Event Prediction:

Model Implementation and Evaluation

2022-05-03 11:00AM (EST)

Huigyu Yang
Sungkyunkwan University
huigyu@skku.edu

Presentation Outline

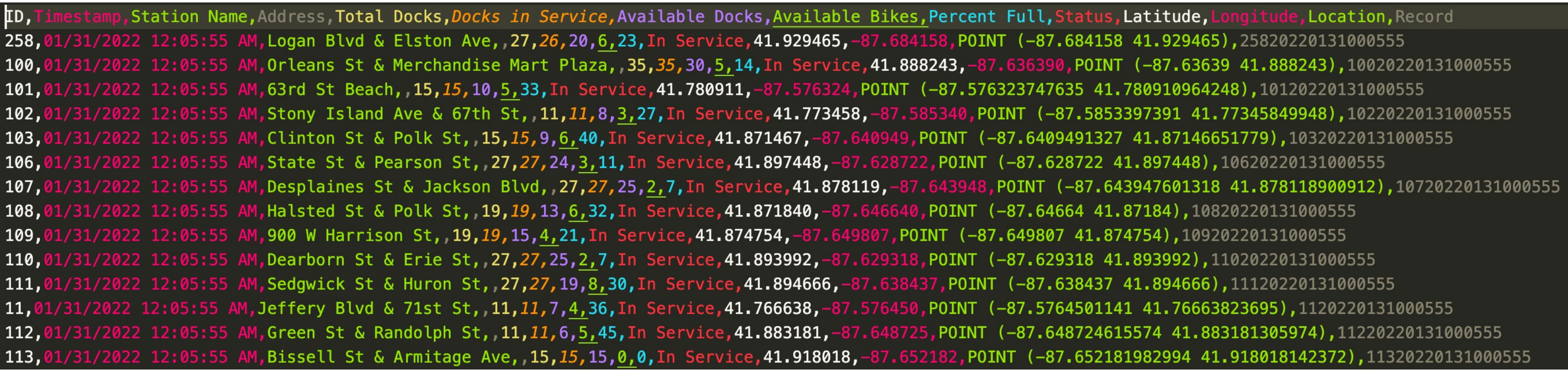
- Chicago Bike Station Dataset
- Model Implementation
- Evaluation
- Future Work

Event Dataset (1/3)

- Datasets in the studies contain only bike rental and return location of user trips
  - New York, Washington DC, Singapore, and Taipei
- Chicago city bike dataset has abundant information among above datasets
  - User trips: rental station, return station, time
  - Bike Routes: GPS information of common traveling routes
  - Historical bike stations: station, capacity, utilization of bike docks

Event Dataset (3/3)

