Các task

Phần I là phần miêu tả dữ liệu. Trong phần II, mình sẽ viết tinh gọn các task mọi người sẽ thử sức mình. Phần III là phần mô tả chi tiết hơn các task. Phần IV là các tài liệu tham khảo có mục đích là giải thích ý nghĩa các con số trong các trường.
Trong tài liệu này, **'mẩu tin'**, **'bài đăng'** hay **'bài post'** là các khái niệm tương đương và được sử dụng luân phiên.

I. Miêu tả về dữ liệu

Dữ liệu các bạn nhận được là file data.json. Mỗi entry trong file này là một mẩu tin bất động sản được lấy từ các trang rao bất động sản của Vietnam. Mỗi entry sẽ có các trường và ý nghĩa của chúng như bảng sau:

Tên trường	Ý nghĩa
ID	ID của mẩu tin trong database <Có thể tạm thời không quan tâm>
content	Nội dung của mẩu tin rao <Đây là phần mọi người sẽ tập trung xử lý>
realestate_type	Loại bất động sản. VD: 1 --> đất, 2 --> nhà...
floor	Số tầng của bất động sản. Được dùng ở task số 2
attributes	Là một danh sách các các từ/cụm từ đã được đánh tag. Mọi người xem dữ liệu mẫu trong file **sample.json**

II. Tổng quan các task

1. Nhận diện và phân biệt các loại hình BĐS

Các bạn cần cập nhật lại thuộc tính **realestate_type** của từng bài đăng

Mỗi loại bất động sản sau khi được phân loại sẽ có một giá trị trị **realestate_type** mới.

1.1. Phân biệt nhóm đất

Các bài đăng về đất sẽ có trường **realestate_type = 1**.

Bởi vì có nhiều loại đất khác nhau như đất nông nghiệp, đất thổ cư, đất nền nên yêu cầu được đặt ra là phân loại các loại đất đó.

Hướng làm: Các bạn phân tích trường **content** của các bài đăng có **realestate_type = 1** để tìm các yếu tố đề cập tới bài đăng đang nói về đất thổ cư, đất nền hoặc đất nông nghiệp.

VD:

- Đất thổ cư: **realestate_type mới = 1**

> Tôi cần bán gấp lô **đất thổ cư** hẻm xe hơi đường 8 linh xuân thủ đức.Diện tích 632 m2 trong đó **thổ cư** 585 m2.Ngang 8m, nở hậu 20m.Đường rộng 5m thích hợp xây nhà trọ, xưởng, nhà trẻ...Giá bán 27 tỷ còn thương lượng. Liên hệ chính chủ để xem đất

- Đất nền: **realestate_type mới = 10**

> Đất Mặt tiền Kênh 8 Đường Vườn Thơm - Bình Lợi bình chánh Diện tích: 50m x 125 = 6300 mét Sổ đỏ **đất nông nghiệp** Cách Đường vườn thơm đúng 100m Cách Trần văn giàu khoảng 1.5km Giá : 2.3 triệu/m2Chỉ tiếp khách thực sự mua Liên hệ: A.Sang: 0909.103.008 0918.352.338

- Đất nông nghiệp: **realestate_type mới = 9**

> Đất Mặt tiền Kênh 8 Đường Vườn Thơm - Bình Lợi bình chánh Diện tích: 50m x 125 = 6300 mét Sổ đỏ **đất nông nghiệp** Cách Đường vườn thơm đúng 100m Cách Trần văn giàu khoảng 1.5km Giá : 2.3 triệu/m2Chỉ tiếp khách thực sự mua Liên hệ: A.Sang: 0909.103.008 0918.352.338

1.2. Phân biệt nhóm nhà trọ

Tương tự với nhóm đất, nhóm nhà trọ cũng sẽ được phân thành 2 loại:

- **nhà trọ - dãy nhà trọ - dãy trọ**: **realestate_type mới = 2**

> Chính chủ cần bán nà và **dãy trọ** trần xuân soạn p tân hưng q7. Dt 5,1*27 nở hậu L (công nhận 158m2). Thiết kết trệt 2lau. Tầng trệt nhà ở,tầng 2-3 có 6 căn phong trọ cho thuê thu nhập ổn định hơn 12tr/tháng. Nhà chính chủ hơn 20 năm. Giá 8,5ty tl. Lh 0916456451

