IIW 7 – Record Linkage

REPORT

Task 2

1) **What is your dataset?**
   The dataset selected is the restaurant dataset of Fodor's and Zagat's.
2) **What are the fields in your datasets? Describe them**.
   Each record in the dataset has following field:
   r_name: Name of the restaurant
   r_address: Address of restaurant
   r_number: Phone number of the restaurant
   r_type: Type of restaurant
   row_num: This was added temporarily for report purpose and have no relationship with the record linkage whatsoever.

Task 4

1) **What were the tools you used or were there any libraries you used to write your own code to perform record linkage? Describe your methodology.**
   First my own python code was written to separate field from string of data. This was manually verified.
   Secondly FRIL tool was used. In this tool the two csv files as dataset1 and dataset2 was provided as input. Each csv files had fields descripted in Task 2. Required configuration was done, all the columns were selected in the output. Different similarity algorithm was performed with mix and match of single field and groups of field with different weights. Also different join methods were performed and finally output was taken in a csv file and checked manually the results with the groundtruth file.
2) **Describe for each pair of fields between your datasets that were compared together, what was the similarity measure used and why did you choose to use it.**
   The comparison was done on similar kind of field in the two dataset. In some case a multiple fields (number and name) or(number and address) and individual field were compared.
   dataset1.r_name -> dataset2.r_name
   dataset1.r_address -> dataset2.r_address
   dataset1.r_number -> dataset2.r_number
   dataset1.r_type -> dataset2.r_type

   The comparison was also done on name field with Jaro similarity algorithm with Blocking Join type which gave only 13 true positive results rather than giving atleast 20 true positive result.

Hence the need to use some other similarity algorithm. Various different algorithm and fields were used and below was the final result along with the algorithm used.

The final comparison was made on **dataset1.r_name ->dataset2.r_name** with 30% weight and distance measure as **soundex** and **dataset1.r_number->dataset2.r_number** with 70% weight and distance measure as **edit distance** and **Blocking** join method on number. The blocking join in number helps to eliminate large number of comparisons between record in the datasets.
Why edit distance?
It provided better results. Since the number was well formatted except for the difference of "-" and "/" and provided high level of similarity score. In many case only one translation would be needed.
Why soundex?
Soundex is applied on name and it can provide better results even if there is small misspell in the name of the restaurant which can be a common mistake.

Table

Record in dataset 1	Record in dataset2 (groundtext.txt)	Record in dataset2 (result)	Matched?
3	3	3	Yes
4	4	4	Yes
34	7	Null	(Did not get record in the result. This is false negative.)
10	17	17	Yes
13	19	19	Yes
16	22	22	Yes
17	23	23	Yes
27	29	Null	(Did not get record in the result. This is false negative.)
30	30	30	Yes
31	31	31	Yes
61	43	Null	(Did not get record in the result. This is false negative.)
41	52	52	Yes
45	58	58	Yes
68	67	Null	(Did not get record in the result. This is false negative.)
52	69	69	Yes
54	70	70	Yes
55	73	73	Yes
62	77	77	Yes
65	83	83	Yes

73	89	89	Yes
40	47	47	Yes

Precision = (tp/(tp+fp))

= 17(17+0)

=1

Total records in groundtruth = 21
Total record predicted by tool = 19
Number of records in ground truth which matches the output of tool = 17
Number of record in groundtruth not predicted by tool = 4

There was no record which was falsely predicted by the tool. However, there were total of 19 record linkage provided by the tool out of which 17 record were there in goundtruth.txt. Also there were 4 record in groundtruth.txt which were not predicted by the tool. They are termed as false negative.

The performance of the tool was good. The precision 1 tells that the system did not detect any false positive. Though the system detected some false negative the overall performance was good.