Data Management & Gathering - NLP

Prediction Task

We intend to predict the **true** coordinates of a bus stop.

From our data exploration, we've discovered some uncertainty in the data, in the following image, we've created a heatmap for all the reported stops at bus stop `415`.

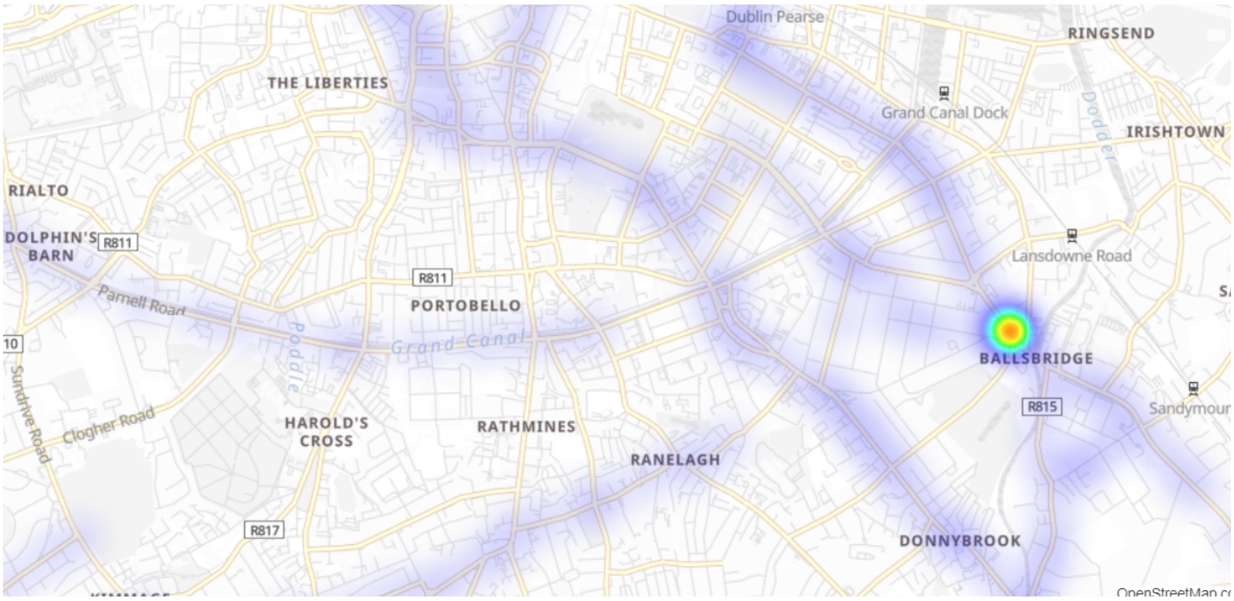


Figure 1. Bus Stop 415 Heatmap

From previous labs, we suspect that many reports report the **wrong** bus stop number.

Textual Data Integration

We've used "HERE Maps" REST-API to *reverse geocode* (translating coordinates to street address) all `atStop` observations from the data (`50 million`).

The mentioned API has a strong limitation of 50 requests per second, if we were to retrieve all street addresses, it would've take two weeks.

We were able to reduce the amount of required requests from the API to `300,000`, by reducing the coordinates precision to 5 decimal degrees (`1 squared meter` error).

Then we've added (from RTPi) all the bus stations names, both in English & Gaeilge (Irish).

We plan to calculate the *distance* between the reported bus station name, and the actual street address derived from the coordinates.

