### Từ email hôm 2/9

Vấn đề hiện tại, là địa chỉ lat/long của google với số "123 Tỗ Hữu" không đúng lắm.

Câu hỏi là, mình có thể trả về lat/long chính xác hơn google không?

Nếu có, đây là đồ án mang tính ứng dụng cao, và mang tính lý thuyết cao (machine learning on categorical/sequential data)

Và đây là đồ án nằm trong tầm tay của em, hoàn thành trong một nốt nhạc.

Em xem nhé

---------- Forwarded message ----------  
From: **Henry Tu** <[tutrunghieu@gmail.com](mailto:tutrunghieu@gmail.com)>  
Date: Fri, Sep 2, 2016 at 7:15 PM  
Subject: averaged location: a machine learning method to correct Google lat/long from the address  
To: TU TRUNG HIEU <[tutrunghieu@gmail.com](mailto:tutrunghieu@gmail.com)>

+ the problem with existing location search (we don't have the right location)

+ we hypothesize that, by averaging the nearest locations, we hope to have better location for each address.

+ how to do that, what are similar locations?

+ how to combine the lat/long of similar locations? can we have averaging?

-- This is what we will address in this address

------ This is the regression problem of machine learning, to build f:String -> (L, L)

-------------- we will solve the problem with kNN linear regression

------- we answer: what is the center of the city?

------------ Framing/casting

address 1 >> LL1

address 2 >> LL2

address 3 >> LL3

address 4 >> LL4

----------- then

address X >> LL???

### Bài toán

112 – Trần Duy Hưng >> (21.010648, 105.799417)

114 – Trần Duy Hưng >> (21.010689, 105.799317)

113 – Trần Duy Hưng >> (?, ?)

### Cách nhìn để giải quyết

+ Cách giải quyết nội suy tuyến tính

X1 -> y1

X2 -> y2

0.5\*x1 + 0.5\*x2 >> ?

### Liệu nội suy tuyến tính có giải quyết được không?

+ giả định là đa số các đường đều là thẳng, do đó nội suy tuyến tính có thể hoạt động không có lỗi

+ với các khúc cong, vẫn có đủ dữ liêu, các phương pháp kNN hoặc nội suy tuyến tính sẽ hoạt động tốt (giả đinh).

+ với các khúc cong, không đủ dữ liêu, sẽ có sai số



### Dữ liệu thử nghiệm (task 1, tải dữ liệu về)

+ tạo ra một bảng có dạng

|  |  |  |
| --- | --- | --- |
| địa chỉ 1 | lat 1 | long 1 |
| địa chỉ 1 | lat 1 | long 1 |
| địa chỉ 1 | lat 1 | long 1 |
|  |  |  |

### Thuật toán (task 2, cài đặt thuật toán)

+ người dùng nhập “113 – Trần Duy Hưng”

+ tìm trong CSDL để lấy được các địa chỉ gần nhất (“113 – Trần Duy Hưng”, Lat1, long1) và (“117 – Trần Duy Hưng”, lat2, long2)

+ kết hợp lại bằng trung bình cộng

Out\_lat = avg(lat1, lat2, .., latk)

Out\_long = avg(llong1, long2, .., longk)

### Kiểm tra thuật toán

+ lấy 1 điểm có sẵn trên bản đồ, biết rõ lat/long rồi,

+ đưa điểm đó làm đầu vào thuật toán, rồi lấy đầu ra và so sánh với đầu ra đã biết (ground-truth)

Ví dụ (212 Trần Duy Hưng) phải cho ra (21.007013, 105.794361)

### Không phải nội suy tuyến tính

+ Đường trái / phải sẽ có số chẵn lẻ khác nhau

+ Đường là chữ, không phải số, không sắp xếp trình tự và nội suy được. Nếu là số

lat(0.3\*x1 + 0.7\*x2) = 0.3\*lat(x1) + 0.7\*lat(x2) giả định lat là tuyến tính

### Có thể suy luận đươc số còn lại của bên phải đường, nếu biết bên trái đường

+ nếu biết rất nhiều điểm ở bên trái đường. lại chỉ biết vài điểm bên phải đường, có thể suy ra các điểm còn lại bên phải đường không?

