Big Data Management Fall 2016

Project 3

This project contains both technical and critical questions. Technical questions are identified with a T, and critical questions with a C. Questions that require both a critical and a technical perspective are identified with T/C.

Each group should prepare a single document that answers the project questions. This document should be submitted on learnit both in the technical class and in the critical class by November 30th at 23:55. Document length is up to you but remember "brevity is the soul of wit".

In this project your challenge is to develop a plan to (a) assess and (b) improve mobility for city residents. In the dataset we only have cars and people - so assume there are no bikes and no public transport in this city (very boring). Please consider ONE of the three following scenarios and develop your research questions and data views in accordance:

1) You are working with city planning and traffic management. Your job is to determine how to spend money on improving mobility infrastructure in the city for all residents. Assume that a fair percentage of the cars in the city (and in your dataset) are equipped with significant networking capabilities (latest volvo or tesla for example). Come up with a plan for the city government complete with demonstrations of the utility (and limitations) of the data and a solid argument for why your proposal is a good investment.

2) You are a startup brainstorming potential applications and services for city residents to improve their journeys. The city is collaborating with you by making available the mobility data it can gather. Come up with one or two examples of applications or services that convincingly demonstrate the utility (and limitations) of the data, make an argument for why people would use your app and how this would be useful to the city (convince both sides that your app is a good investment).

3) You are in charge of city incident and emergency response services. You would like to streamline incident response for your teams and develop a way to engage citizens to ensure their safety. The city is making available the data it has on resident mobility. Come up with one or two proposals for improvements and investments that would make the city residents safer, demonstrate the utility (and limitations) of the data, make an argument for why these are crucial improvements/investments to make.

You are given a data set about the movement of persons in Copenhagen. The dataset can be accessed at https://ituniversity-my.sharepoint.com/personal/drom_itu_dk/Documents/FCDOutput It takes about 1 hour to download it from the ITU.

This dataset is the result of a simulation process. It was generated using the SUMO toolbox: http://sumo.dlr.de/daily/userdoc/SUMO_User_Documentation.html, by 3 MSc students at IT University of Copenhagen. They describe the simulation steps as follows:

1. To generate that specific data, we downloaded the OpenStreetMap xml data for Copenhagen from https://www.openstreetmap.org/

2. We used NETCONVERT utility to transform into a SUMO map. The utility transforms OSM data into edges (streets), lanes (lanes on edges), nodes (intersections) and traffic light logic which can be used for SUMO simulations. http://sumo.dlr.de/daily/userdoc/NETCONVERT.html

3. Using randomTrips.py we generated random Trips on the map. Trips include just Origin and Destination edges. http://sumo.dlr.de/daily/userdoc/Tools/Trip.html#randomTrips.py

4. Using DUAROUTER we transformed the Trips into SUMO Routes: http://sumo.dlr.de/daily/userdoc/DUAROUTER.html

5. We ran a SUMO simulation using Routes and Map files. Commands for the generation are:MAP: netconvert --osm-files CopenhagenMap.osm.xml -o copenhagen.net.xmlVEHICLES: randomTrips.py -n copenhagen.net.xml -o copenhagen.trips.xml -r copenhagen.rou.xml -L -v --period 0.05PEDESTRIANS: randomTrips.py -n copenhagen.net.xml -o copenhagen.ped.trips.xml -r copenhagen.ped.rou.xml --pedestrians -v --period 0.05 SUMO: sumo -c copenhagen.sumocfg --fcd-output FCDOutput.txt --fcd-output.geo -e 7200   The data is stored in an XML file that contains data for approximately 75.000 cars and 75.000 persons moving around in Copenhagen. The file structure is simple and contains one parent element (timestep) with 2 different child elements (person and vehicle). The attributes are as follows: - Timestep element defines the timestep of the simulation and has an attribute "time" which takes a float value in seconds (from 0.00 to 7200.00) - Person element defines the location of a person with a specific id in the current timestep. Its attributes are: unique id; x and y (lat/long in GPS coordinates); angle; speed; pos (where on the edge is located); edge (street transformed from OpenStreetMap data); slope. - Vehicle element is identical to the Person element, except that it also includes the "lane" attribute which defines on which lane of the edge it is traveling and a "type" attribute which identifies the type of vehicle used. Note that the SUMO simulator "teleports" vehicles once in a while. http://sumo.dlr.de/daily/userdoc/Simulation/Why_Vehicles_are_teleporting.html The data has not been cleaned for when vehicles are teleporting.

For some background on these types of datasets and examples please refer to:
http://vehicular-mobility-trace.github.io
http://kolntrace.project.citi-lab.fr/

Question 1 (C): Defining your goals
Define mobility tracking and transport visibility in the context of a city. What kind of transport infrastructure is visible in such a dataset? Identify potential areas where city person mobility and transport management can be improved. Depending on the scenario you select these potential areas will be different (hint: think of scenarios describing movement through the city and transport scenarios,

relate these to the data you have). Following your selected scenario, identify possible improvements (define improvement - is it efficiency in resource consumption, greater convenience for residents or some combination thereof, etc).

Develop 1-2 scenario-relevant proposals that will drive the project. Keep in mind the following questions: What would you need to know to identify problems and to devise potential changes? Can you know this given the data you have been given? What else might you need for this?

Develop research questions that will help you engage the dataset. Document these initial research questions and how these might change as you engage the dataset.

Question 2 (C/T): Batch layer
Define three views that can be used to get insights about this data set.

Question 3 (T): Master data set
The data is made available in XML.

A. Describe the pros and cons of two different systems to store and manipulate this data (to answer this question you will need to make assumptions about the kind of processing required that are consistent with your answers for Question 2A)

B. Ingest the data into the system of your choice in a way that supports the definition of views defined in Question 2A.

C. Implement the derivation processes that produce the views defined in Question 2

Question 4 (C/T): Data processing
Produce visualizations that can illustrate how the available data addresses your research questions

Question 5 (C/T): Log
List the problems/challenges you faced during this project and explain how you tackled them.

Question 6 (C/T): Selling it
Produce a pitch for your proposal relevant to your selected scenario (this might be a pitch to VCs for the startup or a proposal for the government from transport authority or emergency management). Please address the potential benefits and costs of your proposal. By costs we mean both financial consequences as well as social and ethical concerns. Present visualizations derived from the dataset as support for your pitch. Develop any other relevant documentation that you think may help your case.