Project on Similarity Search

Ankit Dhall

The query was pre-defined into variables. The categorical variables of this query were then encoded into numeric values to enable distance calculation.

This encoding was done on the basis of the encoding used in the MIF generation code.

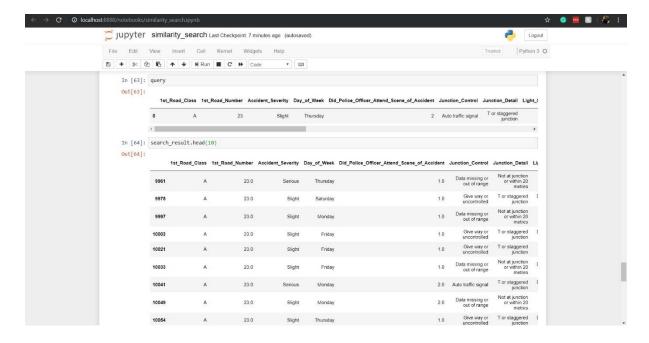
Reference points were also defined on the basis of the reference points used in the MIF generation code.

After running the search algorithm, the nearest/most similar data objects from the input file were retrieved and stored in a csv file.

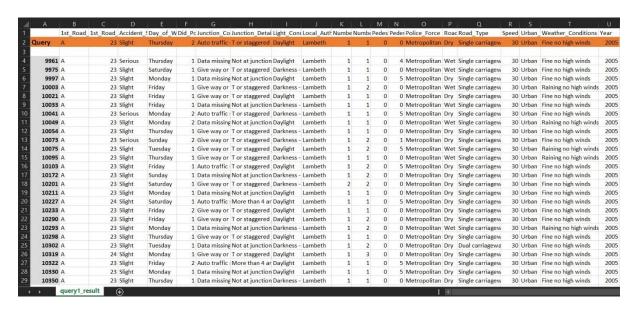
Query 1:

```
1st Road Class = "A"
1st_Road_Number = "23"
Accident_Severity = "Slight"
Day of Week = "Thursday"
Did_Police_Officer_Attend_Scene_of_Accident = "2"
Junction_Control = "Auto traffic signal"
Junction_Detail = "T or staggered junction"
Light_Conditions = "Daylight"
Local_Authority_(District) = "Lambeth"
Number of Casualties = "1"
Number_of_Vehicles = "1"
Pedestrian Crossing-Human Control = "0"
Pedestrian_Crossing-Physical_Facilities = "0"
Police_Force = "Metropolitan Police"
Road_Surface_Conditions = "Dry"
Road Type = "Single carriageway"
Speed_limit = "30"
Urban_or_Rural_Area = "Urban"
Weather_Conditions = "Fine no high winds"
Year = "2005"
```

Execution of the code was done in a Jupyter Notebook and the following query and results were obtained.



These search results were stored in a CSV file along with the query



In the above screenshot, the query has been highlighted in orange. The indices of the data objects (from the original data file) retrieved in the search results are highlighted in grey.