CS 4400X: Introduction to Database Systems Spring 2021

Final Project: Entity matching

Instructor: Xu Chu

1 Task description

You are given two tables (left table and right table) of electronic products. Each table is from a different shopping website. Each row in a table represents a product instance. For every pair of tuples (L_i, R_j), where L_i is a tuple in the left table and R_j is a tuple in the right table, it is either a *match* or a *non-match*. A pair of tuples is a match is they refer to the same real-world entity.

Three files are provided in data.zip: ltable.csv (the left table), rtable.csv (the right table), and train.csv (the training set). The training set contains a subset of tuple pairs, where some of them are matches and some of them are non-matches. The training set has three columns "ltable_id", "rtable_id", and "label". "label" being 1/0 denotes match/non-match.

The task is to find all remaining matching pairs like (L_i,R_j) in the two tables, **excluding those matches already found in the training set.**

2 Evaluation

Your solution will be evaluated using F1-score:

$$\text{F1} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} \tag{1}$$

where precision is:

$$\text{precision} = \frac{\text{number of predicted matches that are correct}}{\text{number of predicted matches}} \tag{2}$$

and recall is:

$$\text{recall} = \frac{\text{number of predicted matches that are correct}}{\text{number of matches in groundtruth}} \tag{3}$$

Note the evaluation will exclude the groundtruth matches already in the training set.

We will rank all solutions based on the F1-score, and assign grades for this project accordingly. The best solution will get 100/100, and any valid solution (different from the sample solution) will get 60/100. The remaining grades will be within [60,100], and will be assigned according to the distribution of the F1-score.

3 Submission

You need to submit a single project.zip file on canvas. The project.zip file should include the the following:

- A PDF file that describes the outline of your solution. In the PDF file, please provide a link to a **public github** repository that hosts your solution. We will check your repository to make sure that you have actually implemented your solution.
- A csv file that contains ONLY the matching pairs found by your system, **excluding those matching pairs already in the training set**. The csv file must have two columns: "ltable_id" and "rtable_id".

4 Sample Solution

We have provided a sample submission **CS4401X-Spring2021-Project-Sample-Solution.zip**. Of course, the sample solution is a very simple solution to get you started.

Here are some notes that might be helpful for you to improve upon the sample solution:

- The precision, recall, F1 for this sample solution is 0.017, 0.466, 0.032.
- can feature engineering/model training be improved?
- is the current way of blocking missing any matching pairs?