Proposal for Data Mining Project

Group - 6

2022 / 04 / 08

## Phần 1 - Giới thiệu

Tất cả chúng ta đều đã bị kẹt xe ở một cột đèn giao thông, chỉ được dành vài giây để đi qua một ngã tư, phía sau một đoàn người đi khác. Đó là vấn đề cần phải giải quyết, để có thể giúp các nhà lập kế hoạch thành phố và chính phủ dự đoán trước các điểm nóng về giao thông và giảm bớt căng thẳng khi dừng và đi của hàng triệu người khi tham gia giao thông.

Dựa vào môn học Khai Thác Dữ Liệu để phân tích và trả lời các câu hỏi:

1. Dùng random forest xem bài toán có khả thi không?
2. Những siêu tham số nào quan trọng đối với bài toán kẹt xe?
3. Thử áp dụng các thuật toán phân cụm(k-mean, Hierarchical, dbscan) vào bài toán thử xem nó có cải thiện được độ chính xác không?

Nguồn lấy từ Kaggle.

Link: https:[//www.kaggle.com/competitions/bigquery-geotab-intersection-congestion/data](http://www.kaggle.com/competitions/bigquery-geotab-intersection-congestion/data)

Bộ dữ liệu cho cuộc thi này bao gồm thông tin tổng hợp về các phương tiện đã dừng và thời gian chờ ở giao lộ. Nhiệm vụ trong bài là dự đoán tình trạng tắc nghẽn, dựa trên thước đo tổng hợp về khoảng cách dừng xe và thời gian chờ, tại các giao lộ ở 4 thành phố lớn của Hoa Kỳ: Atlanta, Boston, Chicago & Philadelphia.

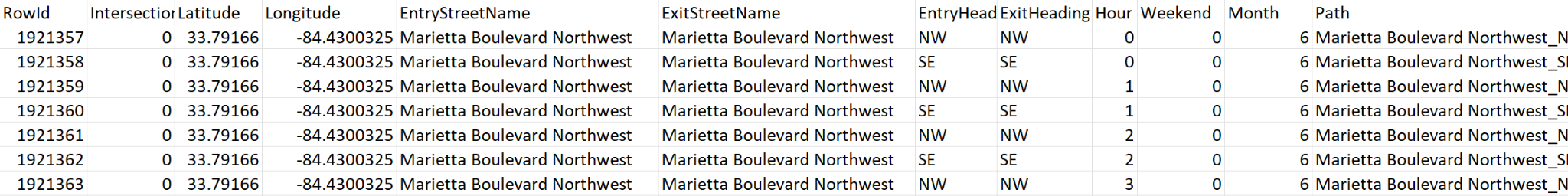
## Phần 2 - Dữ liệu

Giải thích các biến:

* RowId ID hàng
* IntersectionId ID Giao lộ
* Latitude Vĩ độ
* Longitude Kinh độ
* EntryStreetName Tên đường phố nhập
* ExitStreetName Tên đường phố thoát
* EntryHeading Tiêu đề mục nhập
* ExitHeading Thoát tiêu đề
* Hour Giờ
* Weekend Ngày cuối tuần
* Month Tháng
* Path Đường dẫn
* TotalTimeStopped\_p20 Tổng thời gian đã dừng\_p20
* TotalTimeStopped\_p40 Tổng thời gian đã dừng\_p40
* TotalTimeStopped\_p50 Tổng thời gian đã dừng\_p50
* TotalTimeStopped\_p60 Tổng thời gian đã dừng\_p60
* TotalTimeStopped\_p80 Tổng thời gian đã dừng\_p80
* TimeFromFirstStop\_p20 Thời gian từ điểm dừng đầu tiên\_p20
* TimeFromFirstStop\_p40 Thời gian từ điểm dừng đầu tiên\_p40
* TimeFromFirstStop\_p50 Thời gian từ điểm dừng đầu tiên\_p50
* TimeFromFirstStop\_p60 Thời gian từ điểm dừng đầu tiên\_p60
* TimeFromFirstStop\_p80 Thời gian từ điểm dừng đầu tiên\_p80
* DistanceToFirstStop\_p20 Khoảng cách đến điểm dừng đầu tiên\_p20
* DistanceToFirstStop\_p40 Khoảng cách đến điểm dừng đầu tiên\_p40
* DistanceToFirstStop\_p50 Khoảng cách đến điểm dừng đầu tiên\_p50
* DistanceToFirstStop\_p60 Khoảng cách đến điểm dừng đầu tiên\_p60
* DistanceToFirstStop\_p80 Khoảng cách đến điểm dừng đầu tiên\_p80
* City Thành phố

Tập dữ liệu có 2 phần: train và test

* Tập train có 856387 dòng dữ liệu.
* Tập test có 1921357 dòng dữ liệu.
* Các dòng dữ liệu đầu:



## Phần 3 - Kế hoạch phân tích dữ liệu

* Input: tất cả các biến trong tập train
* Output: phần trăm thứ 20, 50 và 80 cho tổng thời gian dừng lại tại một giao lộ và khoảng cách giữa giao lộ và địa điểm đầu tiên mà một chiếc xe dừng lại trong khi chờ. Có thể coi mục tiêu của mình là tóm tắt sự phân bố thời gian chờ và khoảng cách dừng tại mỗi giao lộ.
* (Các) phương pháp mà nhóm tin rằng sẽ hữu ích trong việc trả lời (các) câu hỏi đặt ra: Nhóm dự định test kết quả với Random Forest, Clustering,... để dự đoán, nếu kết quả không tốt như mong đợi, nhóm sẽ bổ sung thêm thuật toán.
* Kế hoạch thực hiện và phân công việc.