# Ý tổng quan

## Problems:

Có **2 lỗi chính** của bảng excel: địa chỉ đường bị rối và không consistent font để mapping. Vì mapping cần biến thống nhất để phân biệt các bệnh viện khác nhau, mình cần thống nhất key street address và tên bệnh viện. Nhưng đồng thời, các sheet lại không thống nhất một format chung, nên key address của anh Kiên sẽ không ra kêt quả đồng đều

## Hướng đề xuất:

Hỗ trợ key street address đã có sẵn của anh Kiên Mình sẽ Clean Data cho từng excel sheet trước sử dụng Automation trong excel, và áp dụng web scraping với Python để lấy một format chung của các địa chỉ mình cho là có “Vấn đề”. Vì áp dụng cả web scraping, quá trình sẽ mất thời gian. Assumption của em: vì hỗ trợ bệnh viện và chi nhánh mình sẽ dùng lâu dài, và quá trình để thêm liên kết bảo hiểm với set bệnh viện+địa chỉ mới sẽ mất thời gian, nên mình sẽ có thời gian để automate quá trình sửa lại và clean địa chỉ.

# Chi tiết quá trình:

### Sửa từng bảng chỗ lỗi Error

(có đính kèm hướng dẫn ở sheet guidelines trong file excel)

##### Viết lại địa chỉ thành bộ phận nhỏ với dấu phẩy:

Thêm dấu phẩy trước các từ phố, phường, quận, huyện,….

**Function sử dụng**:

= SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(

SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(

SUBSTITUTE(SUBSTITUTE(

TRIM (LOWER([ĐỊA CHỈ GỐC])),

" phường", ", phường"), " quận Q.", ", quận Q."),

" P.", ", P."), " p.", ", p."), " rd.", ", rd."), " rd", ", rd"),

" phố", ", phố"), " Phố", ", Phố"),

" tp.", ", tp.")," tp", ", tp")

**Em khuyên** mình không nên sử dụng VBA macro để tách các chữ trên. VBA không nhận dạng tiếng Việt tốt, và các program unikey cho VBA cũng không hoạt động tốt. Một cách khác là bỏ dấu tiếng Việt. Cách bỏ dấu tiếng Việt là áp dụng một macro khác:

[VBA bỏ dấu TV](https://docs.google.com/document/d/1tkwX_l1gis89O3uNJp3UzDgQRp72_lORXH8lrGoyqzE/edit)

Tuy nhiên, **Em Khuyên** mình cẩn trọng bước này. Sẽ có những tên đường hay phường có tên giống nhau hoặc gần giống nhau (Phường Tân Quy với Tân Quý). Nếu mình kết hợp với tên bệnh viện thì vấn đề bớt lo ngại hơn, nhưng vẫn nên cẩn trọng nếu mình sử dụng địa chỉ cho analytics.

##### Phân tách địa chỉ thành bộ phận nhỏ

Tách địa chỉ thành các bộ phận nhỏ để dễ xử lý, sử dụng **Delimiter.** Delimitor sẽ tách cột địa chỉ thành các cột với các thành phần nhỏ hơn của địa chỉ. **Delimiter** có thể phân biệt 2 dấu phẩy liền kề xuất hiện ở bước trên.

**Future work** của project sẽ là hỗ trộ key address của anh Kiên. Cột được tách đầu tiên thường là cột về số nhà với tên đường, vốn là key info của bệnh viện. Đây là cột mình sẽ chú tâm nhiều nhất. Chủ yếu mình sẽ clean data ở cột này khi kết hợp tên bệnh viện. Các cột sau được tách ra nhằm để reference thêm.

##### Detect những địa chỉ “đặc biệt”

Địa chỉ đặt biệt là những đia chỉ tách ra không được 3 value, và được biểu hiện nếu số value bé hơn 3 hoặc lớn hơn 3.

**Giải thích**:

Nếu địa chỉ tách ra được mỗi hai phần thì địa chỉ hoặc ko phân dấu, hoặc ko rõ thành thị (như “...hai bà trưng, hà nội”).

Với địa chỉ tách ra được thành 3 giá trị, khả năng cao số nhà đã được tách khỏi tên đường (ví dụ “số 88, đường cao thắng, phường gì đó, tp. gì đó”).

