ICT4SM Laboratory 1 Description 2024/25

Kai Huang

E-mail: kai.huang@polito.it

Acknowledgement

Marco Mellia

Table of content

Lab goals

The laboratory part of the ICT4SM focuses on the analysis of free-floating car sharing data collected from open systems on the Internet. Data has been collected from websites of FFCS systems and made available through a MongoDB database. The goal of the laboratory is twofold:

- Allow students to get used to ICT technologies typically used in the backend of smart society applications – database and remote server access, and writing analytics using simple scripts
- Allow students to work on real mobility data, trying to extract useful information, including geographical and temporal ones – and get used to data pre-processing and filtering.

Accessing the data

Data is stored in a MongoDB server running on the server `bidgatadb.polito.it:27017`. It offers read-only access to clients connected to a network to the following database:

- Collection name: Carsharing
- User: ictts
- Password: Ict4SM22!
- Requires SSL with self-signed certificates

You can use command line interfaces (e.g., the mongo shell) or GUIs (e.g., the Robo3T application) if properly configured. For the mongo shell, you can use

```
mongo --host bigdatadb.polito.it --port 27017 --ssl \
      --sslAllowInvalidCertificates -u ictts –p \
      --authenticationDatabase carsharing
```

Collections

The system exposes 4 collections for the car sharing provider **Car2Go**, which are updated in real time. Those are:

- "ActiveBookings": Contains cars that are currently booked and not available
- "ActiveParkings": Contains cars that are currently parked and available
- “PermanentBookings": Contains all booking periods recorded so far
- "PermanentParkings": Contains all parking periods recorded so far

The same collections are available for for the car sharing provider **Enjoy**. Names are self-explanatory:

- "enjoy_ActiveBookings": Contains cars that are currently booked and not available
- "enjoy_ActiveParkings": Contains cars that are currently parked and available
- ”enjoy_PermanentBookings": Contains all booking periods recorded so far
- "enjoy_PermanentParkings": Contains all parking periods recorded so far

For the cities of Torino and Milano, the system augments the booking information with additional data obtained from Google Maps service: walking, traveling, and public transportation alternative possibilities. Not all of them are available, due to the limited number of queries Google allows.

Lab description

Each group is assigned three cities to analyse. Each group has to work on the project assignment and submit a report.

- Max 6 pages which describes what you have done and the findings.
- Add one header page with the group number, members, etc.
- You can add additional results, figures, etc. in the appendix (might not be evaluated).
- Code, scripts, etc., must be added as separate files

Step 1 – Preliminary data analysis

To get used to both MongoDB and the data at your disposal, investigate first the collections and get used to the document and field stored in each.

- How many documents are present in each collection?
- Why the number of documents in PermanentParkings and PermanentBooking is similar?
- For which cities the system is collecting data?
- When did each collection start? When did each collection end?
- What about the timezone of the init_date and init_time timestamps? Which timezone do they refer to?

Considering the **three cities** assigned to your group.

- What is the total number of cars seen in the whole period in each city? How can you estimate the fleet size in a given period, e.g., one week? How does this relate to the total number of vehicles seen in the whole collection?
- How many bookings have been recorded in December 2017 in each city?
- How many bookings have the alternative transportation modes recorded in each city?

To solve these questions, use MongoDB query. Report queries, their results, and comment your choices and the results.

Step 2 –Car sharing usage characterization

Consider each city of your group, and the period of time of November 1st 2017 – January 31st 2018. Consider the time series (city, timestamp, duration, locations). Process it to further analyse it by producing the following plots and results:

1. Derive the distribution of booking/parking duration and plot them. Show the distributions as empirical Cumulative Distribution Functions (CDF). Which consideration can you derive from the results?
    a. Which city has more density for larger values of duration? Is this expected? Does the CDF suggest the presence of some outliers?
    b. How do you interpret the differences in the CDFs?

   c. Does the CDF change over time? E.g., aggregate per different weeks of data, or per different days. Are these CDFs different? Why?
2. Consider the system utilization over time: aggregate rentals per hour of the day, and then plot the number of booked/parked cars (or percentage of booked/parked cars) per hour versus time of day. Do you notice any outliers? Can you explain them?
3. Derive a criterion to filter possible outliers (e.g., booking periods that are too short/too long), to obtain *rentals* from bookings, filtering system issues or problems with the data collection.
4. Filtering data as above, consider the system utilization over time. How do they change compared to the unfiltered versions? Are you able to filter outliers efficiently for both bookings and parkings? Consider also to plot the CDF of the filtered events. How do these compare to the unfiltered versions?
5. Filtering the data as above, compute the average, median, standard deviation, and percentiles of the booking/parking duration over time (e.g., per each day of the collection).
   a. Do these figures change over time?
   b. Is it possible to spot any periodicity (e.g., weekends vs weekdays, holidays versus working periods)?
   c. Is it possible to spot any trend (e.g., increasing, decreasing, holiday periods)?
6. For the city of Milano, correlate the probability of a rental with the availability of other transport means.
   a. Extract those valid rentals for which there is also the data for alternative transport systems.
   b. Consider one alternative transport system, e.g., public transports. Take the duration, and divide it into time bins, e.g., [0,5)min, [5,10)min, [10,15)min, … Compute then the number of rentals for each bin, i.e., the probability of seeing a rental given the duration of public transport would be in a given interval. Plot the obtained histogram and comment the results.
      Pay attention: do you know the probability of observing a rental of a given duration?
      Remember the Bayes Theorem: P{A|B}=P{B|A}P{A}/P{B}. Which probability are you observing?
7. After filtering the data, consider one city of your collection and check the position of the cars when booked and returned. Then compute the density of cars at rental starting and ending time during different hours of the day.
   a. Plot the origin and destination position of cars at different times using a mapping service of your preference. For instance, how different are the destination zones on Mondays between 8- 10 and 18-20? Or the same time, but on a different day? Or weekends and weekday?
   b. Choose a proper way to divide the area. Estimate travel demand by computing trip generation for origin and destination. Plot the results using a heatmap (i.e., assigning a different colour to each zone to represent the densities of cars).
   c. How different are the destination zones on Mondays between 8- 10 and 18-20? Or the same time, but on a different day? Or weekends and weekday?