hugomcp@di.ubi.pt, 2021/22

Data Science
Practical Project

LSTW: Large-Scale Traffic and Weather Events Dataset

Available from https://smoosavi.org/datasets/lstw, the LSTW is a large-scale, country-wide dataset for transportation and traffic research, which contains traffic and weather event data for the United States. In terms of traffic, the dataset supplies several types of events including accident, congestion, construction, etc. In terms of weather events, we have several types including rain, snow, storm, cold weather event, etc.

The dataset is *continuously* being collected from August 2016, and today it contains about **37 million** traffic and weather events in total (31.4 million traffic and 5.6 million weather events).

Traffic Events

A traffic event is a spatiotemporal entity, where each entity is associated with location and time, and one of the following event types:

- Accident: Refers to traffic accident which can involve one or more vehicles.
- Broken-Vehicle: Refers to the situation when there is one (or more) disabled vehicle(s) in a road.
- Congestion: Refers to the situation when the speed of traffic is lower than the expected speed.
- Construction: Refers to an on-going construction, maintenance, or re-paring project in a road.
- Event: Refers to the situations such as *sport event*, *concert*, and *demonstration*.
- Lane-Blocked: Refers to the cases when we have blocked lane(s) due to traffic or weather condition.
- Flow-Incident: Refers to all other types of traffic events. Examples are *broken traffic light*and *animal in the road*.

The data is provided in terms of a “csv” file with 19 attributes, describing the type of event, its severity, the starting and end time and the location (see full list of attributes in annex).

Weather Events

Again, the weather event is also a spatiotemporal entity, where each element is associated with location and time, and one of the following types:

- **Severe-Cold**: The case of having extremely low temperature, with temperature below *-23.7* degrees of Celsius.
- **Fog**: The case where there is low visibility condition as a result of *fog* or *haze*.

- **Hail**: The case of having solid precipitation including *ice pellets* and *hail*.
- **Rain**: The case of having rain, ranging from light to heavy.
- **Snow**: The case of having snow, ranging from light to heavy.
- **Storm**: The extremely windy condition, where the wind speed is at least *60 km/h*.
- **Other Precipitation**: Any other type of precipitation which cannot be assigned to previously described event types.

Similarly, the data is provided in a "csv" file, with 13 attributes, listed in annex.

Tasks

1. ***Google* Colab** – This is a free service, based in *Jupyter Notebooks,* that supplies free GPU processing. It can be used, in a web browser, in the development of Python projects related to Computer Vision, Machine Learning and Data Science domains, using tools/libraries such as (Keras, PyTorch, TensorFlow,...).

   Using one of the many resources (tutorials) available at the web (e.g., https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c), the first task is to configure the Google Colab service and execute a simple "Hello World" notebook, using Python.

2. Considering the typical phases associated to a Data Science project, the main goal of this practical work is to infer the maximum possible rules/knowledge from this data, such as:

   - What traffic events are the most related to weather conditions?
   - What are the most dangerous hours for driving, at a particular location?
   - What is the probability of suffering traffic problem, when travelling from "A" to "B" at day "X"?
   - Can we predict one possible traffic event, at a specific place? And what about the weather events, can we predict the next relevant event?
   - In terms of visualization, how can we illustrate the traffic and weather events frequency distribution in the most appealing (graphical) way?

The project should be developed using Jupiter Notebooks, in a Google Collab environment. Students should produce a report that describes the most important decisions taken, and the most relevant steps of the work carried out, emphasizing the phases typically associated to Data Science projects.

Annex

List of traffic events attributes

#	Attribute	Description	Nullable
1	EventId	This is the identifier of a record	No
2	Type	The type of an event; examples are *accident* and *congestion*.	No
3	Severity	The severity is a value between 0 and 4, where 0 indicates the least impact on traffic (i.e., short delay as a result of the event) and 4 indicates a significant impact on traffic (i.e., long delay).	No
4	TMC	Each traffic event has a Traffic Message Channel (TMC) code which provides a more detailed description on type of the event.	No
5	Description	The natural language description of an event.	No
6	StartTime (UTC)	The start time of an event in UTC time zone.	No
7	EndTime (UTC)	The end time of an event in UTC time zone.	No
8	TimeZone	The US-based timezone based on the location of an event (eastern, central, mountain, and pacific).	No
9	LocationLat	The latitude in GPS coordinate.	Yes
10	LocationLng	The longitude in GPS coordinate.	Yes
11	Distance (mi)	The length of the road extent affected by the event.	Yes
12	AirportCode	The closest airport station to the location of a traffic event.	Yes
13	Number	The street number in address record.	Yes
14	Street	The street name in address record.	Yes
15	Side	The relative side of a street (R/L) in address record.	Yes
16	City	The city in address record.	Yes
17	County	The county in address record.	Yes
18	State	The state in address record.	Yes
19	ZipCode	The zipcode in address record.	Yes

List of weather events attributes

#	Attribute	Description	Nullable
1	EventId	This is the identifier of a record	No
2	Type	The type of an event; examples are *rain* and *snow*.	No
3	Severity	The severity of an event, wherever applicable.	Yes
4	StartTime (UTC)	The start time of an event in UTC time zone.	No
5	EndTime (UTC)	The end time of an event in UTC time zone.	No
6	TimeZone	The US-based timezone based on the location of an event (eastern, central, mountain, and pacific).	No
7	LocationLat	The latitude in GPS coordinate.	Yes
8	LocationLng	The longitude in GPS coordinate.	Yes
9	AirportCode	The airport station that a weather event is reported from.	Yes
10	City	The city in address record.	Yes
11	County	The county in address record.	Yes
12	State	The state in address record.	Yes
13	ZipCode	The zipcode in address record.	Yes

References

Moosavi, Sobhan, Mohammad Hossein Samavatian, Arnab Nandi, Srinivasan Parthasarathy, and Rajiv Ramnath. "Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data." In proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019.