Capstone

Data Set loading and Data Summary

(Library Event Data)

Sravan Kumar Borra

Now days all the research has been concentrated on Big data and Internet of things (IoT). Big data is a vast concept used for analyzing and studying the data collected from different sources. IoT is concept about connectivity where everything is connected to internet. All these concepts are digitally revolutionized and the research is in nascent stage. One of the most important concepts which were based on IoT is Smart city.

Smart city are typically based on connectivity where everything is connected to internet. Insights for these cities are based on big data. The data is collected from different sources such as traffic, library events, pollution etc. All this data is collected and analyzed to provide better services to the people living in the city. Mainly insurance companies are looking forward to collect the data and analyze it.

**Data**

For analyzing the data, one should have data and that data should be a real time data because better insights will be provided by real time data. In this paper, Library event data is selected to provide better insights for managing events for libraries for a city. This data is used by library administration to check the activities of the users of library. They can create daily work reports and assessments based on the users work on that particular day.

Library event data has been fetched from City pulse website. This website provides datasets for various sectors which will help to build smart cities. The administrators are conducting research to build smart city concepts based on these datasets. All these datasets are based on Denmark and other European countries.

Library event dataset is a real time dataset of Aarhus and Denmark from 2013 to 2015. The dataset contains the data of library events from October 2013 to June 2015. This dataset has been provided in raw form in .csv format and TTL format. We are using .csv format to perform analysis in Zeppelin. Zeppelin is platform where we can perform our programming in Spark.

The data contains about the events conducted in various libraries across Denmark from 2013 to 2015. The data has been divided into various categorical variables with different formats. There are about twenty variables which contains city, zip code, url of the event etc. There are about 1500 events conducted in between 2013 and 2015 at different locations in Denmark and Aarhus. The time and date of those events are also provided in this data set. The first column of the data set is library Id. Every library has been allotted with specific Id for identification purpose. This ID is in the form of integer and it is defined as string in Zeppelin.

Following the library ID, the city of the library has been provided which also gives us the information about the city where the event has been conducted. The title of the event is also provided. So the data can be separated based on the title also. End time of the event has been enlisted in the data to give the information about the starting and ending time of the event. A URL has been provided for each event which takes us to the website of the library where we can get more information about that particular event and attractions of that event. Price column has also been provided for every event to give the information about pricing range. There is a column named “Changed”. It gives the information about the event if there are any changes in timings or dates. We can know if event has been postponed or pre-poned.

Library name has been provided in the library column. Different library names were enlisted in that column. Image URL has also been provided in the data set which is particularly a poster which has all the information, important dates, timings and highlights of that particular event. A teaser name was also provided to search the video of the event conducted in a particular library. Street name has also been provided which makes it easy for the people to find the library address and also helps for commuting and GPS purposes. All this information is very useful for performing analyses and built predictive models which help in forming a smart city.

But all the data has not been one sided which concentrates on smart city, connectivity, GPS etc. Longitude and Latitude of the location has also been provided so that people can use those location addresses and commute using a map. All these events are streamed on internet and on television. Stream time has also been provided where people can watch the event on television at that particular time. Date also has been provided in the streaming column so that the viewers can know the date and time of streaming.

**Sample data**

Sample data has been provided below. In the given table there are library id columns, city name, title of the event, street name of the location of the library, start time and end time of the event, longitude numbers are shown. This is included so that we can know about the format of the data set and the categorical variables.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| lid | city | endtime | title | street | status | longitude | starttime | latitude | \_id | id | streamtime |
| 1 | Aarhus | 11/5/2013 17:00 | Lav din egen 'bogÃ¦der' | MÃ¸llegade 1 | 1 | 10.20018 | 11/5/2013 15:00 | 56.15662 | 258 | 8989 | 12/8/2014 17:44 |
| 42 | Viby J | 12/30/2013 0:00 | Keramiker Susanne Bidstrup udstiller pÃ¥ Viby Bibliotek | Skanderborgvej 170 | 2 | 10.16443 | 12/2/2013 0:00 | 56.1304 | 379 | 9235 | 12/8/2014 17:44 |
| 6 | Beder | 2/27/2014 21:00 | Ã˜lsmagning | Kirkebakken 41 | 2 | 10.21605 | 2/27/2014 19:00 | 56.06032 | 371 | 9227 | 12/8/2014 17:44 |
| 55 | Ã…by | 10/31/2013 0:00 | Susanne Butcher udstiller pÃ¥ Ã…by Bibliotek i oktober | Ludvig Feilbergs Vej 7 | 1 | 10.16252 | 10/1/2013 0:00 | 56.15644 | 233 | 8920 | 12/8/2014 17:44 |

All this data has been loaded into Zeppelin so that we can work on the data set for building a model which can provide all the information for the users in the library. This data set is used to provide information about the events conducted in various libraries across Denmark. A sample model can be developed using this data which will be helpful for building one of the important sectors in Smart city which is Library.