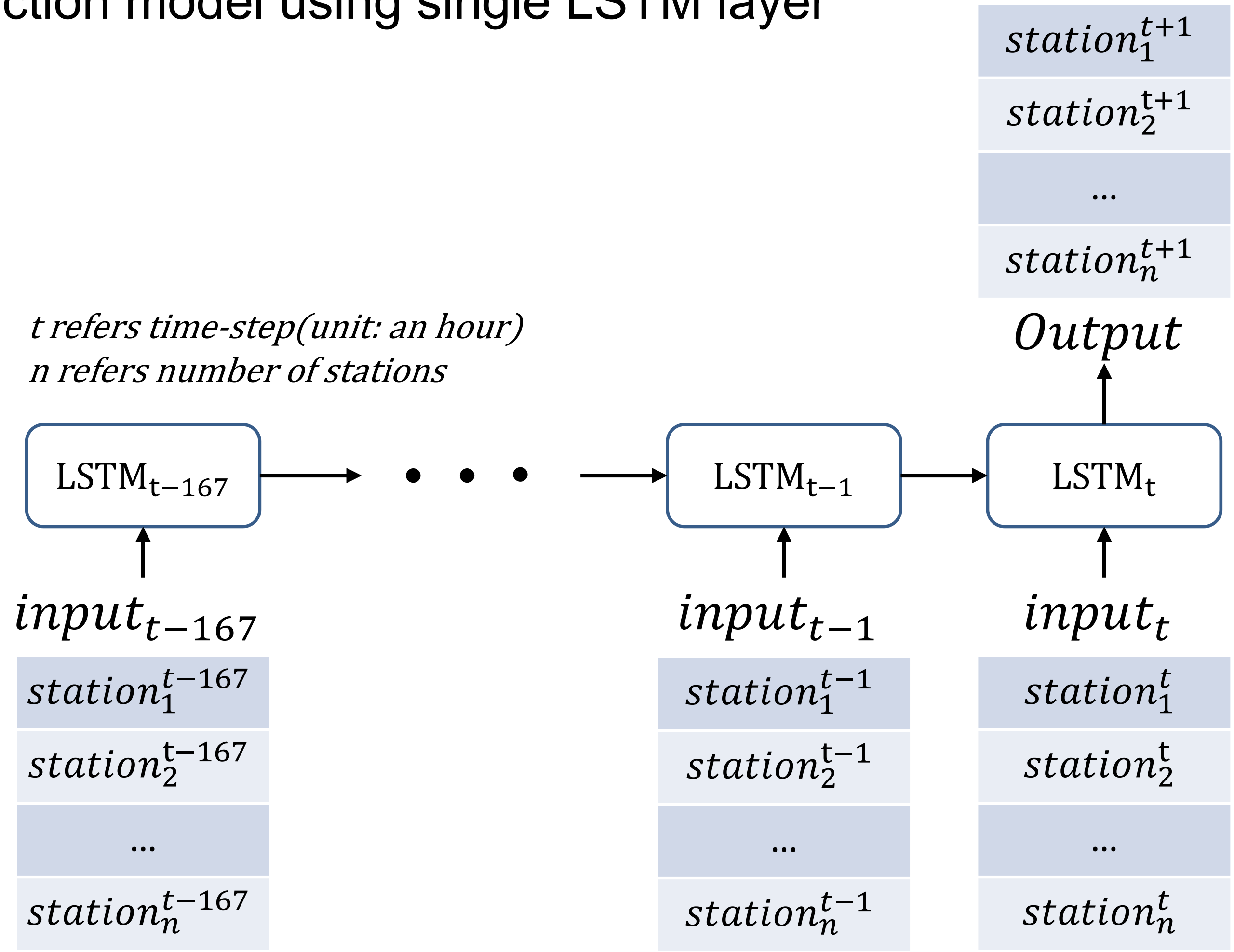
- Chicago historical bike stations dataset
  - 3,451,068 rows for 1 month
  - 842 stations
  - Dock utilization status is logged at every 1 hour



```
ID,Timestamp,Station Name,Address,Total Docks,Docks in Service,Available Docks,Available Bikes,Percent Full,Status,Latitude,Longitude,Location,Record
258,01/31/2022 12:05:55 AM,Logan Blvd & Elston Ave,,27,26,20,6,23,In Service,41.929465,-87.684158,POINT (-87.684158 41.929465),25820220131000555
100,01/31/2022 12:05:55 AM,Orleans St & Merchandise Mart Plaza,,35,35,30,5,14,In Service,41.888243,-87.636390,POINT (-87.63639 41.888243),10020220131000555
101,01/31/2022 12:05:55 AM,63rd St Beach,,15,15,10,5,33,In Service,41.780911,-87.576324,POINT (-87.576323747635 41.780910964248),10120220131000555
102,01/31/2022 12:05:55 AM,Stony Island Ave & 67th St,,11,11,8,3,27,In Service,41.773458,-87.585340,POINT (-87.5853397391 41.77345849948),10220220131000555
103,01/31/2022 12:05:55 AM,Clinton St & Polk St,,15,15,9,6,40,In Service,41.871467,-87.640949,POINT (-87.6409491327 41.87146651779),10320220131000555
106,01/31/2022 12:05:55 AM,State St & Pearson St,,27,27,24,3,11,In Service,41.897448,-87.628722,POINT (-87.628722 41.897448),10620220131000555
107,01/31/2022 12:05:55 AM,Desplaines St & Jackson Blvd,,27,27,25,2,7,In Service,41.878119,-87.643948,POINT (-87.643947601318 41.878118900912),10720220131000555
108,01/31/2022 12:05:55 AM,Halsted St & Polk St,,19,19,13,6,32,In Service,41.871840,-87.646640,POINT (-87.64664 41.87184),10820220131000555
109,01/31/2022 12:05:55 AM,900 W Harrison St,,19,19,15,4,21,In Service,41.874754,-87.649807,POINT (-87.649807 41.874754),10920220131000555
110,01/31/2022 12:05:55 AM,Dearborn St & Erie St,,27,27,25,2,7,In Service,41.893992,-87.629318,POINT (-87.629318 41.893992),11020220131000555
111,01/31/2022 12:05:55 AM,Sedgwick St & Huron St,,27,27,19,8,30,In Service,41.894666,-87.638437,POINT (-87.638437 41.894666),11120220131000555
11,01/31/2022 12:05:55 AM,Jeffery Blvd & 71st St,,11,11,7,4,36,In Service,41.766638,-87.576450,POINT (-87.5764501141 41.76663823695),1120220131000555
112,01/31/2022 12:05:55 AM,Green St & Randolph St,,11,11,6,5,45,In Service,41.883181,-87.648725,POINT (-87.648724615574 41.883181305974),11220220131000555
113,01/31/2022 12:05:55 AM,Bissell St & Armitage Ave,,15,15,15,0,0,In Service,41.918018,-87.652182,POINT (-87.652181982994 41.918018142372),11320220131000555
```

