VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY

TRUONG DAI HOC KHOA HOC TU NHIEN
KHTN
TP. HO CHI MINH

Faculty of Information Technology

Môn học : Khoa Học Dữ Liệu
Lớp CQ2016/2 – Học kỳ I/2019-2020
GV: Trần Trung Kiên

BÁO CÁO LẦN I- ĐỒ ÁN CUỐI KỲ

<Tiến độ thực hiện đồ án tính đến ngày 02-12-2019>

Nhóm 14: *Đặng Phương Nam – Lê Minh Nghĩa*

NỘI DUNG CHÍNH

Câu hỏi đặt ra?

Thu thập dữ liệu

Cho các thông tin về một căn nhà
(đường, huyện/quận, mặt tiền, đường vào, diện tích, nội thất, số tầng,…)

Căn nhà này có giá là bao nhiêu tiền?

CÂU HỎI ĐẶT RA?

Người bán có thể dự đoán được giá trị căn nhà mà mình muốn bán.

Người mua có thể ước lượng được căn nhà mình muốn mua có giá cả hợp lý hay không?

Nguồn gốc câu hỏi: bắt nguồn từ việc mua, bán nhà ở diễn ra hằng ngày trong lĩnh vực bất động sản.

LỢI ÍCH CỦA CÂU HỎI

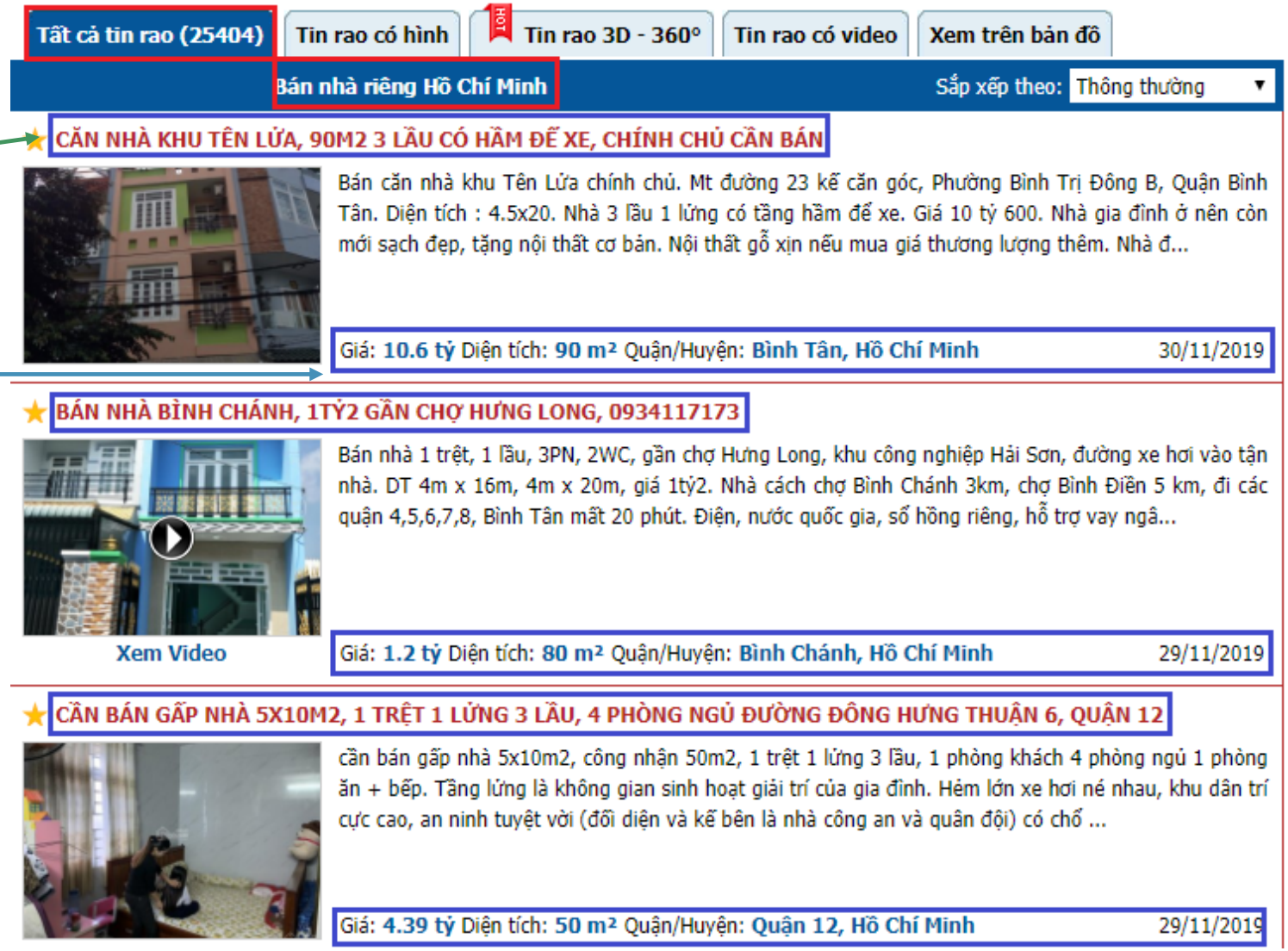
Trang web thu thập dữ liệu: https://batdongsan.com.vn/

