HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

PHÂN TÍCH VÀ DỰ ĐOÁN GIÁ BẤT ĐỘNG SẢN HÀ NỘI

Nhóm 20: Đào Thành Mạnh
Trần Hữu Huy
Nguyễn Đức Thịnh
Nguyễn Anh Quân
Võ Minh Trí

ONE LOVE. ONE FUTURE.

Nội dung

I. Giới thiệu bài toán
II. Thu thập và xử lý dữ liệu
III. Phương pháp thực nghiệm
IV. Kết quả
V. Triển khai

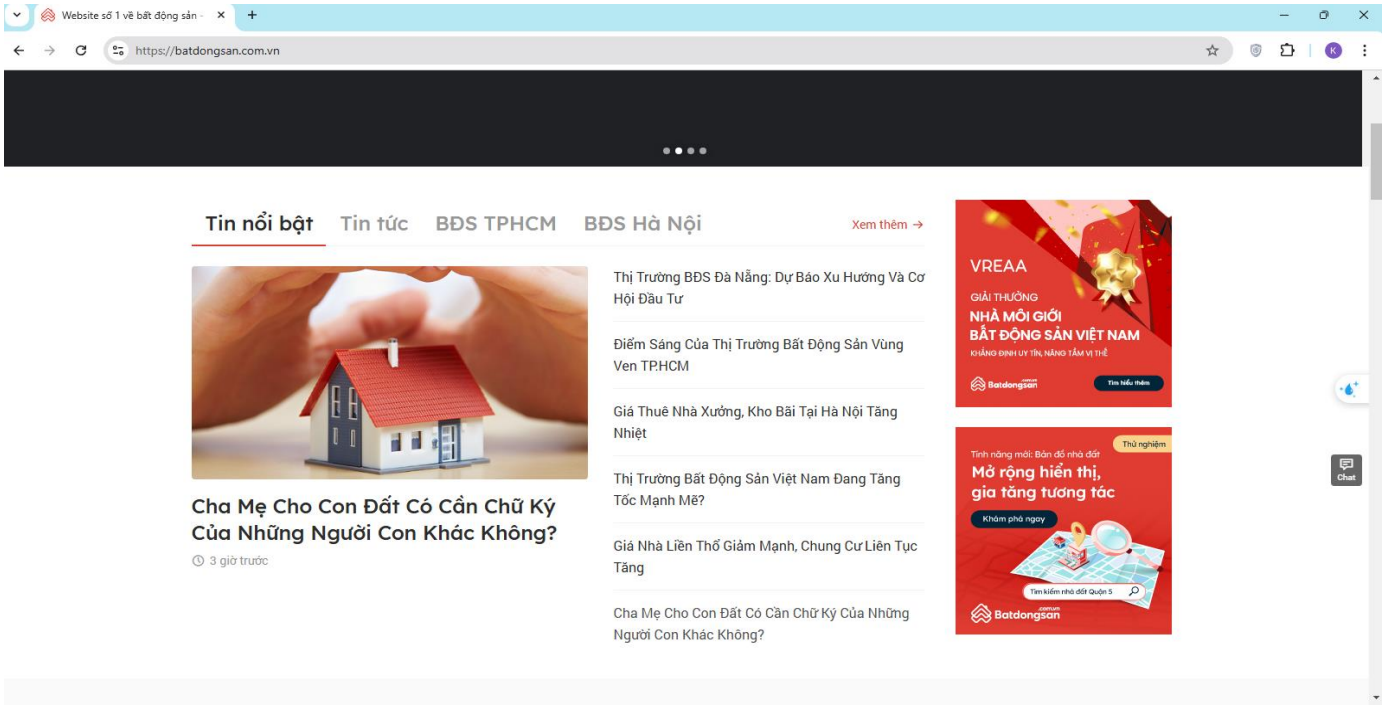
GIỚI THIỆU BÀI TOÁN

- Bất động sản luôn là một trong những lĩnh vực thu hút nhất của nền kinh tế
- Đặc biệt, thị trường bất động sản ở Hà Nội là một thị trường lớn và phức tạp và cung cấp nguồn dữ liệu phong phú
- Để có cái nhìn chính xác nhất về thị trường và từ đó đưa ra quyết định cụ thể

→ Nhóm thực hiện đề tại, áp dụng phương pháp của khoa học dữ liệu để phân tích, dự đoán

Thu thập dữ liệu

- Trang web sử dụng: batdongsan.com.vn
  - Lý do: chứa lịch sử giá
  - Vấn đề: tích hợp Cloudflare (stop bot attacks)



Thu thập dữ liệu

- Công cụ sử dụng: selenium
  - Framework kiểm thử tự động
  - Giả lập hành vi con người (bypass Cloudflare)
  - Nhược điểm: chậm hơn các công cụ crawl chuyên dụng

Thu thập dữ liệu

- Tổng quan dữ liệu: hầu hết các trường dữ liệu chứa thông tin quan trọng như tiêu đề, mô tả, giá, kiểu bất động sản hay địa chỉ đều được điền đầy đủ. Các trường thông tin thêm thì tương đối thưa.

