Bike-sharing stations: profiling and availability prediction

FOSDEM2018 - Bruxelles

Raphaël Delhome(*), Damien Garaud
Introduction
Shared-bike services

- Shared-bike rental service in large cities
- Small-duration rents
- Stations and availability

Velo'v (Lyon)
Major challenges

- Is it possible to classify bike-sharing stations according to bike availability?
- What if there are no bike when we want one?
- What if there are no available bike station when we have to drop our bike off?
- ...
- How to build a complete ETL framework to analyze data and report results?
Outline

(Part 1) Handle open geospatial data
(Part 2) Bike-sharing station unsupervised classification
(Part 3) Bike and station short-term availability prediction
(Part 4) Demo of an API
Data overview
Open geospatial data

Data GrandLyon

Opendata Bordeaux
## Data presentation

<table>
<thead>
<tr>
<th>id</th>
<th>last_timestamp</th>
<th>bs</th>
<th>abs</th>
<th>ab</th>
<th>bonus</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>10063</td>
<td>2017-07-08 23:49:09</td>
<td>34</td>
<td>10</td>
<td>23</td>
<td>Non</td>
<td>OPEN</td>
</tr>
<tr>
<td>10021</td>
<td>2017-07-08 00:30:12</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>Non</td>
<td>CLOSED</td>
</tr>
<tr>
<td>8038</td>
<td>2017-07-08 23:49:26</td>
<td>20</td>
<td>6</td>
<td>14</td>
<td>Non</td>
<td>OPEN</td>
</tr>
<tr>
<td>7045</td>
<td>2017-07-08 23:52:43</td>
<td>20</td>
<td>13</td>
<td>7</td>
<td>Non</td>
<td>OPEN</td>
</tr>
</tbody>
</table>
Data pipeline

- Build a Python data pipeline thanks to **Luigi**
- Get, transform and store the data
  - gather data every ten minutes (*json, xml, shp*)
  - in-base storage (*postgresql, postgis*)
  - feature engineering and ML treatments
Bike-sharing station classification
Objective

- Classify bike-sharing station according to the way they are used by customers
- Main idea = group stations that looks similar
- ... What does it mean? => Focus on the time series
K-means clustering

- Inspired from a similar work of James Lawlor
- One profile = one individual
- Group similar individual together
- Deduce stations profiles

4 clusters have been identified
Clustered station mapping
Shared-bike availability prediction
Objective

- Know if some bikes (resp. stations) will be available in the next few minutes
- *Main idea* = Use available information to predict availability
- *What does it mean?* => Supervised learning to learn an availability probability
Use a boosting tree method:

- to predict $Y$ (availability probability at $H+1$)
- starting from $X$ (hour, day, available bikes at $H$, ...)

XGBoost method

$$f(\text{boy}) = 2 + 0.9 = 2.9$$

$$f(\text{girl}) = -1 + 0.9 = -0.1$$
Results

Without tuning features, RMSE = 0.095
Demo


Bicycle-sharing data analysis

Get and visualize some bicycle-sharing Open Data.

See the project page on Github.

Summary

Available for the following cities

- Bordeaux (France)
- Lyon (France)

Access

Examples:

- Get some information from a city: URL/api/bordeaux/station
- Retrieve details for a given bicycle-sharing station: URL/api/lyon/station/1009

You also can read and try the REST API generated by Flask-RESTPlus and Swagger.
Demo: documentation

**Jitenshea: Bicycle-sharing data analysis**

Retrieve some data related to bicycle-sharing data from some cities.

<table>
<thead>
<tr>
<th>Method</th>
<th>Route</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET</td>
<td>/city</td>
</tr>
<tr>
<td>GET</td>
<td>/(city)/daily/station</td>
</tr>
<tr>
<td>GET</td>
<td>/(city)/daily/station/(ids)</td>
</tr>
<tr>
<td>GET</td>
<td>/(city)/profile/daily/station/(ids)</td>
</tr>
<tr>
<td>GET</td>
<td>/(city)/profile/hourly/station/(ids)</td>
</tr>
<tr>
<td>GET</td>
<td>/(city)/station</td>
</tr>
<tr>
<td>GET</td>
<td>/(city)/station/(ids)</td>
</tr>
<tr>
<td>GET</td>
<td>/(city)/timeseries/station/(ids)</td>
</tr>
</tbody>
</table>
Demo: Lyon page (1/2)

Bicycle-sharing data analysis for Lyon

Map

Stations

<table>
<thead>
<tr>
<th>id</th>
<th>Name</th>
<th>City</th>
<th>Bikes</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001</td>
<td>Terreaux / Terme</td>
<td>lyon 1 er</td>
<td>16</td>
<td>Rue Terme</td>
</tr>
<tr>
<td>1002</td>
<td>Opéra</td>
<td>lyon 1 er</td>
<td>22</td>
<td>Angle Place de la comédie, rue Joseph Sorlin</td>
</tr>
<tr>
<td>1003</td>
<td>Place Sathonay</td>
<td>lyon 1 er</td>
<td>16</td>
<td>4 Place Sathonay</td>
</tr>
</tbody>
</table>
## Demo: Lyon page (2/2)

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>City</th>
<th>Bikes</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>3090</td>
<td>Vinatier</td>
<td>Lyon 3ème</td>
<td>20</td>
<td>95 bd Pinel</td>
</tr>
<tr>
<td>3091</td>
<td>Hôpital Neurologique</td>
<td>Lyon 3ème</td>
<td>30</td>
<td>59, Boulevard Pinel</td>
</tr>
<tr>
<td>3087</td>
<td>Part-Dieu / Druelle / Garibaldi</td>
<td>Lyon 3ème</td>
<td>20</td>
<td>4 Bd Eugène Druelle</td>
</tr>
<tr>
<td>3088</td>
<td>Guichard / Mazenod</td>
<td>Lyon 3ème</td>
<td>16</td>
<td>5, Rue Moncey</td>
</tr>
<tr>
<td>3089</td>
<td>Trarieux / Lacassagne</td>
<td>Lyon 3ème</td>
<td>24</td>
<td>221 avenue Lacassagne</td>
</tr>
</tbody>
</table>

Showing 1 to 348 of 348 entries

**Charts**

Daily transactions for yesterday. Just the most important ones.

![Daily Transactions Chart](chart.png)
Demo: Terreaux page (1/2)

Profiles
Station profiles: transactions mean for each day of the week (resp. hour of the day) for a period of 30 days.

Hourly
Day profile for the station Terreaux / Terme

Daily
Week profile for the station Terreaux / Terme
Demo: Terreaux page (2/2)

Timeseries

Daily Transactions

Number of transactions these previous days.
Conclusion
Conclusion and perspectives

- Addressing some simple research questions with some open geospatial dataset
- From data source to database (ETL-like) with Luigi
- Production of an API to visualize data => towards production?
- Online learning: keep on gathering data, and learn continuously
Thanks for your attention!

Questions?

damiengaraud@oslandia.com
raphaeldelhome@oslandia.com

See more on Oslandia's blog and on
github.com/garaud/jitenshea