THU THẬP DỮ LIỆU – PARSE HTML

Thu thập dữ liệu về các tin rao "Bán nhà riêng tại Tp.Hồ Chí Minh"

Parse HTML để lấy:

+ Title.
+ Detail url.
+ Price.
+ Area.
+ District.
+ Up time.



Ở mỗi lần parse tại một page của trang web nhóm đều thực kiểm tra việc lấy dữ liệu có hợp pháp hay không tại file robots.txt của trang web:

https://batdongsan.com.vn/robots.txt

Ở mỗi bản tin rao bán, nhóm muốn lấy thêm một số thông tin chi tiết:

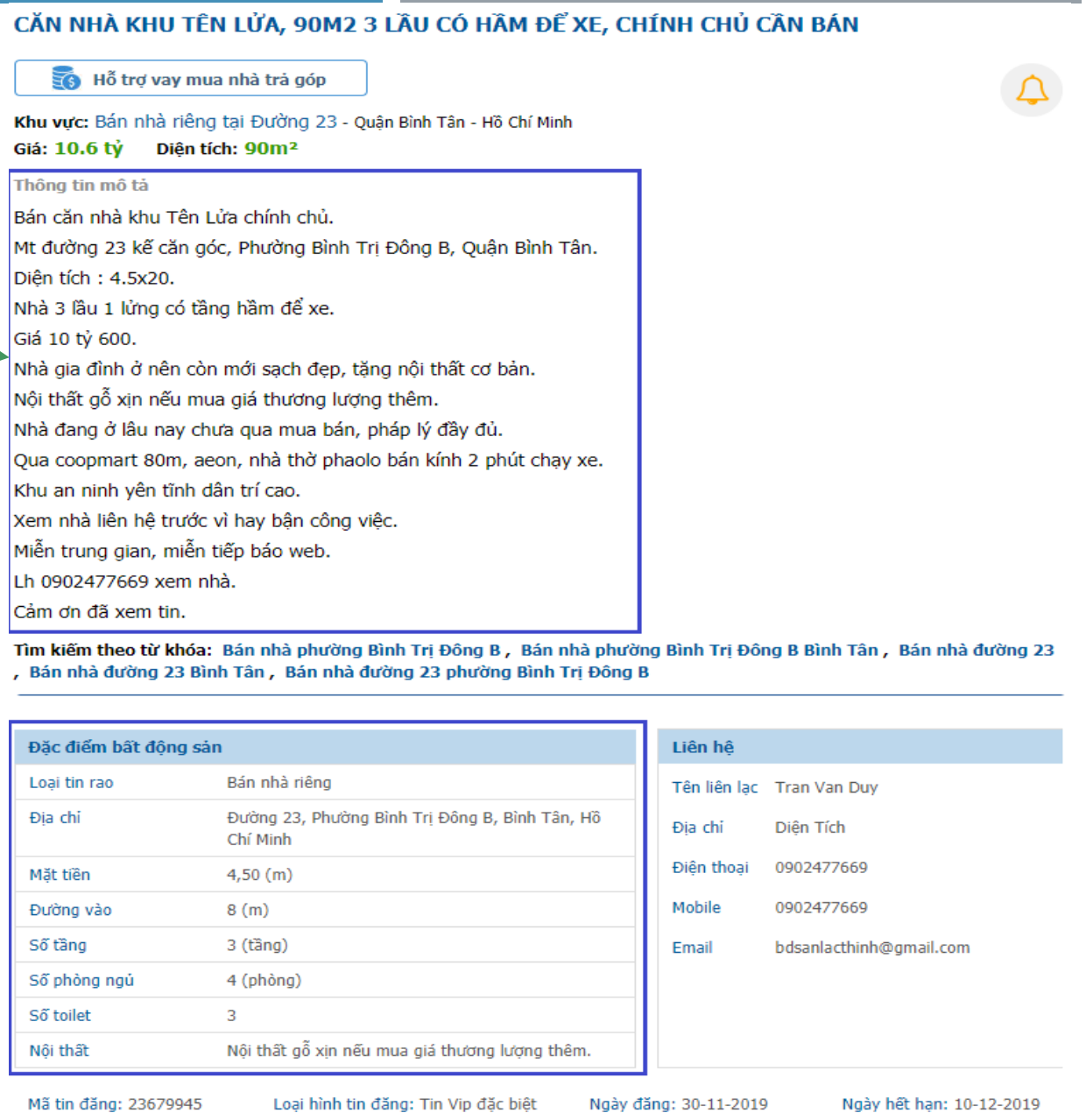
Parse HTML để lấy:

+ Decription.

+ Features:
  - Address.
  - Front.
  - Entrance.
  - ….

Ở mỗi lần parse tại bản tin rao bán của trang web nhóm đều thực kiểm tra việc lấy dữ liệu có hợp pháp hay không tại file robots.txt của trang web:

https://batdongsan.com.vn/robots.txt



CĂN NHÀ KHU TÊN LỬA, 90M2 3 LẦU CÓ HẦM ĐỂ XE, CHÍNH CHỦ CẦN BÁN

Hỗ trợ vay mua nhà trả góp

Khu vực: Bán nhà riêng tại Đường 23 - Quận Bình Tân - Hồ Chí Minh
Giá: 10.6 tỷ **Diện tích:** 90m²

Thông tin mô tả

Bán căn nhà khu Tên Lửa chính chủ.
Mt đường 23 kế căn góc, Phường Bình Trị Đông B, Quận Bình Tân.
Diện tích : 4.5x20.
Nhà 3 lầu 1 lửng có tầng hầm để xe.
Giá 10 tỷ 600.
Nhà gia đình ở nên còn mới sạch đẹp, tặng nội thất cơ bản.
Nội thất gỗ xịn nếu mua giá thương lượng thêm.
Nhà đang ở lâu nay chưa qua mua bán, pháp lý đầy đủ.
Qua coopmart 80m, aeon, nhà thờ phaolo bán kính 2 phút chạy xe.
Khu an ninh yên tĩnh dân trí cao.
Xem nhà liên hệ trước vì hay bận công việc.
Miễn trung gian, miễn tiếp báo web.
Lh 0902477669 xem nhà.
Cảm ơn đã xem tin.

Tìm kiếm theo từ khóa: Bán nhà phường Bình Trị Đông B , Bán nhà phường Bình Trị Đông B Bình Tân , Bán nhà đường 23 , Bán nhà đường 23 Bình Tân , Bán nhà đường 23 phường Bình Trị Đông B

Đặc điểm bất động sản	
Loại tin rao	Bán nhà riêng
Địa chỉ	Đường 23, Phường Bình Trị Đông B, Bình Tân, Hồ Chí Minh
Mặt tiền	4,50 (m)
Đường vào	8 (m)
Số tầng	3 (tầng)
Số phòng ngủ	4 (phòng)
Số toilet	3
Nội thất	Nội thất gỗ xịn nếu mua giá thương lượng thêm.