- **phòng trọ**: **realestate_type mới = 8**

> Tôi cần cho thuê **phòng trọ** cao cấp trong khu Himlam Kênh Tẻ.
>
> Diện tích: 35-40m2, đầy đủ nội thất: giường nệm, tủ quần áo, ti vi, máy lạnh, máy giặt.
>
> Bao điện nước. giờ giấc tự do, có bếp riêng.
>
> Giá thuê: 6 triệu/ tháng
>
> Liên hệ: 0935.08.1685 chị Hương

1.3. Phân biệt nhóm chung cư - căn hộ - nhà riêng

Trong nhóm này chỉ cần phân biệt **căn hộ dịch vụ** mà thôi và để nó ra nhóm riêng.
Căn hộ dịch vụ có **realestate_type mới = 2**

> Cho thuê **căn hộ dịch vụ** gần Vincom Đồng Khởi, lầu 3, giá thuê 9 triệu/tháng.
> Tình trạng: Đang trống.
> Căn hộ được thiết kế hiện đại, nội thất đầy đủ: Sofa, giường, nệm, tủ lạnh, bếp..
> Khách thuê có thể dọn vào ở liền.
> Tọa lạc ngay vị trí trung tâm thích hợp cho khách muốn đi bộ tới chỗ làm giá thuê bao gồm: Internet, cab tv, dọn dẹp 2 lần/tuần; nước sinh hoạt: Điện: 4,000 vnd/kWh.
> Liên hệ: Mr Duong Thai 0916 881125.
> Email: thaiquang.duong@gmail.com

2. Phân biệt diện tích

Mỗi bài post sẽ có nhiều thuộc tính, và một trong số chúng là **area_cal**: **diện tích đất của bất động sản**. Mục tiêu của task này là xác định thuộc tính **area_cal**.

Có 3 diện tích có khả năng xuất hiện:

- **Diện tích đất:** Có thể được ghi dưới một trong các dạng sau:
  + 20m2
  + 5x4
  + 5 x 4
  + 5X4
  + ngang 4m x dài 5m

=> Đây chính là thuộc tính **area_cal**

- **Diện tích công nhận**: <Xem ví dụ ở dưới>

Thường thì **thuộc tính normal đứng trước thuộc tính area** sẽ có cụm **diện tích công nhận/DTCN/CN**

=> Nếu trong bài có xuất hiện cả **diện tích đất** và **diện tích công nhận** thì lấy giá trị **diện tích công nhận** gán cho **area_cal**

- **Diện tích sàn**: không phải một trong hai loại trên và thường có diện tích khá lớn

=> Nếu chỉ có diện tích diện tích sàn thì công thức tính **area_cal** từ diện tích sàn như sau:

area_cal = <diện tích sàn> / floor

Lưu ý: Trong trường hợp là BĐS là đất có diện tích hỗn hợp VD như có 1000m2 trong đó có 300m2 thổ cư thì sẽ lấy phần diện tích tổng là 1000m2

VD về diện tích công nhận:

Chính chủ cần bán lô đất!
* Vị trí:

- Liền kề Vườn Lài Villa với 3 mặt giáp sông Sài Gòn và sông Vàm Thuật, mang lại cho cả khu vực An Phú Đông không khí thoáng mát tự nhiên mà ít có khu vực khác ở TP Hồ Chí Minh có được.
- Cách Quốc lộ 1A 1km, dễ dàng kết nối với trung tâm và các vùng lân cận. Dự án có tiềm năng phát triển tốt, khả năng sinh lời cao.

- Khu dân cư đông đúc, an ninh, dân trí cao.

- Cách 5 phút di chuyển đến ĐH Văn Hiến, đại học Nguyễn Tất Thành, chợ An Phú Đông.

- Cách 15 phút đến E - Mart Gò Vấp, Coop Mart Phan Văn Trị, Vincom Plaza Gò Vấp, bệnh viện quốc tế

Hạnh Phúc.

- Sổ hồng riêng, sang tên công chứng nhanh gọn.

- Diện tích: 6,51x15m (lô góc 2 mặt tiền) chưa bao gồm vỉa hè. Tổng diện tích công nhận trên sổ 95,3m2.

