Australian National University

COMP3430 / COMP8430
Data wrangling

Lab 7: End to End Record Linkage

Objectives of this lab

- Today's lab is the last in a series of five labs where in the past four labs we have implemented the different steps of a complete record linkage program.

- In this lab we will be experimenting and testing how all the different parameters and choices can affect the outcomes of a record linkage project.

- Also, we will be learning how we can write the linkage outcome into a file.

Outline of this lab

- How to build a complete record linkage system
- Explore how different parameter settings and choices of techniques affect the overall performance
- Write a linkage outcome into a file
- Summary

Preliminaries

- Go back over the work from previous labs and remind yourself what we were doing and how the overall program is structured.
- **Before you begin, complete any outstanding task or implementations from the previous labs.**
- You can download the evaluation module with sample solutions in week 5 and use it with your RL program if you find difficulties implementing the required evaluation measures.

Build a complete record linkage system

- In this lab you are given with extra set of data sets for experiments. Download from Wattle the **comp3430_comp8430-reclink-lab7-datasets.zip** archive.

- The zip archive includes data sets with different sizes and quality levels (clean to very dirty).

- See if you modify the main program to run the RL program with different settings including these data sets.
- Experiment with different choices in each of the different components (blocking, comparison, and classification) and different parameter settings for thresholds, different weightings, and so on.

Write linkage result into a file

- Write the output (the record id pairs of predicted true matches) of each parameter setting and function choices into a file.
- You can use the Python program **saveLinkResult.py** which can be downloaded from Wattle to write the linkage output into a file.
- Have a look at the function **save_linkage_set()** to see what the inputs and outputs are of this function.
- Call this function from the main Python program **recordLinkage.py** to write the result into a file.

Questions to consider

- For different evaluation metrics, which parameter settings produce the best results?
- Do these best settings behave differently for the different data sets with different sizes and different data quality (corruption) levels?
- Do some of these parameter settings trade-off one evaluation metric against another?
- See the tutorial document for more questions that we recommend you to explore.

Summary

- In this lab we experimented our RL program with different parameter settings and different data sets.

- We also learnt how we can write the linkage output into a file.