{'_adr': 'Đường Nguyễn Hoàng, Phường Mỹ Đình 2, Nam Từ Liêm, Hà Nội',
'need': 'Cho thuê',
'city': 'Hà Nội',
'dstr': 'Nam Từ Liêm',
'addr': 'Văn phòng tại đường Nguyễn Hoàng',
'_ttl': 'CHO THUÊ VĂN PHÒNG RẺ, ĐẸP NHẤT MỸ ĐÌNH, VỊ TRÍ ĐẮC ĐỊA, DIỆN TÍCH SỬ DỤNG 1,908 m/sàn',
'_dsr': '\n Cho thuê tòa nhà Suced, tọa lạc tại vị trí đắc địa trên...Ms. Vân Anh 0915 055 ***\n',
'locT': '21.031360957798,105.774467319522',
'area': '1.908 m²',
'cost': '230 triệu/m²',
'flor': '12 tầng',
'lgal': 'Sổ đỏ/ Sổ hồng',
'strd': '15/10/2024',
'endd': '30/10/2024',
'pttl': 'Tin VIP Vàng',
'prid': '41217655'},

Thu thập dữ liệu

BỘ TÀI NGUYÊN VÀ MÔI TRƯỜNG

Số: 10 /2022/TT-BTNMT

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Hà Nội, ngày 29 tháng 9 năm 2022

THÔNG TƯ

Ban hành Danh mục địa danh dân cư, sơn văn, thuỷ văn, kinh tế - xã hội phục vụ công tác thành lập bản đồ thành phố Hà Nội

Căn cứ Luật Đo đạc và bản đồ ngày 14 tháng 6 năm 2018;

Căn cứ Nghị định số 68/2022/NĐ-CP ngày 22 tháng 9 năm 2022 của Chính phủ quy định chức năng, nhiệm vụ, quyền hạn và cơ cấu tổ chức của Bộ Tài nguyên và Môi trường;

Theo đề nghị của Cục trưởng Cục Đo đạc, Bản đồ và Thông tin địa lý Việt Nam và Vụ trưởng Vụ Pháp chế;

Bộ trưởng Bộ Tài nguyên và Môi trường ban hành Thông tư ban hành Danh mục địa danh dân cư, sơn văn, thuỷ văn, kinh tế - xã hội phục vụ công tác thành lập bản đồ thành phố Hà Nội.

Điều 1. Ban hành kèm theo Thông tư này Danh mục địa danh dân cư, sơn văn, thuỷ văn, kinh tế - xã hội phục vụ công tác thành lập bản đồ thành phố Hà Nội.

Điều 2. Thông tư này có hiệu lực thi hành kể từ ngày *15* tháng *1* năm 2022.

Điều 3. Bộ, cơ quan ngang Bộ, cơ quan thuộc Chính phủ, Ủy ban nhân dân các tỉnh, thành phố trực thuộc Trung ương và các tổ chức, cá nhân có liên quan chịu trách nhiệm thi hành Thông tư này./.

Nơi nhận:
- Văn phòng Quốc hội;
- Văn phòng Chính phủ;
- Các Bộ, cơ quan ngang Bộ, cơ quan thuộc Chính phủ;
- UBND các tỉnh, thành phố trực thuộc Trung ương;
- Sở Nội vụ và Sở TN&MT thành phố Hà Nội;
- Cục Kiểm tra văn bản QPPL (Bộ Tư pháp);
- Bộ trưởng và các Thứ trưởng Bộ TN&MT;
- Các đơn vị trực thuộc Bộ TN&MT, Cổng Thông tin điện tử Bộ TN&MT;
- Công báo, Cổng Thông tin điện tử Chính phủ;
- Lưu: VT, PC, ĐĐBĐVN.