- Hỗ trợ vay vốn ngân hàng từ 50 đến 70% giá trị giao dịch lô đất.

- Giá bán: 4,95 tỷ TL.

Liên hệ: 0906 6565 54 gặp anh Huy để được hỗ trợ xem đất.

3. Cải thiện và phân biệt giá bán

Đã hoàn thành các doc cần thiết cho team :)

4. Cải thiện vị trí

Mọi người đọc phần chi tiết ở dưới rồi sẽ có assign việc cụ thể cho từng bạn

III. Chi tiết các task

1. Nhân diện và phân biệt loại hình BĐS

- Hiện thời đã phân biệt được
  - Đất (realestate_type = 1)
  - Nhà/Nhà riêng (realestate_type = 2)
  - Căn hộ/chung cư/penhouse (realestate_type = 3)
  - Nhà xưởng (realestate_type = 5)
  - Nhà trọ (realestate_type = 7)
  - Khác (realestate_type = 6) : chưa phân loại
  - Dự án (realestate_type = 4)
- Tuy nhiên yêu cầu phân biệt thêm
  - Trong đất thì có 3 mục đất cần phân ra
    - **Đất hoặc đất thổ cư (realestate_type = 1):** đất thổ cư là đất bình thường được quy định để xây dựng nên nhà ở
    - **Đất nông nghiệp, đất trồng cây, đất lâm nghiệp, … (realestate_type = 9)**: đây là đất dùng để làm nông nghiệp, trồng cây.
    - **Đất nền, đất dự án (realestate_type = 10):** đây là dạng 1 miếng đất lớn có thể là đất nông nghiệp (được chuyển đổi mục đích sử dụng đất thành đất thổ cư) hoặc đất thổ cư được chủ dự án tách ra thành từng lô nhỏ và bán cho người mua đất từng lô. VD mỗi lô diện tích 80m2, 100m2, 120m2, …
  - Cách để phân biệt 3 loại đất trên
    - Mặc định sẽ là đất hoặc đất thổ cư hoặc có chữ ‘đất thổ cư’ sẽ là đất thổ cư, hoặc có chữ viết tắt ‘TC’ sẽ là đất thổ cư, hoặc có chữ ‘đất ở tại đô thị’ , ‘đất ở tại nông thôn’ là đất thổ cư. VD :

      Tôi cần bán gấp lô đất thổ cư hẻm xe hơi đường 8 linh xuân thủ đức.Diện tích 632 m2 trong đó thổ cư 585 m2.Ngang 8m, nở hậu 20m.Đường rộng 5m thích hợp xây nhà trọ, xưởng, nhà trẻ...Giá bán 27 tỷ còn thương lượng. Liên hệ chính chủ để xem đất

- Đất nông nghiệp, đất trồng cây, đất lâm nghiệp: thường các đất này có diện tích lớn trên 100m2 và có dấu hiện nhận biết ‘đất nông nghiệp', ‘đất trồng cây'. VD:

  Đất Mặt tiền Kênh 8 Đường Vườn Thơm - Bình Lợi bình chánh Diện tích: 50m x 125 = 6300 mét Sổ đỏ đất nông nghiệp Cách Đường vườn thơm đúng 100m Cách Trần văn giàu khoảng 1.5km Giá : 2.3 triệu/m2Chỉ tiếp khách thực sự mua Liên hệ: A.Sang: 0909.103.008 0918.352.338

- Đất nền, đất dự án: thường các đất này các dấu hiệu nhận biết sau: ‘đất nền’, ‘đất dự án', ‘phân lô', ‘nền đất’,

- Ngoài ra còn có một số dạng đất hỗn hợp bao gồm đất vừa thổ cư vừa có đất nông nghiệp: tạm thời đẩy loại đất này sang đất thổ cư

- Trong nhà trọ thì phân biệt ra 2 loại
  - **Nhà trọ hoặc dãy nhà trọ hoặc dãy trọ (realestate_type = 7)** : thường là BĐS dùng cho mục đích bán. Cách nhận biết: các cụm từ: ‘nhà trọ’, ‘dãy trọ', ‘dãy nhà trọ'. VD:

