Homework 7: Record Linkage

Due Date: Friday, 2016/11/04 @11:59pm on Blackboard.

Summary

In this homework you will find matching records among two datasets by record linkage, and evaluate the performance of your methodology.

Task 1 – Choosing your datasets

In this HW, you are encouraged to use datasets you have extracted for your project. You can also use datasets you have extracted for previous HW if appropriate.

Note 1: If for your project you need to do de-duplication instead of record linkage, then adapt the tasks below accordingly.

Note 2: If you believe the datasets of your project is not suitable for this HW, you may use other datasets, but they should follow the same requirements as described below. (One possible dataset you may choose to use is the Fodor's and Zagat's Restaurant dataset.)

Write a program to separate every record in your dataset into different fields if it is not already separated.

You should choose two datasets that contain records that can be matched. Your datasets should contain **at least 100 records**. You should be able identify **at least 20 matching records** manually from the dataset.

For your submission, please submit the datasets named `dataset1` and `dataset2` with appropriate file extensions.

Task 2 – Examine your datasets

In your report, describe the datasets you have chosen, include the following points:

1. What are your datasets?
2. What are the fields in your datasets? Describe them.

You should also submit a textfile groundtruth.txt containing the 20 matching records you have identified in the format:

dataset1:2 dataset2:3

dataset1:3 dataset2:6

dataset1:5 dataset2:9

which, for the first row says the 2nd record in `dataset1` matches with the 3rd record in `dataset2`, and so on.

Task 3 – Perform record linkage

Use any tools of your preference or write your own code to match records among your two chosen datasets. Example tools you can use are FRIL (http://fril.sourceforge.net) and dedupe (https://github.com/datamade/dedupe). Try different similarity measures on different fields so that you can get most correct matches.

Task 4 – Describe your methodology

Write the answers to the following questions in your report file:

(a) 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.
(b) 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.

Task 5 – Evaluation

In your report file, create a table listing the 20 records you found matches manually for in dataset1, its corresponding record in dataset2, the record in dataset2 that it was matched to by your record linkage method, and whether the results match your groundtruth. Below is an example of how your table should look like. Refer to records by their line number.

Record in dataset1	Record in dataset2 (groundtruth.txt)	Record in dataset2 (result)	Matched?
2	3	3	Yes
3	6	7	No
…	…	…	…

Also compute the precision using the table you have filled in. Show your steps.

Was the performance good? If not, what do you think is the problem? (Bad performance with no reasonable justification would result in point loss for task 3-4.)

Submission Instructions

You must submit the following files (totally at least 5 files) in a single .zip archive named `Firstname_Lastname_hw7.zip` and submit it via Blackboard:

- The report PDF file `Firstname_Lastname_hw7.pdf` containing the answers to Task 2,4,5.
- The data files `dataset1 dataset2 groundtruth.txt` containing the data as specified in Task 1.
- The data file `results` containing the full results of performing record linkage on your two datasets.
- (Optional) Any other code you wrote for this HW. Also describe what they are in your report PDF file `Firstname_Lastname_hw7.pdf`

Grading

Task 1 and 2 (**40%)**: Finding appropriate datasets, description of datasets, manually creating the groundtruth consisting of at least 20 records.
Task 3 and 4 (**40%)**: Performing record linkage on your datasets, description of methodology, explain reasons for choices of similarity measures
Task 5 (**20%)**: Comparison of results to groundtruth, correct computation of precision. Comments on performance. (Bad performance with no reasonable justification would result in point loss for task 3-4.)

Ground Rules

This homework must be done individually. You can ask others for help with the tools, but the submitted homework needs to be your own work.