KT. BỘ TRƯỞNG
THỨ TRƯỞNG

Nguyễn Thị Phương Hoa

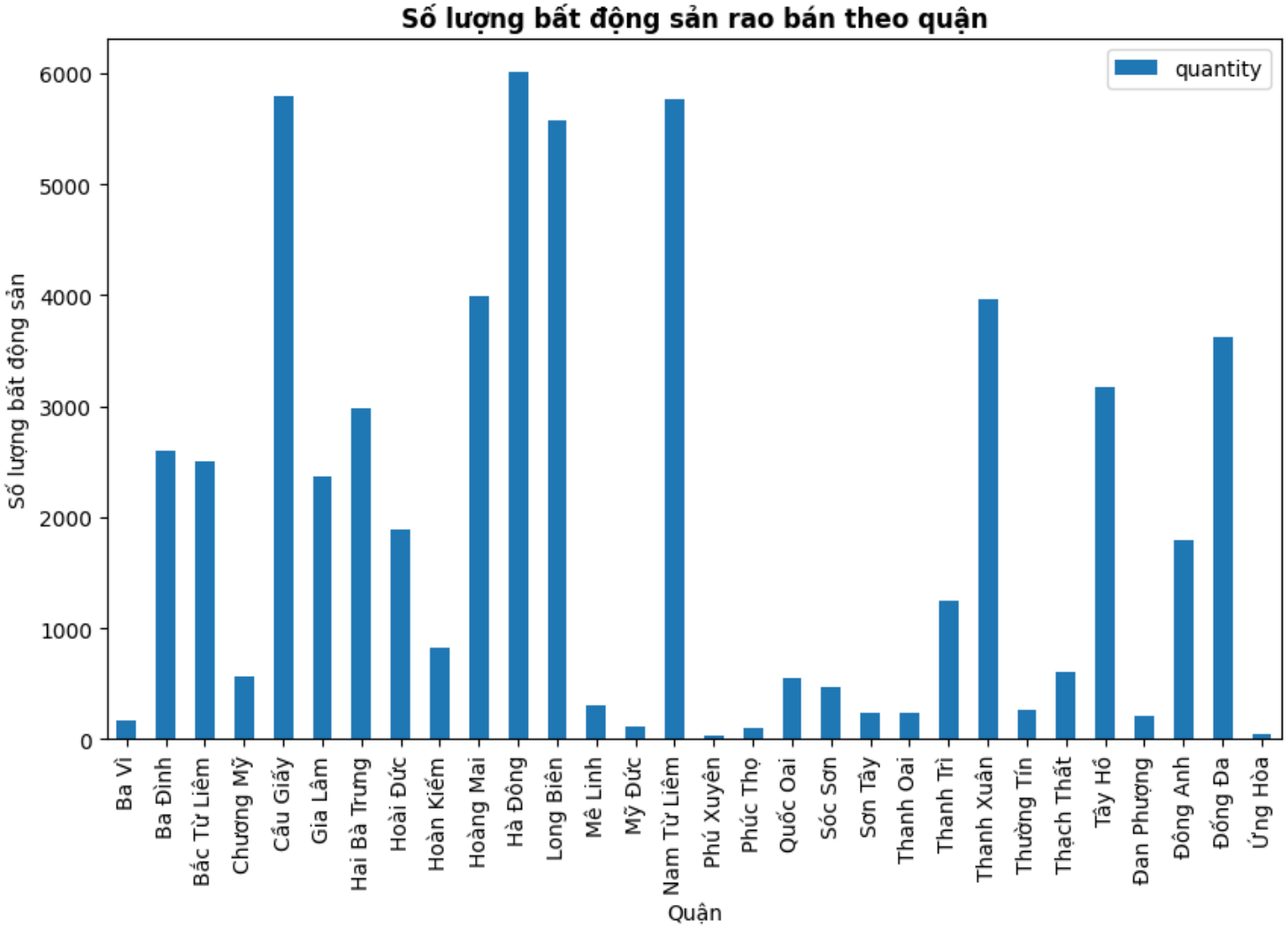
→

```
Địa danh,Nhóm đối tượng,Tên ĐVHC cấp xã,Tên ĐVHC cấp huyện,Vĩ độ,Kinh độ,Vĩ độ.1,Kinh độ.1,Vĩ độ.2,Kinh độ.2
quảng trường Ba Đình,KX,P. Điện Biên,Q. Ba Đình,21° 02' 18'',105° 50' 03'',,,,
thành cổ Hà Nội,KX,P. Điện Biên,Q. Ba Đình,21° 02' 06'',105° 50' 18'',,,,
Toà nhà trụ sở Bộ Ngoại giao,KX,P. Điện Biên,Q. Ba Đình,21° 02' 12'',105° 50' 06'',,,,
Trung tâm Hội nghị Quốc tế,KX,P. Điện Biên,Q. Ba Đình,21° 02' 04'',105° 49' 59'',,,,
Bảo tàng Chiến thắng B52,KX,P. Đội Cấn,Q. Ba Đình,21° 02' 10'',105° 49' 27'',,,,
Bảo tàng Hồ Chí Minh,KX,P. Đội Cấn,Q. Ba Đình,21° 02' 11'',105° 49' 51'',,,,
Bộ Khoa học và Công nghệ,KX,P. Đội Cấn,Q. Ba Đình,21° 02' 08'',105° 49' 12'',,,,
chợ Ngọc Hà,KX,P. Đội Cấn,Q. Ba Đình,21° 02' 05'',105° 49' 47'',,,,
chùa Bát Tháp,KX,P. Đội Cấn,Q. Ba Đình,21° 02' 13'',105° 49' 14'',,,,
chùa Một Cột (chùa Diên Hựu),KX,P. Đội Cấn,Q. Ba Đình,21° 02' 13'',105° 49' 54'',,,,
đền Miếu Trắng,KX,P. Đội Cấn,Q. Ba Đình,21° 02' 06'',105° 49' 19'',,,,
đình Vạn Phúc,KX,P. Đội Cấn,Q. Ba Đình,21° 02' 06'',105° 49' 20'',,,,
Đài Truyền hình Việt Nam,KX,P. Giảng Võ,Q. Ba Đình,21° 01' 38'',105° 48' 40'',,,,
đình Giảng Võ,KX,P. Giảng Võ,Q. Ba Đình,21° 01' 34'',105° 48' 53'',,,,
hồ Giảng Võ,TV,P. Giảng Võ,Q. Ba Đình,21° 01' 45'',105° 49' 04'',,,,
Bộ Tư lệnh Thông tin,KX,P. Kim Mã,Q. Ba Đình,21° 01' 55'',105° 49' 32'',,,,
Bộ Y tế,KX,P. Kim Mã,Q. Ba Đình,21° 01' 47'',105° 49' 23'',,,,
chùa Kim Sơn,KX,P. Kim Mã,Q. Ba Đình,21° 01' 57'',105° 49' 29'',,,,
đình Kim Mã,KX,P. Kim Mã,Q. Ba Đình,21° 01' 56'',105° 49' 16'',,,,
đình Xuân Biểu,KX,P. Kim Mã,Q. Ba Đình,21° 02' 02'',105° 49' 44'',,,,
nhà hát Chèo Việt Nam,KX,P. Kim Mã,Q. Ba Đình,21° 01' 58'',105° 49' 31'',,,,
chùa Vĩnh Khánh,KX,P. Liễu Giai,Q. Ba Đình,21° 02' 31'',105° 49' 02'',,,,
Cung thể thao tổng hợp Quần Ngựa,KX,P. Liễu Giai,Q. Ba Đình,21° 02' 29'',105° 48' 45'',,,,
đền Liễu Giai,KX,P. Liễu Giai,Q. Ba Đình,21° 02' 17'',105° 48' 56'',,,,
đền Vĩnh Phúc,KX,P. Liễu Giai,Q. Ba Đình,21° 02' 30'',105° 49' 01'',,,,
đình Liễu Giai,KX,P. Liễu Giai,Q. Ba Đình,21° 02' 17'',105° 48' 55'',,,,
```

