The Australian National University **Research School of Computer Science, CECS**

COMP3430/COMP8430 – Data Wrangling – 2019

Lab 7: End to End Record Linkage Week 9

Overview and Objectives

We have spent the past four labs (3 to 6) constructing the different steps of a complete record linkage program, and 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.

Lab Questions

There are not really any implementation tasks for today's lab beyond making use of the record linkage system and experimenting with it. **Note that the focus of this lab is about your understanding of how different blocking, comparison, and classification techniques can be used to build a complete record linkage system; and how to select the best performing combination of different techniques based on the obtained linkage result. This will be important for the upcoming final record linkage project (for COMP3430 students only) in the Data Wrangling course.**

Please first complete any outstanding task or implementations from the previous labs. Once you are done, please download from Wattle the `comp3430_comp8430_reclink-lab7-datasets.zip` archive (in Week 9) that contains an extra set of new data sets for you to experiment with. Then please run the record linkage program on the data sets of different sizes and quality levels (clean to very dirty). Ideally modify your main record linkage program in such a way that it runs a linkage on all provided data sets in one go.

Experiment with different function choices in each of the different components (blocking, comparison, and classification). Try different parameter settings for thresholds, different attribute choices for blocking and comparison, different weightings for the classification step, and so on. Some questions you may wish to consider include:

- For each of the different evaluation metrics, which choices produce the best results (or the best results you can find)?
- Are these still the best choices for the different data sets with different sizes and different data quality (corruption) levels?
- Do some of these choices trade-off one evaluation metric against another (i.e. they produce a good result for one evaluation measure but are poor for some others)?
- How significantly does blocking improve the performance (in terms of time)? Does this become more important as the data sets get larger?
- Are there some parameter settings or functions that are worse than others for all data sets and on all evaluation metrics?
- Can you spot any patterns in the results? Are there any functions that seem to work well on different data types? Do certain parameters seem to require a particular range in order to achieve reasonable results (e.g. the similarity thresholds)? Could you use these patterns to justify your choice of functions and parameter settings in the future?

Please note that for some parameter settings and function choices, the program may be very slow, especially on the larger data sets. Please terminate a run early rather than spending the whole lab waiting for it to complete (for some choices it may not finish at all).

In addition there are some other things you may wish to use this lab for such as:

- 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 in Week 9 to write the linkage output into a file. Once you downloaded this Python file have a look at the function `save_linkage_set()` which we have provided, to see what the inputs and outputs are of this function. Then call this function from the main Python program `recordLinkage.py` to write the result into a file. Make sure you import `saveLinkResult.py` within `recordLinkage.py` before you call the function `save_linkage_set()`.
- If there are any pieces of the code you did not fully understand or complete, please have another look at them today and ask for assistance from your tutor if you would like any further explanations.
- If there were any of the extension exercises from labs 3 to 6 that you partially implemented, then this lab is another opportunity to complete them.
- If you would like to improve some of the components or measurements such as timing or the information printed out, please feel free to do so. You are welcome to customise the program as much as you desire.

Also note that you will have to make use of this program for the upcoming final project (COMP3430 students only), so please keep this in mind while you are experimenting and make sure you are comfortable with the entire program and know how everything works.