

Thu thập và phân tích dữ liệu bất động sản

Từ website batdongsan.com.vn sử dụng Selenium và MongoDB

Sinh viên thực hiện:

Phạm Trường Phát
Đinh Quốc Khánh
Lê Đình Nhật Thắng

Giảng viên hướng dẫn:

ThS. Lê Nhật Tùng

Trường Đại học Công nghệ TP. Hồ Chí Minh (HUTECH)
Khoa Công nghệ Thông tin – Khoa học dữ liệu

Năm học 2025–2026

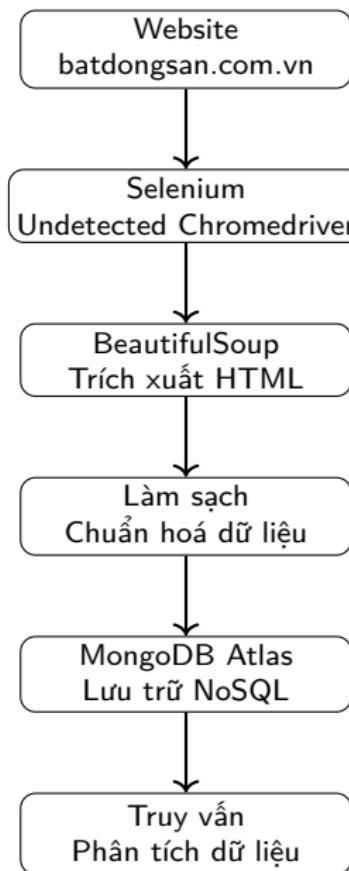
Lý do chọn đề tài

- Dữ liệu bất động sản lớn và biến động
- Thách thức về khai thác
- Sự phù hợp của công nghệ
- Giá trị thực tiễn

Mục tiêu nghiên cứu

- Xây dựng hệ thống thu thập dữ liệu tự động
- Trích xuất dữ liệu đa dạng
- Lưu trữ và quản lý hiệu quả
- Phục vụ phân tích thị trường

Quy trình hệ thống



Thu thập dữ liệu

- **Khởi tạo trình duyệt:** Selenium + Undetected Chromedriver để xử lý website tải động
- **Điều hướng trang:** Thu thập tin bán và cho thuê qua nhiều trang danh sách
- **Trích xuất thông tin:** BeautifulSoup phân tích HTML, lấy dữ liệu chi tiết
- **Kiểm soát trùng lặp:** Sử dụng post_id để tránh lưu trùng dữ liệu

Lưu trữ và truy vấn

- MongoDB Atlas (NoSQL)
- Cấu trúc dữ liệu linh hoạt
- Tối ưu truy vấn với Index
- Khai thác dữ liệu hiệu quả (EDA)

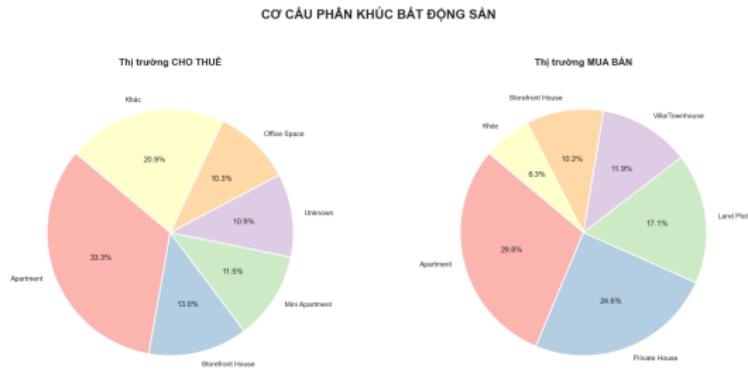
```
_id: ObjectId("604e1760179cd1f57a980ec")
post_id: "44743692"
property_url: "https://batdongsan.com.vn/ban-can-ho-chung-cu-phuong-an-phu-prj-master_"
transaction_type: "sale"
property_type: "Bán - Căn hộ"
title: "Quà Đặng Cấp Sđ hứ view triệu đô tại Masteri Park Place CTS - Sinh L_"
address: "Masteri Park Place, Phường An Phú, Quận 2, Hồ Chí Minh"
latitude: 10.8059150792159
longitude: 106.7713323275044
price: "5,9 tỷ"
price_per_sqm: "-103,51 triệu/m2"
area: "57 m2"
+ spec: Object
  bedroom: "2 phòng"
  bathroom: "1 phòng"
  legal: "Hợp đồng mua bán"
  furniture: "Đầy đủ"
  description: "Chính thức nhận booking 100 triệu/suất (có hoàn lại) - Khách booking s..."
+ images: Array (25)
+ project_info: Object
  date_posted: "22/12/2025"
  date_expired: "29/12/2025"
  news_type: "Tin VIP KTS Cường"
+ contact_info: Object
  name: "Ca Văn Sỹ"
  profile_link: "https://guru.batdongsan.com.vn/pa/cavansy?productType=3&cateId=324&pr..."
  avatar_url: "https://file4.batdongsan.com.vn/resize/200x200/2025/12/12/202512121745_..."
  phone_invisible: "+982 999 *** - Hiện số"
  zalo_url: "https://zalo.me/n82vRrIakkAv34CozaftdL5DC2lhz5d7uV63IkF1LcgsAFKaHJ"
  join_duration: "9 năm"
  listings: "113"
  verified: false
  scraped_at: "2025-12-23T19:14:29.874689"
```

Ví dụ cấu trúc JSON của một bài đăng bất động sản

Kết quả đạt được

- 32.794 tin đăng
- Hiệu suất: 5s/tin
- Tỷ lệ thành công: 98%
- **EDA:** Nhận diện cơ cấu phân khúc theo loại hình

Apartment chiếm tỷ trọng lớn ở cả thị trường thuê và bán.



Cơ cấu phân khúc bất động sản (thuê vs bán)

Kết luận và hướng phát triển

Kết luận

- Xây dựng thành công hệ thống tự động
- Khai thác dữ liệu chuyên sâu
- Lưu trữ hiện đại và nhất quán

Hướng phát triển

- Mở rộng đa luồng, proxy xoay vòng
- Mở rộng phạm vi dữ liệu
- Ứng dụng phân tích và dự đoán giá

Xin cảm ơn!

Chúng em xin chân thành cảm ơn
Quý Thầy/Cô cùng các bạn đã lắng nghe và góp ý

Nhóm sinh viên thực hiện
Phát – Khánh – Thắng