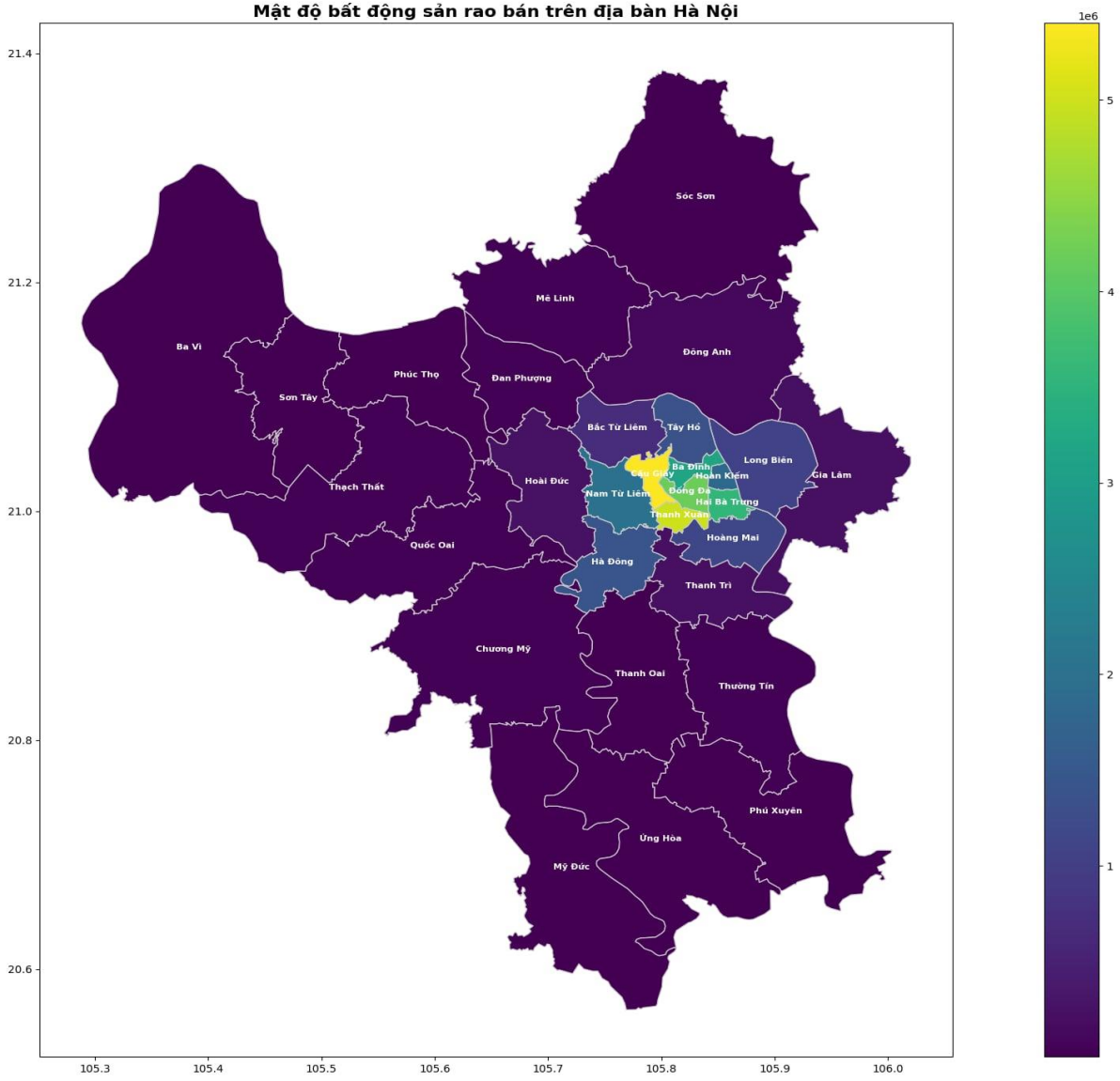
Xử lý dữ liệu

- Loại bỏ các bản ghi trùng lặp
- Xử lý số
  - Chuyển các trường string về float/int
  - Loại bỏ outlier
- Xử lý văn bản
  - Chuẩn hóa text, trích xuất thông tin
  - Chuẩn hóa unicode về dạng NFC
  - Chuẩn hóa bảng mã Unicode tiếng Việt về Unicode dựng sẵn (không phân biệt được bằng mắt thường)
  - Chuẩn hóa kiểu gõ dấu
  - Chuẩn hóa chính tả
  - Loại bỏ các ký tự đặc biệt không mang ý nghĩa ()
  - Loại bỏ các dấu câu không mang ý nghĩa