Tạo một cột mới, cách cột cũ khoảng **6 cột hơn** nhằm chừa chỗ cho delimiter để lưu những địa chỉ đặt biệt này. Cột này chỉ hiển thị những địa chỉ đặt biệt mà không hiển thị những địa chỉ đã “ok”. **Function sử dụng**:

= IF(COUNTA ([RANGE CỘT SAU KHI DELIMITED]

) <> 3, [Ô ĐỊA CHỈ CŨ] , "")

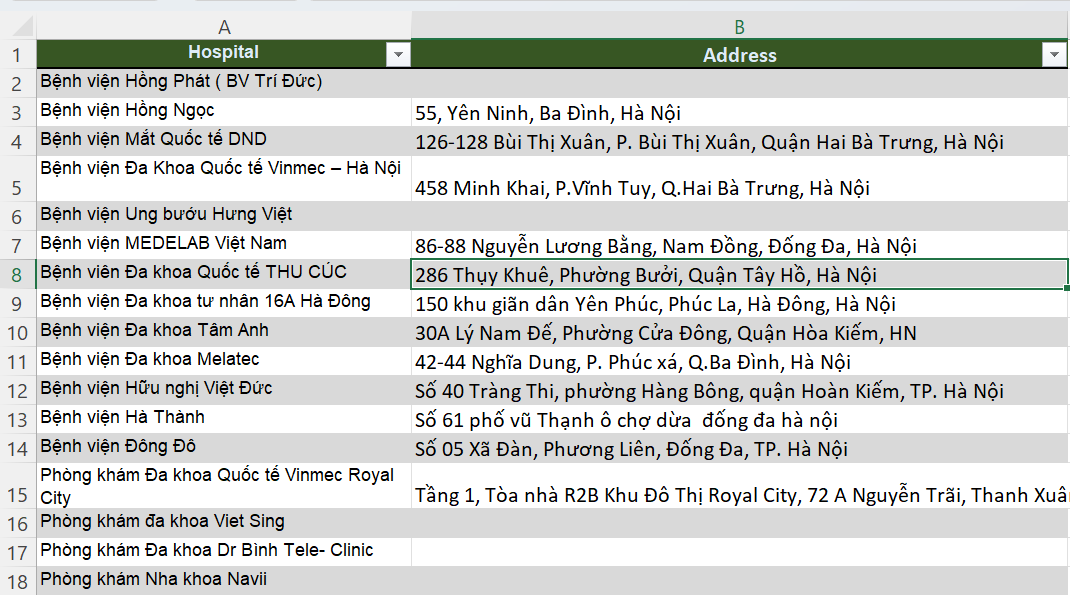
##### Bỏ các địa chỉ đặt biệt vào program Selenium để webscrape Google Map

Vì Google Map sẽ cho một format chung cho tất cả địa chỉ, hoặc thể hiện nếu địa chỉ không xuất hiện được trên Google Maps. Nếu trường hợp này xảy ra, mình sẽ chọn địa chỉ gần nhất hoặc relevance nhất.

Về **accuracy của program**. Chia program thành 3 lần xét lại quá trình. Hiện tại, **cứ 30 địa chỉ** thì có **7 đến 9 địa chỉ** không được nhận ra sau round thứ 2. Đến cuối chương trình, **cứ 30 địa chỉ** thì **9 đến 10 địa chỉ** sẽ không được nhận ra. Program **không nhận diện được các địa chỉ nước ngoài.**

**Để chuẩn bị cho program:**

1. Tải Chrome Web driver: <https://chromedriver.chromium.org/downloads>. Sau bước này, kiểm tra program cho phép webdriver chạy hay không.
2. Instal packages (listed and included in the program). Các package chính cho program: BeautifulSoup, Webdriver
3. File excel lưu dòng các địa chỉ cần chú ý thêm. Cấu trúc của file có dạng:



***(Không cần format table nhìn y chang, em format cho đẹp với rõ)***

Program chia thành 3 round chính:

**First round**: để lấy những địa chỉ xuất hiện rõ ràng trên Google Map, với thời gian chờ là 2 giây mỗi địa chỉ. Mất 20 phút cho 300 address lines.

**2nd round** chạy lại địa chỉ không tìm được từ lần chạy thứ nhất với thời gian chờ page load lâu hơn (4 giây mỗi địa chỉ). Mất 6 phút + 23 phút cho 150 address lines.

**3rd round** lấy địa chỉ gần nhất với search result nếu vẫn không lấy được địa chỉ. Mất 20 phút cho 150 address lines.

Program có lưu các biến với Pickle. **Thời gian chạy tổng cộng**: ~1h 15 mins.

*Nếu anh có chạy demo, anh chỉ cần chạy giúp em phần load của pickle thôi nhé! Em sẽ bỏ cả file pickle vào bên file em gửi.*