INFOH423 – Project
Hack my Ride
Data Mining Project 2021/22

Mahmoud SAKR and Jean-philippe HUBINONT

Since you study in ULB, you must be a regular user of stib-mivb - the company providing public transport in Brussels. This already makes you a domain expert in this data challenge.

News: stib-mivb has an open data portal, on which the scheduling of the trips as well as the real-time vehicle arrival times are continuously published.

We have collected about 3 weeks of this data for you:

- The location of all vehicles every +-30 seconds, encoded in JSON. The JSON format is described at the end of this document
- Esri Shape files describing the map (lines and stops) of stib-mivb network, two snapshots 3 September and 23 September
- GTFS files containing the offline plan/schedule covering the same period of the vehicle location data, two snapshots 3 September and 23 September
- In addition you get a number of GPS tracks for task number 2

Your team is challenged to analyse this data and provide the following insights:

1. Analyze the vehicle speed over the different network segments, how it varies across segments and over time. Present this in a suitable visual way.
2. Analyze the vehicle delays at the different stops, how it varies across stops, and over time. Present this in a suitable visual way.
3. Given a vehicle start time, do arrival time forecasting at a given stop in the route of this vehicle. You should be able to test the accuracy of your

forecasting by randomly splitting the given dataset in disjoint training and testing subsets.

4. The GPS tracks are for real people moving in Brussels. In fact they are from Mahmoud and Jean-Philippe. You are asked to infer the mode of transport of each of these tracks (bus, tram, etc)
5. Think your own of a valuable analysis on this data

Deliverables

You should deliver a report (as a .pdf) containing the following elements:

1. A cover page with the list of group members, including student ID,
2. A description of dataset loading and preprocessing.
3. A description of your data exploration activity; better accompanied with statistics, figures, screenshots, etc.
4. A clear presentation of your solution for every task. Note that for tasks 3, 4, (and possibly 5) you should present automatic solutions in the form of a program. For example, in task 4, it is not accepted that you visualize the tracks and guess the transportation mode yourself.
5. Your model(s) evaluation step with an explanation
6. A presentation and a demo of your solution

Evaluation

The evaluation jury consists of the two course instructors. You will be asked to present your solution and defend it. The grading will consider the following factors:

- Your data management: loading, integration, exploration (10 points)
- For every analysis task (5 points each):
  - Your justification for the analysis pipeline and the parameter setting
  - Your presentation and interpretation of the results
- Your validation/evaluation of the obtained results in tasks 3, 4, (and 5 if relevant) (5 points)

Vehicle location JSON file format

```
{"data":[{
  "time":"1632409236387",
  "responses":[
     {"lines": [
     {"lineId": "1",
     "vehiclePositions": [
          {"directionId": "8161",
          "distanceFromPoint": 1,
          "pointId": "8122"}, ...]}, "lineId": ...]
     },{"lines":...}
  ],}, {...}, ... ]}
```

This data has been collected by invoking the Vehicle Position Real-Time API of stib-mivb. Every 30 seconds, the API was called 9 times, each with 10 lines IDs. The `time` attribute is the time in milliseconds (unix epoch) at which the API was invoked. The `responses` array has the result of the 9 API calls. Every call returns for the given line IDs all their vehicle positions. Note that for one line, there are normally multiple vehicles at the same time, positioned across the route of the line.

A vehicle position, is a triple:

```
{"directionId": "8161",
 "distanceFromPoint": 1,
 "pointId": "8122"}
```

The `directionId` is the identifier of the terminal stop. The `pointId` is the identifier of the last stop traversed by the vehicle. The `distanceFromPoint` is the distance in meters between the vehicle and the last traversed stop.

The stop data (id, location, name, etc), as well as the routes of the lines are given in the Esri Shape files.