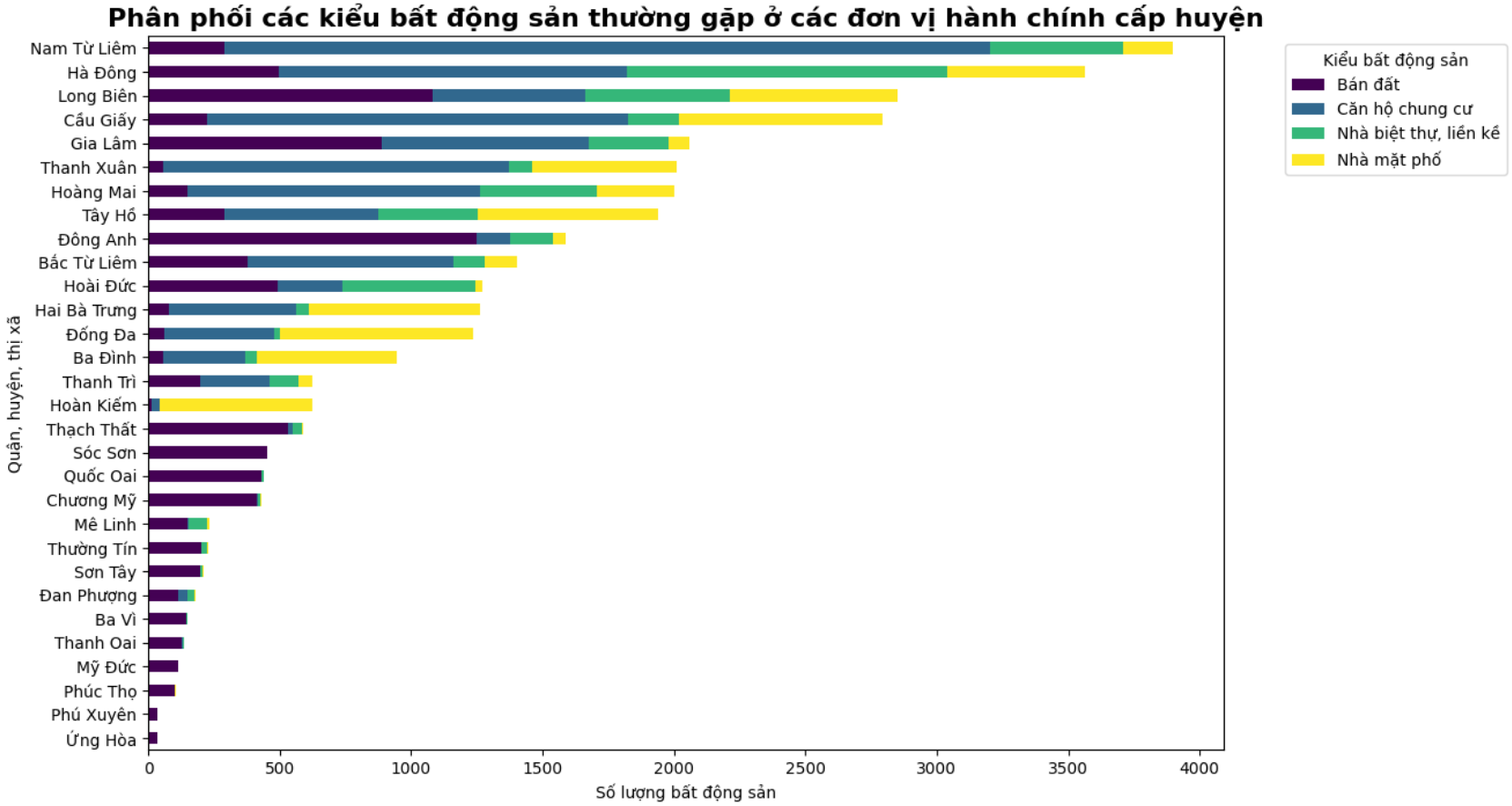
THU THẬP VÀ XỬ LÝ DỮ LIỆU



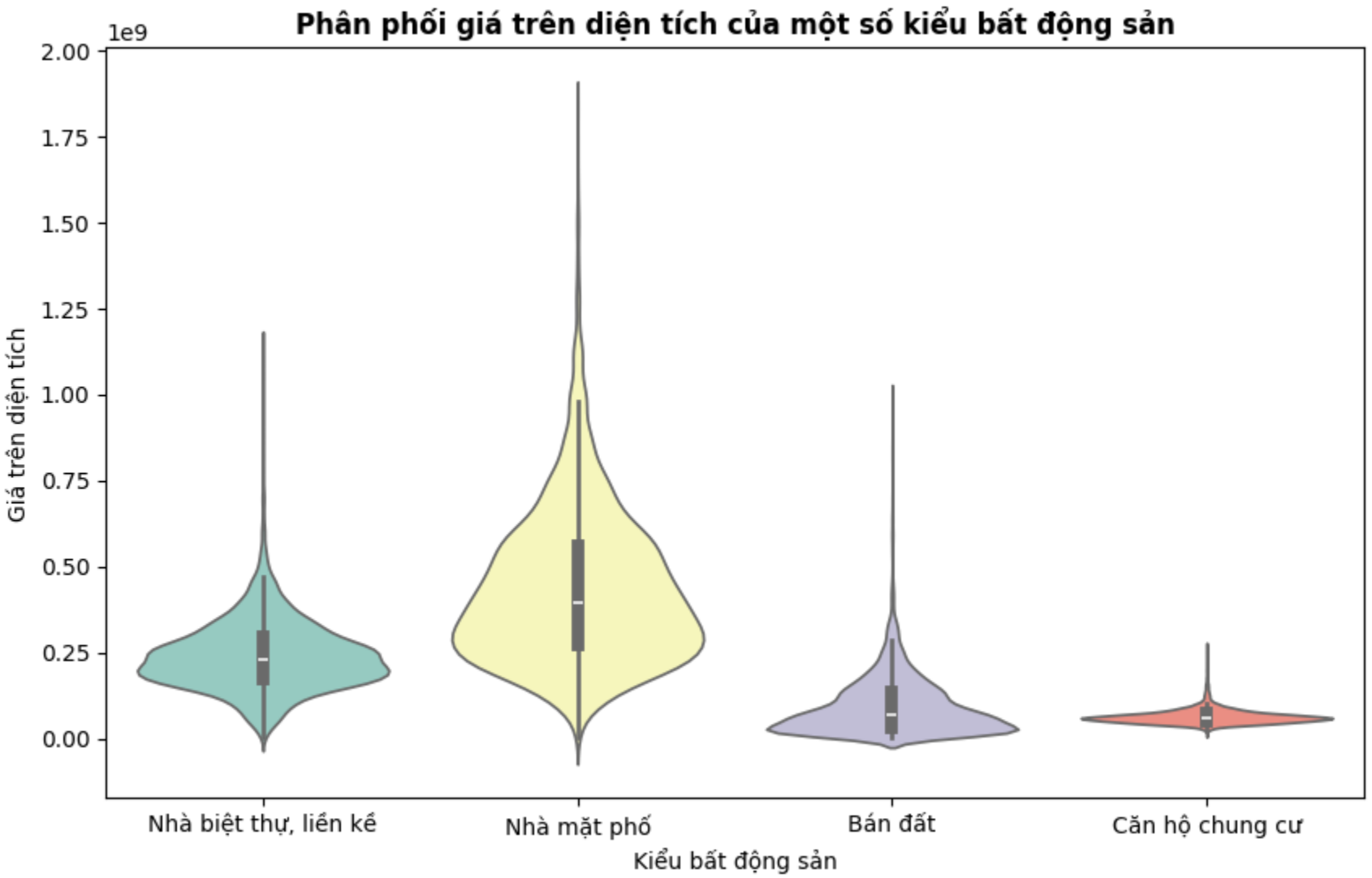
THU THẬP VÀ XỬ LÝ DỮ LIỆU



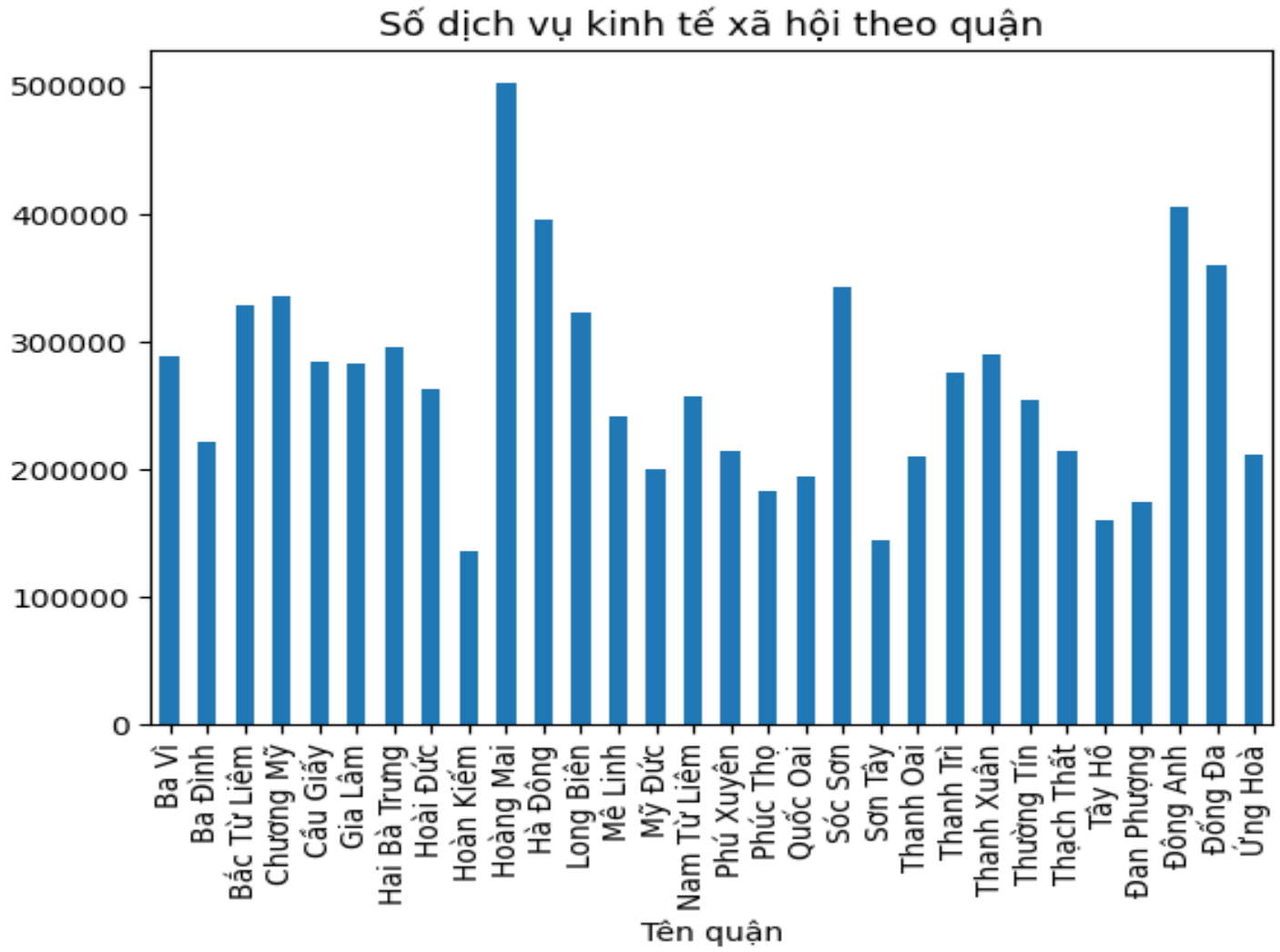
THU THẬP VÀ XỬ LÝ DỮ LIỆU



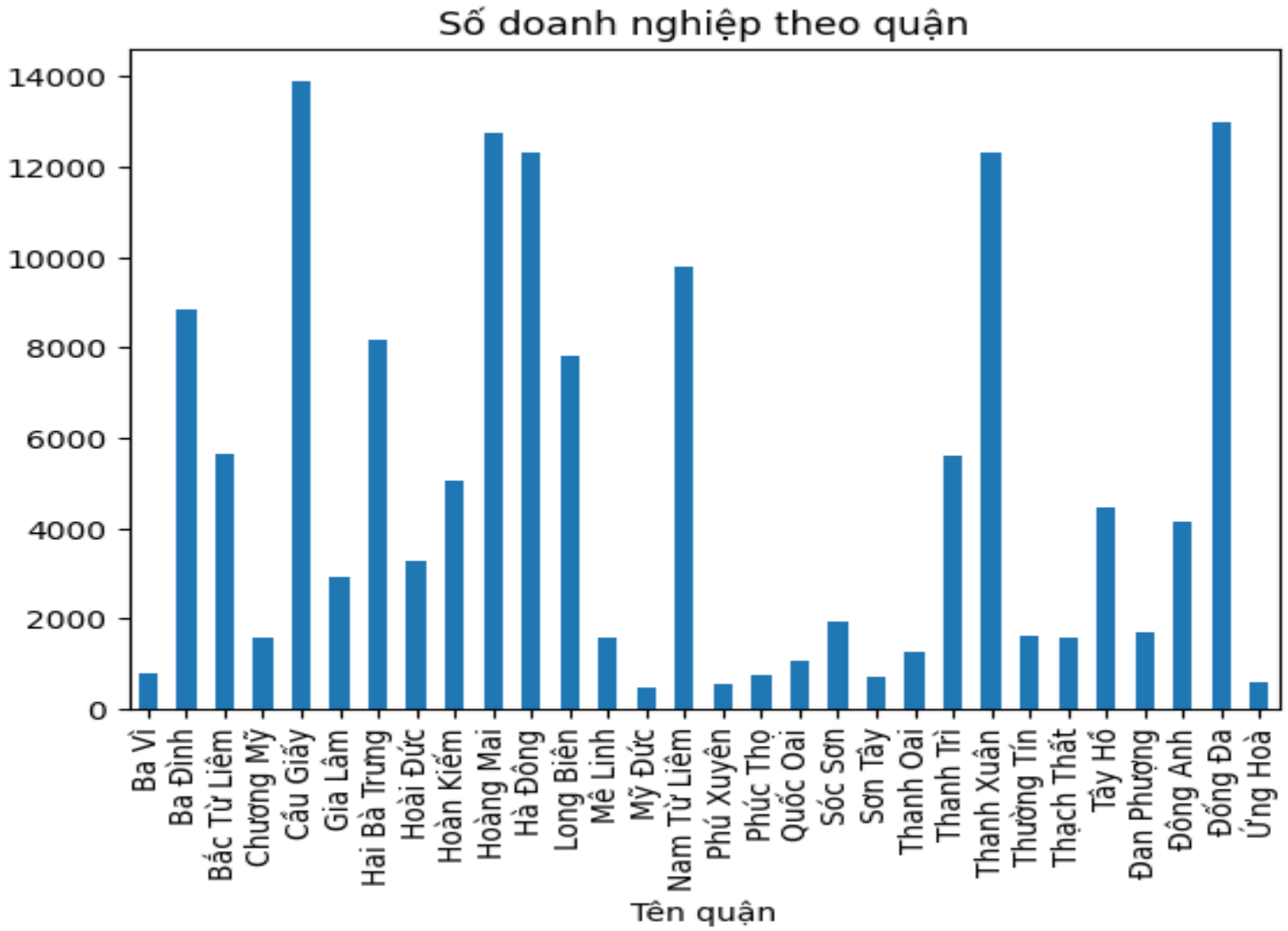
THU THẬP VÀ XỬ LÝ DỮ LIỆU