    *Chính chủ cần bán nà và dãy trọ trần xuân soạn p tân hưng q7. Dt 5,1*27 nở hậu L (công nhận 158m2). Thiết kết trệt 2lau. Tầng trệt nhà ở,tầng 2-3 có 6 căn phong trọ cho thuê thu nhập ổn định hơn 12tr/tháng. Nhà chính chủ hơn 20 năm. Giá 8,5ty tl. Lh 0916456451*

  - **Phòng trọ hoặc phòng cho thuê hoặc phòng trong căn hộ cho thuê (realestate_type = 8) :** thường là loại hình dùng cho mục đích thuê hoặc sang nhượng. Cách nhận biết: các cụm từ: ‘phòng trọ’, ‘phòng cho thuê', ‘ phòng trong căn hộ cho thuê'. VD:
    Tôi cần cho thuê phòng trọ cao cấp trong khu Himlam Kênh Tẻ.

    Diện tích: 35-40m2, đầy đủ nội thất: giường nệm, tủ quần áo, ti vi, máy lạnh, máy giặt.

Bao điện nước. giờ giấc tự do, có bếp riêng.

Giá thuê: 6 triệu/ tháng

Liên hệ: 0935.08.1685 chị Hương

- Trong chung cư, căn hộ, nhà riêng thì phân biệt ra 2 loại:
    - **Chung cư, căn hộ, penhouse (realestate_type = 3)**
    - **Căn hộ dịch vụ** : giấu hiệu nhận biết là có cụm từ ‘căn hộ dịch vụ' hoặc ‘CHDV’, và cần được đưa vào mục nhà riêng **(realestate_type = 2)** thay vì mục **Chung cư, căn hộ, penhouse (realestate_type = 3)** như hiện tại

2. Nhận diện và phân biệt diện tích trong 1 bài post BĐS

- Diện tích khi nhận về từ tag service có thể bị sai hoặc ra một số quá lớn. Do đó cần phân biệt 3 loại diện tích sau
    - Diện tích đất: là phần diện tích đất toàn bộ, mặc định là diện tích của bài post (trường area_cal). Diện tích này có được bằng 2 cách : VD như bài post ghi sẵn diện tích (VD: diện tích 20m2) hoặc ghi theo cách 5x4 , 5 x 4, 5X4 và suy đoán ra 20m2. Đôi khi diện tích cũng được ghi như thế này: ngang 2,6m x dài 5,08m
    - Diện tích công nhận (DTCN): Đây là diện tích đất được phép sở hữu. Và trong trường hợp xuất hiện 2 diện tích (diện tích /diện tích và diện tích công nhận/DTCN/ CN thì chọn diện tích công nhận. VD: diện tích đất là 90m2 diện tích đất công nhận là 85m2 thì mặc định chọn là 85m2
    - Diện tích xây dựng, diện tích sàn: Đây không phải là là diện tích của bài post (trường area_cal). Đây là diện tích Trong trường hợp không có 2 loại diện tích kia thì có thể xác định area_cal bằng cách sau: lấy số diện tích này chia cho số tầng lầu
- Trong trường hợp là BĐS là đất có diện tích hỗn hợp VD như có 1000m2 trong đó có 300m2 thổ cư thì sẽ lấy phần diện tích tổng là 1000m2

3. Cải thiện và phân biệt giá bán, giá thuê và bài post là thuê hay là bán

- Nhận lại document từ Huy Vũ về nhận diện bài post là bán hay thuê: **https://docs.google.com/document/d/1-U5Romj24SGgSHgSICZ5Bz1vcSHGXytU7FWLWqmRBIA/edit?usp=sharing**
    - Bán (transaction_type = 1)
    - Thuê (transaction_type = 2)
    - Sang nhượng (transaction_type = 3)
    - Vừa bán vừa thuê (transaction_type = 4): chủ BĐS muốn đăng bán hoặc cho thuê BĐS. Sẽ xuất hiện cả 2 giá bán và giá thuê của BĐS
    - Bán - đang cho thuê và có giá thuê (transaction_type = 5): chủ BĐS muốn đăng bán, BĐS này đang có sẵn hợp đồng thuê và tạo ra thu nhập cho người sở hữu. Sẽ xuất hiện cả 2 giá bán và giá BĐS này đang được cho thuê
    - Bán - đang cho thuê và có giá thuê (transaction_type = 6): tương tự như (transaction_type = 5) nhưng người đăng ko nói rõ giá đang cho thuê là bao nhiêu nên chỉ có giá bán
    - Khác (transaction_type = 7): Không phân loại được các mục trên thì đẩy vào đây, con người sẽ xử lý phân loại
- Nhận lại document từ Huy Vũ về nhận diện giá bán và giá thuê của bài post
    - price_sell
    - price_rent