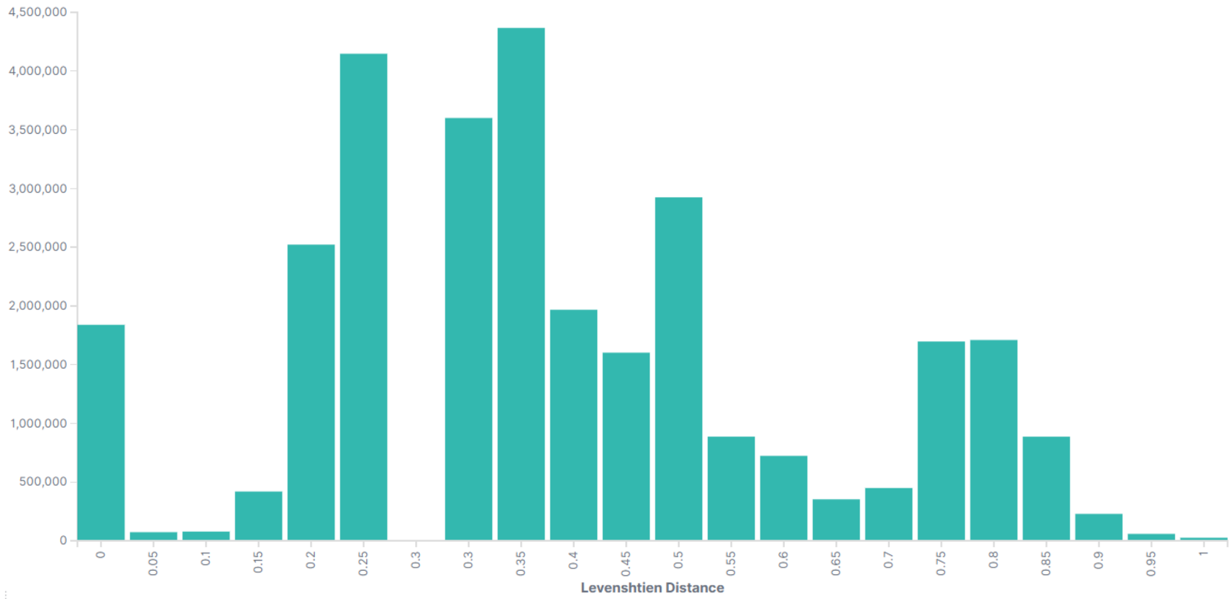
This is under an assumption that most bus stations names are determined by the street they are located in.

Text Distance

We decided to use Levenshtien distance (normalized) to calculate the distance between names (street, bus station). It measures both characters order, and content, while takes into consideration different word lengths.

Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

Levenshtien On Data



Improve the prediction

We will use Levenshtien distance to filter all observations which are not *close* enough to its reported bus station.

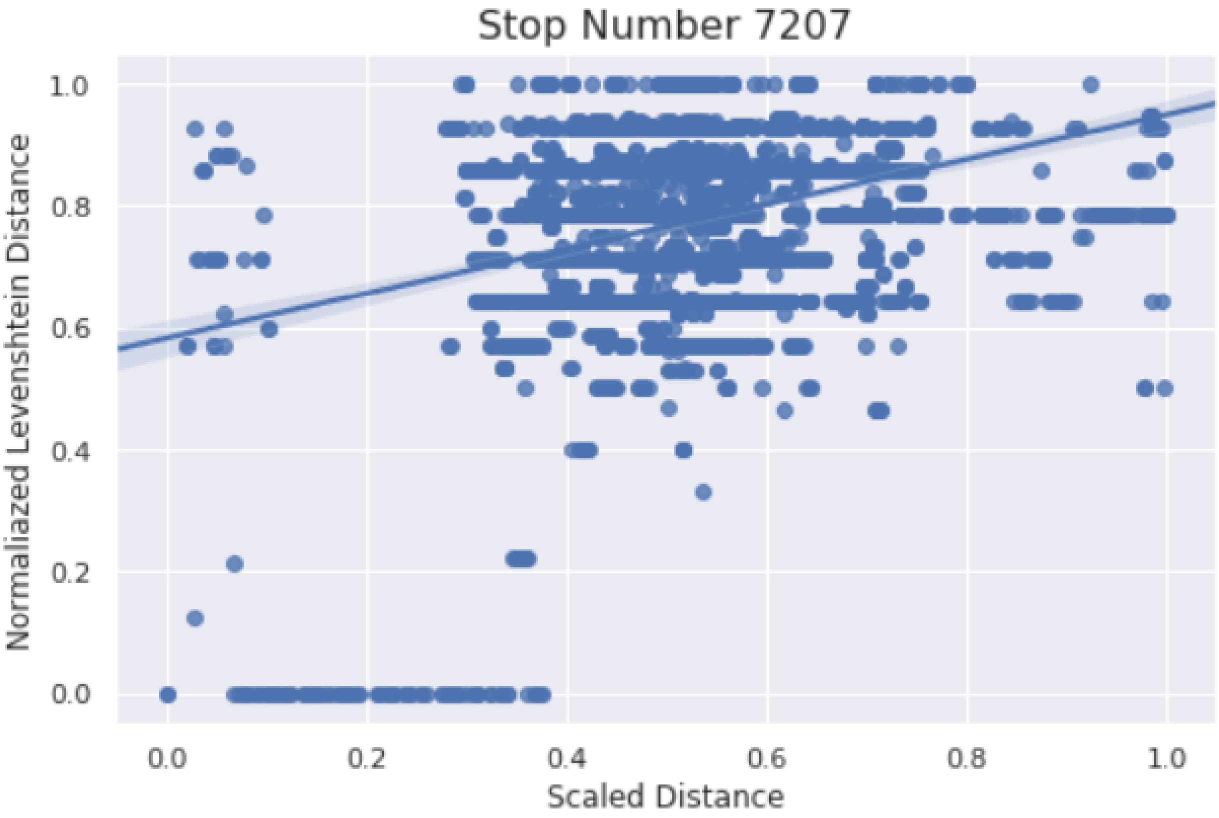
> We defined *close* with Levenshtien distance < `0.5`.