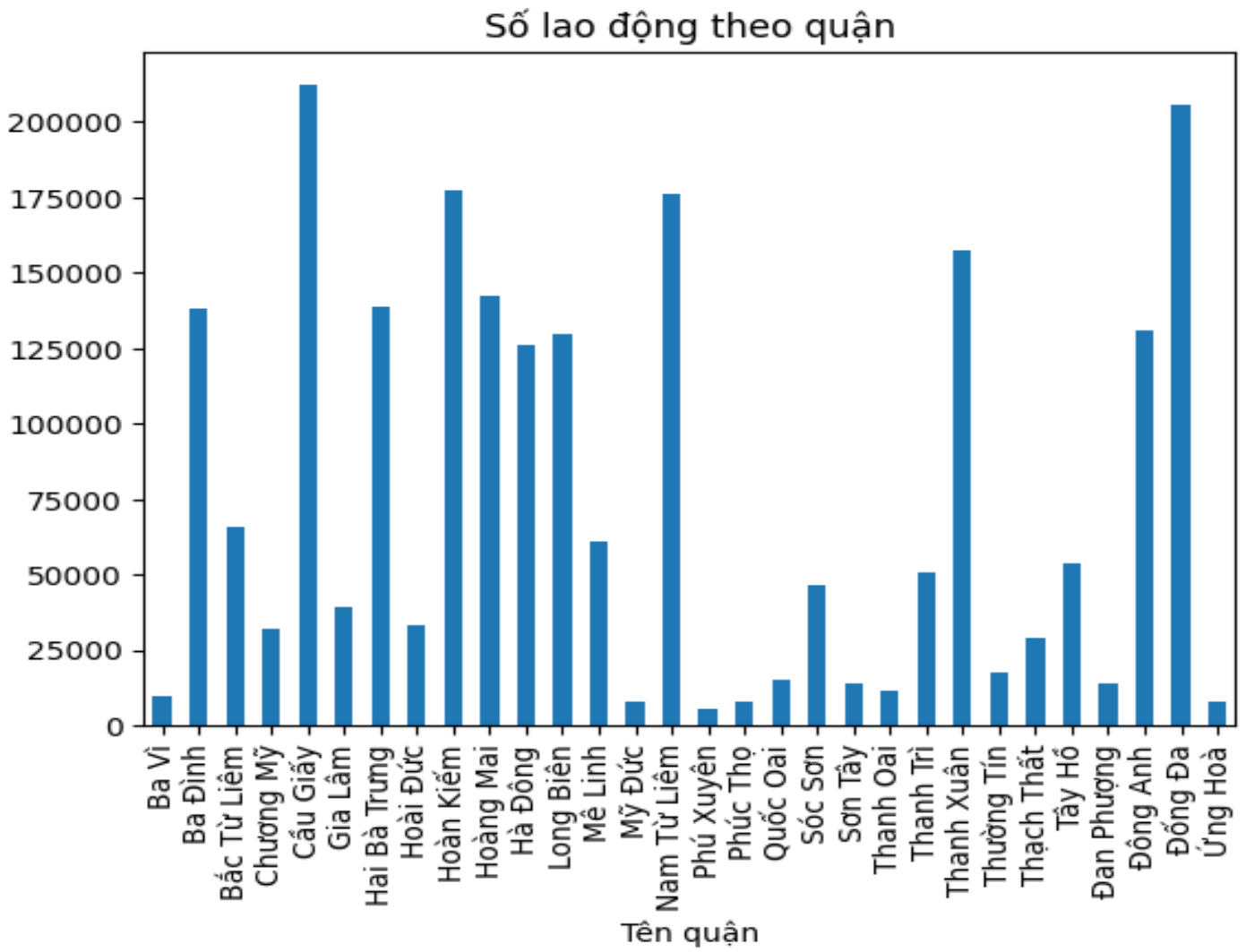
THU THẬP VÀ XỬ LÝ DỮ LIỆU



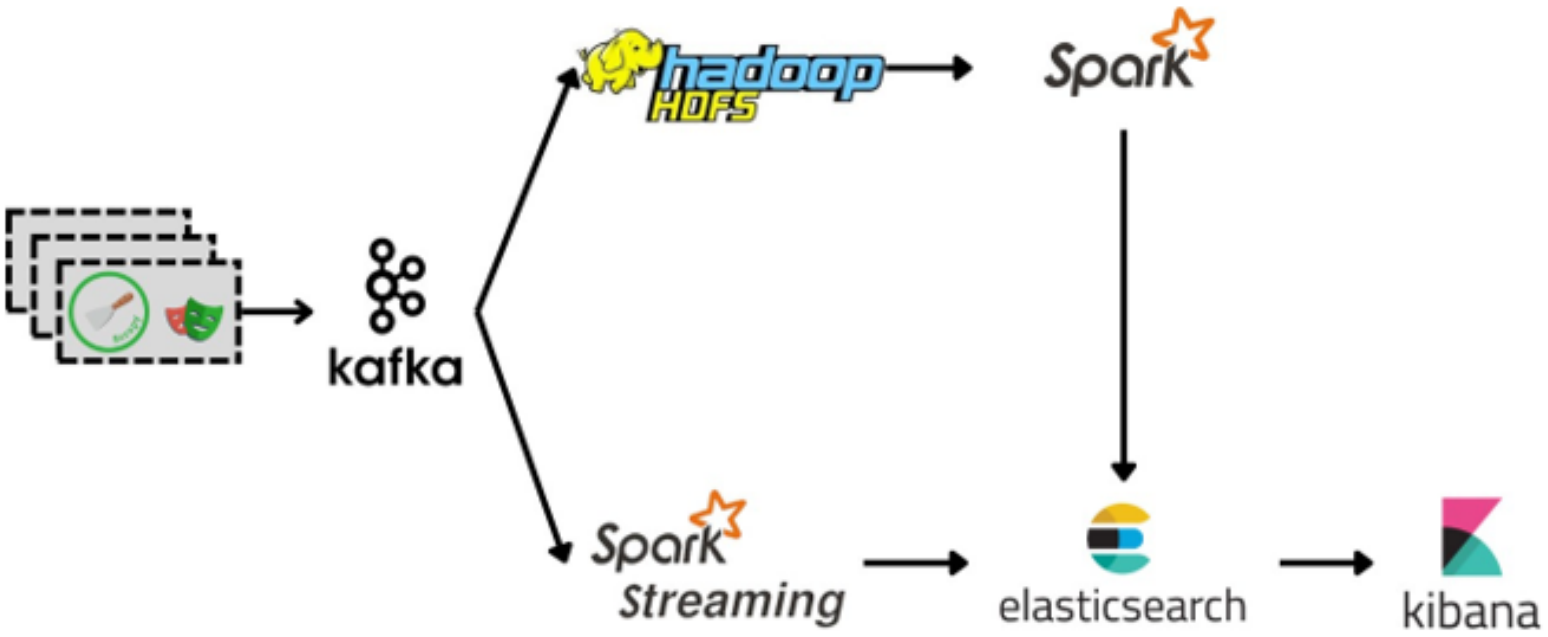
THU THẬP VÀ XỬ LÝ DỮ LIỆU



THU THẬP VÀ XỬ LÝ DỮ LIỆU

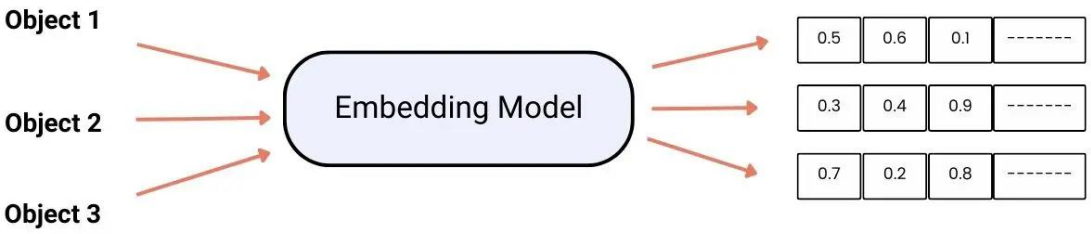


TRIỂN KHAI


hadoop
HDFS
Spark
kafka
Spark