- Cải thiện việc nhận diện bán và thuê và giá bán và giá thuê
    - Hiện tại giá thuê BĐS bị ràng buộc ở khoảng giá rent_range = [299999, 500000000]. Trong khi ở thực tế giá thuê tuỳ vào từng loại hình BĐS có mức giá thuê khác nhau
        - Nhà/Nhà riêng: [299999, 10,000,000,000]
        - Chung cư/Căn hộ: [299999, 100,000,000]
        - Phòng trọ dãy trọ: [199999, 50,000,000]
        - Nhà xưởng: [299999, 10,000,000,000]
        - Đất: [299999, 10,000,000,000]
    - Bị dính số điên thoại thành giá bán hay giá thuê
    - Một bài post có thể xuất hiện nhiều loại giá khác nhau không chỉ đơn giản là một giá bán, một giá thuê bởi vì có thể xuất hiện như thế này:

    - Giá thị trường xung quanh đang bán là 200tr/m2, nhưng nay bán giá 6 tỷ (180tr/m2). BĐS đang có hợp đồng thuê là 19 triệu/tháng
    - Bán lô A 6x15m2 giá 2tỷ5 . Lô B 6x20 giá 3tỷ
- Nhận diện rõ hơn trường hợp transaction_type = 5:
    - Có hiện thị các cụm từ HĐT , hợp đồng thuê, đang cho thuê
    - Và có rõ 2 mức giá bán và giá thuê hợp lý: Công thức để tính ra giá thuê và giá bán có lý:
        - Giá thuê * 12 *100 / Giá bán <= 50 là hợp lý
        - VD: giá thuê 50 triệu, giá bán 500 triệu
            50 *12*100/ 500 = 120 là ko hợp lý

- Nhận diện lại sang nhượng

4. Cải thiện và nhận diện được vị trí của BĐS (position_street)

Có tổng cộng 5 vị trí của BĐS như sau:

- Mặt tiền: description hoặc content có chứa chữ mặt tiền, hoặc MT, hoặc có địa chỉ cụ thể mà ko có / (Ví dụ: nhà 268 Lý thường kiệt,...) thì position_street = 1. Nhớ loại trừ trường hợp 2

- Hai mặt tiền trở lên: description hoặc content có chứa chữ hai mặt tiền, 2 mặt tiền, 3 mặt tiền, ba mặt tiền, hoặc 2 MT, 2MT, 3MT, 3mt ,... thì position_street = 2. Nhớ loại trừ trường hợp 1

- Hẻm/Hẻm 1 sẹc: description hoặc content có chứa chữ hẻm 1 sẹc, hẻm, hoặc có địa chỉ cụ thể mà có 1 cái sẹc VD: 181/9 3 tháng 2. Các trường hợp khác ko xác định được thì mặc định đưa vào đây. position_street = 3

- Hẻm 2 sẹc trở lên: description hoặc content có chứa chữ hẻm 2 sẹc, hẻm 3 sẹc, hẻm, hoặc có địa chỉ cụ thể mà có từ 2 cái sẹc VD: 181/9/36 3 tháng 2. position_street = 4

- Hai mặt tiền hẻm/ Hai mặt hẻm: description hoặc content có chứa chữ hẻm 2 MT hẻm, mặt tiền hẻm, 2 mặt hẻm, 2 mt hẻm. position_street = 4

IV. Tài liệu tham khảo

Các tài liệu bao gồm:

1. Các thông tin cần chuẩn hóa:

https://docs.google.com/document/d/1JuCQhjsQ2rOKvyfGkxlzqN61ROq6c7ZKS6V_NFAnAwE/edit

Đây là file gốc bên công ty gửi qua, 500 anh em đọc thôi chứ đừng chỉnh sửa nhá há há :D