HW3: Entity Resolution Game

Team Members: Shih-Hua Yu sy2734, Lisa Kim lk2715, Chang Ding cd2959

Registered Username and Email on competition site:

shihhuay sy2734@columbia.edu,

lk lk2715@columbia.edu,

cd cd2959@columbia.edu

1. Describe your entity resolution technique, as well as its precision, recall, and F1 score.

In our entity resolution program, we first produced pairwise observations from all rows in the locu and foursquare training datasets. We then calculated the Jaro-Winkler similarity between the following pairs of features from the two datasets: name, phone number, street address, locality, and postal code. We also calculated the absolute difference between longitudes and latitude, respectively.

Secondly, we programmatically skimmed through the matches in the training examples, and excluded observations that appear to be incorrect matches using the following criteria:

We removed matches that satisfy the following conditions simultaneously:

a. Name Jaro-Winkler similarity smaller than 0.9

b. Phone Jaro-Winkler similarity smaller than 1

c. Name Jaro-Winkler similarity * address Jaro-Winkler similarity smaller than 0.4

Thirdly, we applied blocking to our dataset, in order to filter out observations that are clearly non-matches, therefore reducing the number of observations that will be fed into our model later.

Our blocking technique is as following:

For each locu id, we extracted the subset of pairs that contained the locu id. Then, we computed the 15^{th} percentile of Jaro-Winkler similarity and the 85^{th} percentile of Jaccard distance. We removed all the matching pairs that had lower Jaro-Winkler similarity and higher Jaccard distance than the value of the percentiles we obtained.

Lastly, we used random forest classifier combined with grid search to find a best performing model, measured by f1 score. As a result, the performance of our best model using the online competition is as following:

- Precision rate: 100%
- Recall rate: 95.42%
- F1 score: 97.65%.

2. What were the most important features that powered your technique?

We looked at feature importance using the results from our best random forest classifier model, and found our input features ranked from most important to least important in the following order:

Name (most important)

Latitude

Street Address

Phone Number

Postal Code

Locality (least important)

3. How did you avoid pairwise comparison of all venues across both datasets?

In order to avoid pairwise comparison of all possible matchings, we filtered the possible matchings using Jaro-Winkler similarity and Jaccard-distance of the restaurant names between the pairs.

First, we extracted the subset of all matching pairs for each locu id. Then, we computed the 85^{th} percentile of Jaccard distance and 15^{th} percentile of Jaro-Winkler similarity, and we removed all matchings that had lower Jaro-Winkler similarity and higher Jaccard distance. By doing so, we were able to remove 15% of the dataset and reduced the running time while improving the f1 score.