OpenRitardi Datasets

We welcome any sort of analysis on Italian train data. We would be happy to know about your data analysis ideas and even provide you with feedback and support in that. If you obtain some interesting findings or data visualizations, we would be glad to include them on OpenRitardi's website and credit you accordingly.

Introduction

What is OpenRitardi?

OpenRitardi is the first visualization tool of train delays in Italy. You can see the performance of each station and train, and see where the railway service can be improved. The project showcases how open data allows accountability and improves quality of service. OpenRitardi is *open source project* published on GitHub.

Data source

Data is not collected by OpenRitardi directly, but it collected by TrainStats, which is a service that downloads daily data from ViaggiaTreno, the Trenitalia's website that lists train delays. OpenRitardi's contributors have used and analyzed the data, and we are happy to provide you with some guidelines.

Data processing guidelines

Obtaining data

You can download datasets from TrainStats's Mega folder. Choose the days that you want to download (a period of 3-month data is about 10GB). ### Data raw format TrainStats datasets contain:

1. an entry for everytime a train stops at a station in a specific day, and contains information on the train (train class, number, departure, destination), the station, the scheduled time, the actual time and the delay
2. general information on the railway status for a specific day, like Trenitalia or RFI text announcements
3. broad statistics on the number of trains circulated on a specific day

For OpenRitardi usecases we are mainly interested by 1.

Data pre-processing

In order to convert TrainStats' json datasets in a nice tabular dataset which can be used for analysis, you can run our data wrangling script which uses Spark for big-data processing.

In this requirements.txt file you can find the dependencies to be installed in order to run our data wrangling script.

Enjoy!

Now that you have a tabular dataset, where each entry corresponds to the stop of each train in a specific day with the corresponding time and delay, you can put your data scientist hat and try to extract some insightful analysis!

What to do with the data

Feel free to explore it, invent new visualizations or use cases. Below we provide a list of investigations we already carried out, and a list of additional ideas that might be relevant as well.

OpenRitardi's initial analysis

The charts and the data displayed on OpenRitardi's website are computed in this notebook. Some of the analysis we carried out are:

- obtaining the average delay of trains in each station, stratified by day of the week and train class, accessible on OpenRitardi's homepage
- obtaining statistics on delays for the stops of a given train, accessible here
- obtaining the list of best/worst train stations and trains and a regional comparison of train delays, accessive here

Proposed ideas

Something already explored by OpenRitardi for Italy:

- regional differences on average delays
- differences by train type (regional, intercity, arrows) on average delays

or some new analyses and visualizations:

- how long it takes on average to travel in each region (you have the coordinates of the stops and all the schedules, so you can calculate the average speed)
- how many trains there are in each region per inhabitant (perhaps visualized with a cool heatplot)
- average delays per station based on its size (larger stations often have lower average delays for example)
- make a map of Italy where distance is not given by geographical distance but rather by the time it takes to travel around (Naples will be very close to Milan, but very far from Bari, while their geographical distance is comparable)
- compute frequencies of trains connecting cities, for instance Milan is connected to Genova with 20 trains every day while it is connected to Rome 80 times a day. With this, it is possible to make a map of Italy

with segments that connect cities and a segment width or color set by the frequency of connecting trains
- get some inspiration from Chronotrains, BelgianTrains, Swiss Federal Railways, or others

or even machine learning approches like:

- predicting the delay of a train
- using NLP techniques on Trenitalia text announcements to predict specific train delays
- clustering trains or stations and obtain new insights

Contribution and questions

If you'd like to contribute to OpenRitardi or to get some feedback, feel free to open a pull request or an issue on our GitHub.