COMP3430 / COMP8430 – Data Wrangling - 2021

Assignment 3 **Due 11:55pm Friday 22 October 2021**

Worth 20% of the final grade for COMP3430 / COMP8430

Last update September 21, 2021

Overview and Objectives

For this assignment you will be having another look at the record linkage program that you developed in the lab sessions. Specifically, we provide you with two **new** master data sets from where you need to generate two individual data sets to be used for this assignment. We ask you to work with the programs we have developed in the labs, and provide answers based on your findings. As with the previous assignments, the emphasis is on your understanding, descriptions, and justification as much as the raw (numerical) record linkage evaluation results that you are able to achieve.

Important

- The answers to this assignment have to be submitted online in Wattle, see the link **Assignment 3 Submission** in week 11 (18 to 22 October).
- Follow instructions given for maximum text length in free format answers. If your answer is too long it will attract a penalty (for details see the individual questions below and the corresponding answer submission forms in Wattle).
- You can edit your answers many times and they will be saved by Wattle.
- Make sure you submit the **final version** of your assignment answers **before the submission deadline**.
- Note that **Wattle does not allow us to access any earlier edited versions of your answers, so check very carefully what you submit as the final version!**

You can only submit your assignment once!
Make sure you do not forget to submit your assignment!

Penalties

Textual questions have maximum line and maximum word limits. If you write more than these provided limits we will have to apply an over-word-limit penalty. For details of limits see the individual questions below and the corresponding pages in the assignment submission in Wattle.

You will receive a 20% penalty if you failed to submit the assignment before it is due.

Deadlines, Extensions and Late Submissions

The assignment is due 11:55 pm on Friday 22 October 2021.

Students will only be granted an extension on the submission deadline in extenuating circumstances, as defined by ANU policy (http://www.anu.edu.au/students/program-administration/assessments-exams/deferred-examinations).

If you think you have grounds for an extension, you must notify the course convener as soon as possible and provide written evidence in support of your case (such as a medical certificate). The course convener will then decide whether to grant an extension and inform you as soon as practical.

In accordance with the CECS and ANU late submission policy, **no late submissions will be accepted**, except where an extension has been approved by the course convener.

Assignment Structure

The assignment consists of four (4) tasks as described below which can be worth different numbers of marks. Make sure you answer all aspects of each task.

If you have any questions on the assignment please post them on Wattle – **however do not post any partial solutions, program codes, equations, calculations, URLs, etc. or any hints on how to solve any of the assignment tasks.**

Plagiarism

No group work is permitted for this assignment.

We do encourage you to discuss your work, but **we expect you to do the assignment work by yourself**. If you are unsure about what constitutes plagiarism, **make sure you carefully read the ANU Academic Honesty Policy** (http://academichonesty.anu.edu.au/).

If you do include ideas or material from other sources, then you clearly have to make attribution by providing a reference to the material or source in your submitted assignment answers. We do not require a specific referencing format, as long as you are consistent and your references allow us to find the source, should we need to while we are marking your assignment.

Marking

This assignment will be marked out of 20, and it will contribute 20% of your final course mark.

Note that not all tasks and questions are equally difficult. For some of the tasks there is no single right or wrong answer. Marks will be awarded based on your reasoning and the justification of your decisions and explanations, as well as clarity and correctness of writing.

IMPORTANT: We do not accept any type of code, screenshots, or external links as answers. Please do not waste the space given to you to provide answers by writing external links or code in that space. We will not mark such answers and you will lose marks if the correct answer is not inside the text fields.

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 assignment, we will re-mark the entire assignment and your mark may go up or down as a result**.

Data Set Generation for this assignment

As with the previous two assignments, for this assignment each of you will again work on individual data sets that will be based on a pair of new **master data sets** as well as a **ground truth data set** we will provide, and a new **data generation program** we will also provide.

Note that we have generated the master data sets based on real data (such as lookup tables of names, addresses, etc.), and we have then corrupted and modified certain aspects of these data sets. We have intentionally tried to include the types of relationships, features, errors, and other data quality issues that you might find in real data sets. **Any similarity to real persons or places is entirely coincidental**.

Download the two master data sets and the ground truth data set from Wattle (to be made available in week 6) named **dw_assignment_master_rl1.csv.gz**, **dw_assignment_master_rl2.csv.gz**, and **dw_assignment_master_rlgt.csv.gz**, and the new data generation program named **generate-student-datasets-rl.py**. Copy all these files into one folder / directory, and run the code using Python 3 in the following way:

```
python3 generate-student-datasets-rl.py your_ANU_ID
```

The program will generate two data sets named **data_wrangling_rl1_2021_your_ANU_ID.csv** and **data_wrangling_rl2_2021_your_ANU_ID.csv**, as well as a corresponding file that contains the true links (ground truth data), named **data_wrangling_rlgt_2021_your_ANU_ID.csv**, and print some output which contains the following important lines (for example with the ANU ID u1234567):

```
$ python3 generate-student-datasets-rl.py u1234567

Your two student data sets and the ground truth data set for the data wrangling 2021 assignment 3
have been generated and written to:

    data_wrangling_rl1_2021_u1234567.csv
    data_wrangling_rl2_2021_u1234567.csv
    data_wrangling_rlgt_2021_u1234567.csv

  Your ANU ID check code is:             d76225bc
  Your student data sets check code is:  64z47ue6c5

  *** Check this pair of codes is in the list provided on Wattle, if not contact the course convenor.
```

Important

- **Write down your two check codes because you must provide them with the assignment submission.** This will allow us to validate that you have generated and used the correct data set.
- **Check that the pair of check codes** (like in the example above d76225bc and 64z47ue6c5) **is in the list of check codes we will provide on Wattle** (in week 6 under the Assignment 3 document). This will allow you to check that you have generated the correct data sets.

- Note that the check codes are different for the data sets you generated for the previous two assignments and this new pair of data sets.
- **You must use your individual generated pair of data sets, data_wrangling_rl1_2021_your_ANU_ID.csv and data_wrangling_rl2_2021_your_ANU_ID.csv, and the corresponding ground truth data set, data_wrangling_rlgt_2021_your_ANU_ID.csv (generated based on your ANU ID), for all tasks of this assignment**.

Assignment Tasks

The tasks for this assignment are similar to what you had to do in lab 3-7 from weeks 6-10. You are required to run your record linkage program (including any modifications you have made to this program) on your individual two data sets and the ground truth data set (generated as described above), and address the following questions:

Task 1: Blocking (6 marks):

(a) How does blocking affect your results? Specifically, describe your choice of blocking method and choice of blocking keys. Discuss which attributes and/or attribute combination(s) in the given data sets were useful as blocking keys and which were not, and why.

(b) If there is a trade-off between performance (reduction ratio, pairs completeness, and pairs quality) and the quality of the final record linkage results, where do you think the optimal balance is, and why?

(c) Do you think this trade-off would change on different data sets with different levels (both low and high) and characteristics of data quality? If so, how and why?

Write a maximum of 20 lines of text (around 500 words) in total in the corresponding answer field on Wattle. Clearly indicate your answers to (a) to (c).

Task 2: Comparison and Classification (6 marks):

(a) How do different comparison techniques affect linkage results? Discuss and justify how you selected appropriate comparison functions for different attributes, and why these selected functions are suitable while others are not.

(b) How do different classification techniques using different parameter settings affect linkage quality? Discuss and justify how you selected an appropriate classification technique and corresponding parameter settings to obtain high linkage quality and why other classification techniques are not suitable.

(c) As discussed in the lectures in week 8, for suitable linkage quality measures, describe how the final record linkage quality changes with the choice of different parameters and techniques? Is the record linkage quality particularly sensitive to certain parameters, or choice of comparison or classification techniques? If so, why is this the case?

(d) Are there any evaluation measures that are not useful? Describe why these measures are not useful in evaluating the performance of a record linkage project.

(e) 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 task).

Write a maximum of 20 lines of text (around 500 words) in total in the corresponding answer field on Wattle. Clearly indicate your answers to (a) to (e).

Task 3: Optimal Settings (4 marks):

(a) What is the best linkage quality result you are able to achieve, both in the blocking and the classification steps? Why do you think this combination of parameters and techniques worked well for your data set pair?

(b) Are the results equally good for all evaluation measures discussed in the lectures in week 8, or only for some? If the results are good only for some measures, why do you think the results are not good for other measures?

Write a maximum of 250 words (around 10 lines of text) in the corresponding answer field on Wattle. Clearly indicate your answers to (a) and (b).

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

You must 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! We will use a program to check linkage quality using this file to validate what you write in your answers in Wattle. If our program does not work with your submitted file because it does not follow the required file structure then you will loose marks.

Task 4: Data Quality (4 marks):

(a) How dirty are these new data sets you generated for this assignment compared to all the data sets you have worked with in labs 3 to 7? Describe your impression after having conducted the linkage on the different data sets used in the labs.

(b) How can you determine this? Describe the methodology you used to assess the quality of the data sets we provided for this assignment and compare it against the quality of the datasets from labs 3 to 7 (such as any calculations you used, or how you determined the data quality using data exploration and profiling).

Write a maximum of 250 words (around 10 lines of text) in the corresponding answer field on Wattle. Clearly indicate your answers to (a) and (b).

Marking:
For each of the tasks described above you will receive up to the shown mark for appropriately answering the corresponding questions, and describing and justifying what you have done.

For Task 3, we will also compare your answers to the numerical results we obtain from your submitted file of linked records. You will loose marks if the numerical results we obtain differ from what you describe in your textual answers.

English writing mistakes and typographical errors will attract small penalties.