Table 1. MSE Comparsion (Km^2)

Task 3	Task 4
0.10125	*0.2739*

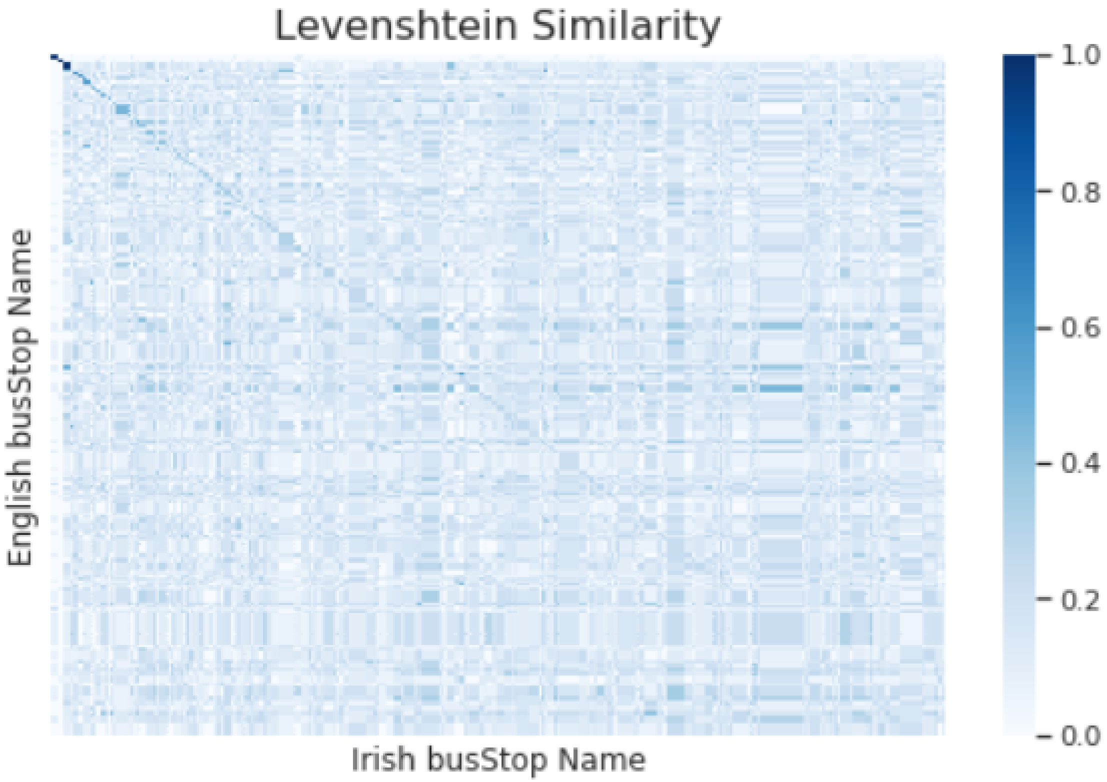
Textual Analysis

Following our prediction task, we were interested in the relation between the textual distance & spatial distance of bus stations.
We computed the Levenshtien distance & hoversine distance from a between all bus stops (to all bus stops).
To visualize it, we will present a scatter plot for bus stop `7207`:



Following our article presentation (GNMT), we were interested wether bus stations names have some similarity between English & Gaeilge (Irish).

We were able to compute a confusion matrix to represent the Levenshtien similarity (normalized distance) between Gaeilge & English station names:


Levenshtein Similarity
English busStop Name
Irish busStop Name