Information Integration on the Web: Homework 10

Due on April 3, 2015

Prof. Ambite & Knoblock

Tushar Tiwari

Problem 1

Describe the performed record linkage

Solution

Data Transformation

After loading both the files, a few transformations were performed in order to get better matches.

- Transform name, addr and type to lower cased strings. This ensures that different casings of the field do not cause a wrong match. For example, "steakhouse" and "Steakhouse" may not give a perfect edit distance score even though they are spelled the same way.
- Trimmed any leading and trailing whitespaces
- Transformed phone num of d1.csv from "xxx/xxx-xxxx" to "xxx-xxx-xxxx" using regex.

Metrics for matching

- In the given data I immediately noticed that a lot of the matches have the exact same phone number assosciated with them. That is a matching pair of records shared the same phone number. This is however not true for all, but for most. Hence, I added a q grams comparision metric on phone no with weight 60% because if they have the exact same phone number then they must be a matching pair. The q value was chosen to be 4 because most restaurants will share the same area code so it must be atleast more than 3.
- After using the above metric, I noticed that the some of the records that did not match from the above metric had same names but different phone nums and hence I added an equal fields boolean distance metric on the name column. Obviously records with exactly the same name will be a matching pair. The weightage I gave this was 30% because the soundex metric explained below added another 10%.
- A phone match was causing lots of matches with similar confidence levels and to put one match ahead of the other match, I added a metric that performed a soundex on the name with weightage 10%. If there were two matches on phone numbers, then the one with more similar soundex name would get selected.
- I chose an acceptance level of 40% because records with same name should get accepted if their phones do not match. 30% from exact and 10% from soundex.

F-Score

- The output file generated (output.csv) by fril contains 112 total matches.
- The groundtruth file also contains 112 matches.
- The number of true positive matches is 109.
- $precision = 109/112 = 0.97321427$
- $recall = 109/112 = 0.97321427$
- $F - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} = 0.97321427$