Streaming
elasticsearch
kibana

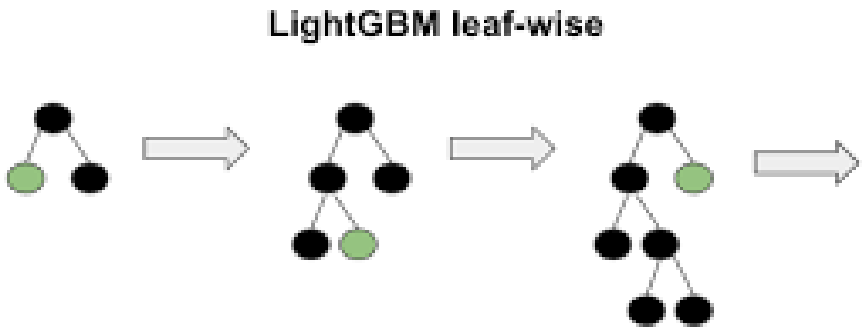
PHƯƠNG PHÁP THỰC NGHIỆM

- Mô hình Embedding: halong_embedding
- Đặc điểm:
  - Halong Embedding là mô hình embedding tiếng Việt được finetuned từ mô hình multilingual-e5-base.



PHƯƠNG PHÁP THỰC NGHIỆM

- Mô hình: LightGBM
- Đặc điểm:
  - LightGBM có khả năng xử lý lượng lớn dữ liệu và yêu cầu ít bộ nhớ hơn so với các mô hình tăng cường cây khác (như XGBoost).



MSE	R-Squared Error	MAPE
2.5319	0.9057	30.4639

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !