For many of the questions there is no right or wrong answer. Marks will be awarded based on your reasoning, and the justifications of your decisions and explanations.

We will endeavour to release your marks and feedback within two teaching weeks after the submission deadline. If you feel we have made an error in marking, you have two weeks following the release of marks to raise any issues with the course convener, after which time your mark will be considered final. If you request that we re-mark your project, we will re-mark the entire project and your mark may go up or down as a result.

Project Questions

For this project, we provide you with the following two data sets, dataset A.csv and dataset B.csv, as well as a truth data set true matches.csv, available for download from Wattle in week 8.

The tasks for this project are similar to what you had to do in lab 7 in week 9. You are required to run your record linkage program (including any modifications you have made to this program) on the two data sets provided, and write a report which addresses the following questions:

1. Blocking (6 marks):
    - How does blocking affect your results? Specifically, describe your choice of blocking method and choice of blocking keys. Discuss which attributes in the given data sets were useful as blocking keys and which were not, and why.
    - If there is a trade-off between performance (reduction ratio, pairs completenes and pairs quality) and the quality of the final record linkage results, where do you think the optimal balance is, and why?
    - Do you think this tradeoff would change on different data sets with different data quality levels? If so, how and why?
2. Linkage Quality (6 marks):
    - How does the record linkage quality change with the choice of parameters and techniques? Specifically, discuss and justify how you selected appropriate comparison functions for certain attributes and why these selected functions are suitable while others were not.
    - Is the record linkage quality particularly sensitive to certain parameters or choice of techniques? If so, why is this the case?
    - Provide the numerical linkage evaluation results for other (not optimal, see below) parameter settings that you have used (you only have to provide the output file for your best obtained linkage results – see next question). Ideally you include tables or plots to show linkage quality results for different parameter settings.
3. Optimal Settings (4 marks):
    - What is the best linkage quality result you are able to achieve? Why do you think this combination of parameters and techniques works well?
    - Are the results good for all evaluation metrics or only some? If the results are good for only some metrics, why do you think the results are not good for other metrics?

      In addition to answering this question in your report, you must also submit the output file which contains the linked predicted matching record pairs (as a CSV file) for the best linkage result you were able to obtain.

      Use the Python program saveLinkResult.py which we use in lab 7 to write linkage output into a file. Your submitted output file must exactly follow this CSV file format!

4. Data Quality (4 marks):
    - How dirty are these data sets relative to those you experimented on during the labs 3 to 7? Describe your impression after having conducted the linkage project.
    - How can you determine this? Describe your methodology to assess the data quality of the data sets we provided for this project (such as any calculations you used, or how you determined the data quality using data exploration and profiling).

Visualisations: You should use appropriate data visualisations such as tables, plots, etc. in your descriptions to all the above questions. Please assume you are presenting your record linkage project to an audience without a strong technical background, so make sure you adequately explain any visualisations you use (i.e. describe what tables and figures show and interpret the content of the obtained results).