Liên hệ	
Tên liên lạc	Tran Van Duy
Địa chỉ	Diện Tích
Điện thoại	0902477669
Mobile	0902477669
Email	bdsanlacthinh@gmail.com

Mã tin đăng: 23679945 Loại hình tin đăng: Tin Vip đặc biệt Ngày đăng: 30-11-2019 Ngày hết hạn: 10-12-2019

```
$ wc -l *
  26220 2019-11-29_23_18_43.json
  19957 2019-11-29_23_18_43-removed-duplicate.json
  46177 total
```

Dữ liệu thu được từ trang web là 26220 bản tin, sau khi xóa các bản tin trùng nhau thì thu lại được 19957 bản tin.

Tiến hành lấy các thông tin cần quan tâm từ 19957 bản tin thu được dữ liệu có dạng như sau:

	area	district	address	floors	bedrooms	toilets	front	entrance	house_aspect	balcony_aspect	interior	near_center	owner	alley	villa	new	price
0	80.0	Bình Chánh	Đường Đoàn Nguyễn Tuấn, Xã Hưng Long	2.0	3.0	2.0	4.00	6.0	Đông	NaN	NaN	0	1	0	0	0	1200.0
1	50.0	Quận 12	Đường Đông Hưng Thuận 6, Phường Tân Hưng Thuận	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0	0	1	0	0	4390.0

THU THẬP DỮ LIỆU – LẤY DỮ LIỆU CẦN QUAN TÂM

1. are: diện tích nhà (m^2)
2. district: quận/huyện.
3. address: địa chỉ nhà.
4. floors: số lượng tầng.
5. bedroms: số lượng phòng ngủ.
6. toilets: số lượng toilet.
7. front: mặt tiền của nhà (m).
8. entrance: đường vào nhà (m).
9. house_aspect: hướng nhà.

Dữ liệu có 19443 dòng và 17 cột

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19443 entries, 0 to 19442
Data columns (total 17 columns):
area              18261 non-null float64
district          19443 non-null object
address           19073 non-null object
floors            12277 non-null float64
bedrooms          8828 non-null float64
toilets           7907 non-null float64
front             8257 non-null float64
entrance          10182 non-null float64
house_aspect      3511 non-null object
balcony_aspect    1604 non-null object
interior          3457 non-null float64
near_center       19443 non-null int64
owner             19443 non-null int64
alley             19443 non-null int64
villa             19443 non-null int64
new               19443 non-null int64
price             19443 non-null float64
dtypes: float64(8), int64(5), object(4)
memory usage: 2.5+ MB
```

Từ 19957 bản tin, trong quá lấy thông tin thì những bản tin không có "giá nhà" nhóm sẽ bỏ đi không lấy nên chỉ còn 19443 dòng.

10. balcony_aspect: hướng ban công.
11. interior: nội thất nhà (1/0).
12. near_center: gần trung tâm (1/0).
13. owner: chính chủ (1/0).
14. alley: hẻm (1/0)
15. villa: biệt thự (1/0).
16. new: nhà mới (1/0).
17. price: giá nhà (triệu đồng).

THU THẬP DỮ LIỆU – THÔNG TIN DỮ LIỆU LẤY ĐƯỢC

Các cột có giá trị thiếu là:
- are thiếu **1182** giá trị.
- address thiếu **470** giá trị
- floors thiếu **7166** giá trị
- bedrooms thiếu **10615** giá trị.
- toilets thiếu **11546** giá trị.
- entrance thiếu **9261** giá trị
- house_aspect thiếu **15932** giá trị.
- balcony_aspect thiếu **17839** giá trị.
- interior thiếu **15986** giá trị

Lý do có nhiều giá trị thiếu tại các cột bên là do người muốn bán nhà lúc đăng tin không nhập đầy đủ các trường thông tin trên mà thường chỉ mô tả chi tiết tại phần description của bản tin hay vì lý do nào đó mà bản thân người bán cũng không biết rõ các thông tin trên.

Giải pháp: Sẽ bỏ đi các cột thực sự không cần thiết mà có nhiều giá trị thiếu như bedrooms, toilets, house_aspect, balcony_aspect. Và phần giá trị thiếu còn lại, nhóm sẽ tiếp tục thử tìm trong phần description của mỗi bản tin rao bán từ phần dữ liệu đã lấy được ban đầu.

THU THẬP DỮ LIỆU – CÁC GIÁ TRỊ THIẾU

THANK YOU FOR WATCHING