CS548 - Homework 10

Name: Apurv Upasani USC ID: 4839-3102-30 Email: aupasani@usc.edu

"I, Apurv Upasani , declare that the submitted work is original and adheres to all University policies and acknowledge the consequences that may result from a violation of those rules"

1) Metrics used:

Name

I used **Jaro-Winkler** metric. This is because in both the datasets, I found the names of the restaurants which match actually have a common prefix. Initially, I thought of using edit distance as the names would be similar. However for some cases eg. Park avenue café & park avenue café (new York city), the edit distance would be too large. However, the prefixes were the same and hence I thought I should use that information. I also assigned the **weight** of **30** for the name as I found after adjusting lot of weights that it resulted in best possible results.

Metrics	Records found	Correct records	Total Records	Precision	Recall	F-score
Base line metric **(Edit distance)**	**79**	**79**	**111**	**1.00**	**0.71**	**0.793**
Used metric **(Jaro Winkler)**	**96**	**92**	**111**	**0.96**	**0.83**	**0.89**

City

I used the **Jaro-Winkler** metric for this field for the similar reasons given above. Initially, street address seemed to be the best choice. However, I found out that in many records, the information was slightly different eg. New York City & New York. Intuitively, I thought, Q grams would should have produced good results. However, I was getting really low number of linkages, using Q-grams. I put the highest **weight** for city, **40** , as I thought it would be the most stable field to link.

Metrics	Records found	Correct records	Total Records	Precision	Recall	F-score
Base line metric **(Edit distance)**	**61**	**61**	**111**	**1.00**	**0.55**	**0.71**
Used metric **(Jaro Winkler)**	**96**	**92**	**111**	**0.96**	**0.83**	**0.89**

Address

I used **Jaro-Winkler** metric for this field. I was able to find out that most of the linked addresses have just a very small offset. But many have a common prefix. I wanted to utilize this information and hence, I used Jaro-Winkler. I assigned a **weight** of **25** for address as it had more fluctuations and special characters.

Metrics	Records found	Correct records	Total Records	Precision	Recall	F-score
Base line metric **(Street address distance)**	**69**	**69**	**111**	**1.00**	**0.62**	**0.76**
Used metric **(Jaro Winkler)**	**96**	**92**	**111**	**0.96**	**0.83**	**0.89**

Type

I used **Jaro- Winkler** metric for this field. Initially, I tried to used edit distance and Q-grams metric. However, I soon realized that some of the linkages are completely different in both datasets and would often result in false negatives. Hence, I kept a minimal **weight** of **5** for this metric and used Jaro-Winkler to just boost records which have similar or same types.

Metrics	Records found	Correct records	Total Records	Precision	Recall	F-score
Base line metric **(Edit distance)**	**84**	**83**	**111**	**0.98**	**0.74**	**0.843**
Used metric **(Jaro Winkler)**	**96**	**92**	**111**	**0.96**	**0.83**	**0.89**

Finally, I used an acceptance level of 90 to limit the output. Initially, I had set an acceptance level of 80. However, I was getting a lot of false positives, mainly due to the values generated by address and name weights.

Results

I was able to obtain a total of 96 linkages after the linkage process. After manual evaluation with Groundtruth.csv, I found that there are 92 correct linkages and 4 false positives. The false positives are due to the similarities in address and name and have same types and cities.

Record Linkage Types	Number of records
Correct Linkages	**92**
False Positives	4
Total	96