Implementation

- A prediction model using single LSTM layer



t refers time-step(unit: an hour)
n refers number of stations

$station_1^{t+1}$
$station_2^{t+1}$
…
$station_n^{t+1}$

$Output$

LSTM_{t-167} → • • • → LSTM_{t-1} → LSTM_{t}

$input_{t-167}$

$station_1^{t-167}$
$station_2^{t-167}$
…
$station_n^{t-167}$

$input_{t-1}$

$station_1^{t-1}$
$station_2^{t-1}$
…
$station_n^{t-1}$

$input_t$

$station_1^{t}$
$station_2^{t}$
…
$station_n^{t}$

Implementation

- A preprocessing algorithm uses two columns which are “station name” and “available bikes” of the historical dataset
- The algorithm collects raw data per station name and generates the list of available bikes

- The prediction model requires input data to normalized into scale [0,1]
- The output of the model can be normalized in two different ways
  - Dock utilization percentage-based input data
    - $Station_n^t = (\#\ of\ Available\ Bikes\ at\ time\ t)/\mathrm{Capacity}(station_n)$
  - Max value-based normalization
    - $Station_n^t = (\#\ of\ Available\ Bikes\ at\ time\ t)/\mathrm{Max}(station_{1:n}^{(t-167):t})$

Implementation: errors in previous trial

- Glob library loads raw historical data from single directory where the multiple files of daily historical datasets
- The library of latest version randomly selects files in allocated directory path
- This caused training and testing data to be concatenated without any temporal continuity

- This logical error in previous experiments has been fixed using function Sorted()

Evaluation

- The outputs of the prediction model is converted to Boolean typed values
- Dock utilization percentage-based input data
  - $If\ station_n^{t+1} \geq \mathrm{boundary_{upper}}$ or $station_n^{t+1} \leq \mathrm{boundary}_{lower}$:

    $\mathrm{Output_n} = \mathrm{True}$
  - *Else:* $\mathrm{Output_n} = \mathrm{False}$
- Available bike number and max normalization-based input data
  - $\mathrm{Temp_n} = station_n^{t+1} \times \max(station_{1:n}) \div \mathrm{capacity}(station_n)$
  - $If\ \mathrm{Temp_n} \geq \mathrm{boundary_{upper}}$ or $\mathrm{Temp_n} \leq \mathrm{boundary}_{lower}$:

    $\mathrm{Output_n} = \mathrm{True}$
  - *Else:* $\mathrm{Output_n} = \mathrm{False}$
- The number of match between converted labels and outputs are divided by total number of test cases to calculate accuracy result

Evaluation

- Percentage based normalization
  - Input: 168 hours, Output: 1 hour
  - Hidden dimension(LSTM): 256
  - Accuracy: 93.46

- Max value-based normalization
  - Input: 168 hours, Output: 1 hour
  - Hidden dimension(LSTM): 256
  - Accuracy: 83.46

- Number of LSTM layers, input length didn’t affect on the result

Future Work

- Find time period when both historical and trip data are available

- Generate bike transition matrix using trip data

- Add fusion layers in the model to train the matrix with historical data