Final exam

P.-L. Bourbonnais and N. Saunier

December 4, 2013

Note the scale (the total score is out of 20) and the indicative time to devote to each exercise. Pay special attention to writing, defining the notations you use and clearly identifying your results and responses.

You have access to the course moodle site and two grade sheets of your choice. Statistical tables are available on the moodle website.

Exercise 1: data models 30 min (/4pts)

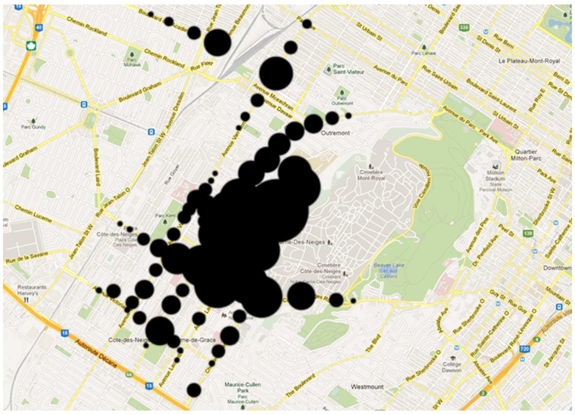
We ask you to present a data model for a transport system. You have two choices: a car and truck rental agency or a centrally managed carpooling service. Your model must include at least 5 entities and constitute a coherent whole (which does not miss an important element necessary for the main functionality of the system). Present the relational model for such a system. Clearly indicate:

1. the primary keys;
2. foreign keys (if the attribute name is not explicit);
3. the most relevant attributes;
4. the data types of these attributes;
5. the cardinality of each relation;
6. the description of the relations;
7. you must have at least one relation of type *many-to-many* (*n-m*).

Exercise 2: GTFS format 45 min (/6pts)

We give you the GTFS files from the Société de transport de Montréal (STM) for the weekday service in January 2014. The STM gives you the mandate to produce an accessibility map. You are asked to draw the map that allows you to visualize how far a person can get in 60 minutes on a weekday from Polytechnique using either walking or public transport (TC) or both (with or without transfers).

The following image is an example of an accessibility map already produced for Polytechnique (anything in black is accessible within the chosen time period):



Here is the relevant data for this question:

- Walking speed: $v_{walking}$ (in m/s)
- Minimum safety time before boarding a vehicle (in addition to walking time): t_{min} (in minutes)
- Consider that the walk is in a straight line (distances as the crow flies)
- Departure at 9:00 am in the morning
- You have the latitude / longitude coordinates of Polytechnique (P_{lat}, P_{lon})
- Transfers are possible (one or more CT lines can be used)
- You don't have to use the `calendar.txt` and `calendar_dates.txt` files, only the week times will be in your `stop_times.txt` file (the GTFS format reference page is available on the moodle website).

Your software can do the following for you:

- Calculate the distance between two lat / lon points
- Draw a radius of x meters around a lat / lon point

What operations must you perform, in order, to obtain the requested accessibility card? Each time, indicate the GTFS file used and the fields you must select to perform your operations. Be specific: a programmer must be able to use your process directly to automate the process of creating the map.

Do not draw the map: we want the algorithm / procedure that allows you to draw it.

Exercise 3: regular expressions 20 min (/ 2pts)

1. Which regular expression allows you to invert two groups of 2 digits followed by a letter, separated by a comma? (1 point) Example:

   ```
   74a, 56b
   ```

   and replace it with:

   ```
   56b, 74a
   ```

Give the regular expression to enter in the search field and the expression to enter in the replacement field.

2. What regular expression detects a Canadian postal code (format: letter number letter number letter number)? (1 point) Example:

```
A1B 2C3
```

Important: the postal code can contain lowercase or uppercase letters and the presence of a space is optional.

Exercise 4: accident data set 55 min (/8)

This exercise is based on a set of 3000 accidents involving a pedestrian and a vehicle between 2003 and 2006 in the city of Montreal (the file is available on moodle). The data is in the form of a text file (with the fields separated by a tab), and each accident is described by the attributes described in the table 1.

Attribute	Description
EVENT	accident number
RDCLASS	road classification (5: motorway; 4: numbered road; 3: collector; 4: artery; 5: local)
SPD_KM	speed according to road classification
MED_INC	median income in the accident area
pop_dens_200	population density within 200 m
veh_type	type of vehicle (“ car ”: car; motorcycle; “ VTB ”: van, truck (“ truck ”) or bus; “ EMS ”: emergency vehicle))
BAD_WEAT	bad weather indicator variable
SEVERITY	severity of the accident (3: fatal; 2: serious injury; 1: slight injury; 0: no injury)
DARK	dummy variable of the night
Park_10	presence of a park 10 m from the accident
hosp_50	presence of a hospital within a radius of 50 m
veh_mvt	movement of the vehicle involved (“ straight ”; “ backup ”; “ leftturn ”; “ rightturn ”)
Comm _Per	percentage of business activity
Res _Per	percentage of residential activity
Inter _Acc	occurrence of the accident at a crossroads

Table 1: Accident attributes

1. Discuss statistical models that can be used to study the association of these attributes with crash severity. (1 point)

2. Describe the processing required to use nominal data in a regression analysis (eg linear regression). (1 point)

3. By creating a new binary variable representing fatal and serious injury accidents (the variable is 1 if the accident is fatal or with serious injury, 0 otherwise), choose a statistical model to study the factors contributing to the probability of a fatal or

serious accident: estimate the model (with Tanagra), clearly present the significant attributes and discuss the results. (4 points)

4. Draw a histogram of the distributions according to the road class of the number of fatal and serious accidents on the one hand, and the number of accidents with minor injuries and without injuries on the other hand: apply a statistical test to determine whether the two distributions are different. (2 points)

Bonus point (/1pts)

Describe the information that the 511Open format can represent (indicate examples of attributes of the data that could be saved in this format).