching it and use it when starting.              matching='.+', # A regex can be used to filter all the posts matching a certain pattern (here, we accept anything)              not\_matching='^Warning', # And likewise those that don't fit a pattern (here, we filter out all posts starting with "Warning")              # keys=[              #     'post\_id',              #     'text',              #     'timestamp',              #     'time',              #     'user\_id',              # ], # List of the keys that should be saved for each post, will save all keys if not set              format='csv', # Output file format, can be csv or json, defaults to csv              days\_limit=3650 # Number of days for the oldest post to fetch, defaults to 3650          )      print("Time run: ", datetime.datetime.now() - start\_time)      print("Time end: ", datetime.datetime.now())  except Exception as e:      print(e)      print("Time run: ", datetime.datetime.now() - start\_time)      print("Time end: ", datetime.datetime.now()) |

# Tổng hợp dữ liệu và xử lí đã thu thập

## Tổng hợp dữ liệu

### Facebook

Dưới đây là tổng hợp thu thập dữ liệu từ các page và group facebook đã thống kê từ phần 3:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Stt | Page id | Tên page | n=1  (10 bài) | n=2  (20 bài) | n=5  (50 bài) | n=8  (80 bài) | n=10  (100 bài) |
| 1 | PTITNews | **PTIT News** | **49.3s** | **1m24.5s** | **3m25.8s** | **4m40.4s** | **17m** |
| 2 | PtitCfsOfficial | PTIT Confessions | 34.7s | 49.9s | 2m6s | 2m37.2s | 3m |
| 3 | DRAMAPTIT | Drama PTIT | 29.1s | 1m5.5s | 2m19s | 3m14.4s | 4m |
| 4 | Antiptit | Hội anti HV công nghệ BƯU ĐIỆN | 31.8s | 1m7.7s | 2m53s | 4m19.6s | 11m |

* Dữ liệu thu thập từ các page trong tháng 11 có tổng hơn 1000 bài viết

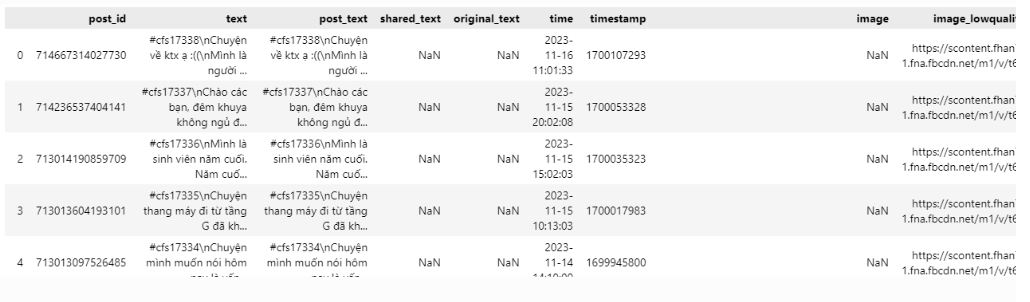
|  |  |  |  |
| --- | --- | --- | --- |
| Stt | Group id | Tên page | Kịch bản đã thực hiện |
| 1 | 518074720443382 | [**D23 Học Viện Công Nghệ Bưu Chính Viễn Thông Hà Nội (PTIT)**](https://www.facebook.com/groups/d23hvcnbcvt/) | Timeout=60s, n=200   * (7270-3764) bài   Time run: 14:10:02  Time end: 2023-11-25 08:06:53   * 850m(18h hôm trước-8h06p hôm sau) |
| 2 | 182737627268456 | [**D22 Học Viện Công Nghệ Bưu Chính Viễn Thông Hà Nội (PTIT)**](https://www.facebook.com/groups/182737627268456/) |  |
| 3 | 662206768144479 | [**D22 PTIT Hà Nội**](https://www.facebook.com/groups/d22ptithanoi/) | Timeout=40s, n=300   * (3164-1576) bài   Time run: 7:35:59  Time end: 2023-11-25 17:51:01   * 455m 59.8s |
| 4 | d23ptithanoi | [**D23 PTIT Hà Nội**](https://www.facebook.com/groups/d23ptithanoi/) | Timeout=50s, n=300   * (4584-1928)bài   Time run: 10:15:54  Time end: 2023-11-25 19:36:25   * 615m 54.7s |
| 5 | 964834897893077 | [**Tân Sinh Viên PTIT (PTIT D23)**](https://www.facebook.com/groups/964834897893077/) |  |
| 6 | buuchinhvienthongptit | [**2K5 Học viện Công nghệ Bưu chính Viễn thông (D21 PTIT, D22 PTIT, D23 PTIT)**](https://www.facebook.com/groups/buuchinhvienthongptit/) |  |
| 7 | tuyendungPtit | [**KÊNH TUYỂN DỤNG SINH VIÊN PTIT**](https://www.facebook.com/groups/tuyendungPtit/) |  |
| 8 | 862343463794409 | [**Tuyển dụng PTIT (Official) Miền Bắc**](https://www.facebook.com/groups/862343463794409/) |  |
| 9 | 667258140376081 | [**Chợ Pờ Tít (PTIT's Market) <333**](https://www.facebook.com/groups/667258140376081/) |  |
| 10 | 306831584046186 | [**Học bổng ptit**](https://www.facebook.com/groups/306831584046186/) |  |

* Dữ liệu thu thập từ các groups trong tháng 11 là gần 16100 bài viết

Toàn bộ dữ liệu được lưu trữ dạng csv.

## Xử lí dữ liệu

Dữ liệu thu về có 54 trường: 'post\_id', 'text', 'post\_text', 'shared\_text', 'original\_text', 'time', 'timestamp', 'image', 'image\_lowquality', 'images', 'images\_description', 'images\_lowquality', 'images\_lowquality\_description', 'video', 'video\_duration\_seconds', 'video\_height', 'video\_id', 'video\_quality', 'video\_size\_MB', 'video\_thumbnail', 'video\_watches', 'video\_width', 'likes', 'comments', 'shares', 'post\_url', 'link', 'links', 'user\_id', 'username', 'user\_url', 'is\_live', 'factcheck', 'shared\_post\_id', 'shared\_time', 'shared\_user\_id', 'shared\_username', 'shared\_post\_url', 'available', 'comments\_full', 'reactors', 'w3\_fb\_url', 'reactions', 'reaction\_count', 'with', 'page\_id', 'sharers', 'image\_id', 'image\_ids', 'was\_live', 'fetched\_time', 'video\_ids', 'videos'



Tuy nhiên có những trường dữ liệu không cần thiết nên sau khi trích chọn dữ liệu chỉ còn 12 trường sau: "post\_id", "text", "time", "timestamp", "comments", "shares", "post\_url", "user\_id", "username", "reactions", "reaction\_count", "comments\_full"

A screenshot of a cell phone

Description automatically generated