

**City, University of London in MSc Data Science**

**PROJECT REPORT**

**2022**

**TITLE OF THE THESIS**

Zacharias Detorakis ([zacharias.detorakis@city.ac.uk](mailto:zacharias.detorakis@city.ac.uk))

Supervised by: Ernesto Jiménez-Ruiz ([ernesto.jimenez-ruiz@city.ac.uk](mailto:ernesto.jimenez-ruiz@city.ac.uk))

*By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.*

*Signed:*

**Abstract**

This project aims to enhance a set of data given as input (e.g. tabular data) with semantic meaning using existing Knowledge Graphs (KG) as reference. The approach that has been implemented is inspired by ColNet, a system implemented as part of the SemTab challenge, and comprises a series of modules for i) parsing the tabular data ii) using the (cell) values to identify candidate entities with lexical similarity and their candidate KG classes iii) training a set of binary cnn classifiers (one per class) and then employing different ways of predicting the class of each column using the above. A pretrained word2vec model was used to train the cnns and transform the column values to inputs for predication.

WILL WE DO A CEA?

The results indicate that…

Table of Contents

[1 Introduction and Objectives 4](#_Toc87024741)

[1.1 s 4](#_Toc87024742)

[2 Context 4](#_Toc87024743)

[3 Methods 4](#_Toc87024744)

[4 Results 5](#_Toc87024745)

[5 Discussion 6](#_Toc87024746)

[6 Evaluation, Reflections, and Conclusions 6](#_Toc87024747)

# Introduction and Objectives

## Objective 1

***Chapter 1*** − ***Introduction and Objectives:*** This chapter should set the scene for the reader. It must outline the background to the problem, give your reasons for the choice of project, and identify the project’s beneficiaries. Your objectives need to be precisely stated, together with the tests that will show, at the end of the project, that they have been met (or not been met). You need also to outline your methods in broad terms, along with y work plan with sufficient detail to show how you planned to meet the objectives. Outline any major changes of goals or methods that happened during the project. Finally, outline the structure of the report, showing how it fits together.

# Context

***Chapter 2*** − ***Context:*** This chapter explains the current state of your topic, in practice and theory. This is the state of the world which you intend to improve, and the state of knowledge on top of which you build your advances and from which you learn knowledge to apply and constraints on your work. So, you will report and analyse what is known about a certain topic, as reported in reference literature and published scientific literature; if you are developing a product, you will need to report about comparable or competing products over which you intend to improve or from which you will obtain ideas; you may need to describe legal or societal situation within which your work takes place; etc.

It is important to demonstrate scholarship, *i.e*. the ability to read about a subject area in a range of sources, assimilate the material and then discuss it intelligently.

You should demonstrate that you understand what you have read by providing some analysis or commentary in view of the goals of your project: it is not enough simply to provide summaries of what you have read. References should be cited following the Harvard Referencing Style. **You must also explain, both in this chapter and, as appropriate, in others, how the results of the studies to which you make reference inform your project work.** To gain a passing grade, your report MUST demonstrate adequate engagement with academic literature and any other sources necessary for the work to be well informed**.**

# Methods

***Chapter 3*** − ***Methods:*** This chapter describes in detail the methods for whatever activities were necessary for your project – e.g., data gathering, data analysis, requirements analysis, design, implementation, testing/evaluation, *etc*. Your choice of methods should be discussed and justified in view of the project objectives, and with reference to the pertinent literature. Report not only what methods you applied in generic terms, but what you actually did: sufficient information about dates and details for your reader to understand how you ran *your* project, rather than just how one *could* run any similar project.

Report in this chapter what you did, not what you produced or found as a result (which goes under Results).

Note: only use the word ‘methodology’ if you know what it means!

# Results

***Chapter 4*** − ***Results:*** This chapter presents the outputs that you produced, by applying the methods that you have selected, including e.g. analysis, design, prototyping, experimental work, evaluation, *etc.*

How you report these results will depend on the nature of the work. It may be helpful to divide them into basic data (*e.g.,* for a project that developed a software product, requirements specification, test data, *etc*.) and analysis of the data (*e.g*. statistical analyses, evaluation analyses, *etc*.). Remember that you are informing the reader of what you have produced and found and emphasising the interesting parts, so summarising at the end of each major section is useful.

It is usually very helpful for the readers to include graphics and diagrams, for instance to clarify software design or requirements, identify key trends and relationships in empirical data, etc. If you do so, be sure to refer to these figures in the text and use them as evidence to support what you are explaining or arguing; and be sure that your figures are well designed and clearly presented – do not just use default settings of the software you are using in producing them.

It is essential that you identify clearly what you accomplished or produced yourself, as opposed to what existed before you started your individual project or was provided by others. For instance, some projects build new software on top of an existing code base, add new data to an existing body of data, or are executed by a student as a member of a team. It is essential to indicate what parts of the activities and results which you report are your own work. If this is left unclear, the markers are instructed not to give credit for work that they cannot attribute to you. Ambiguity would attract penalties for poor academic practice, with delays caused by any investigation (deception would be treated as academic misconduct, of course, which may lead to expulsion).

# Discussion

***Chapter 5*** − ***Discussion:*** This chapter examines your results in comparison with your objectives, and then in the wider perspective of other theoretical and applied work relevant to your project, as covered in your review in Chapter 2. For instance, for a software product you will discuss how well it satisfies the user needs that it addresses, its performance and dependability, aspects of design, implementation or assessment that have proved good choices or that instead you would change if you were to repeat the project knowing what you now know. For novel research results or any other knowledge obtained through the project, you will discuss your confidence in the results, their validity, scope and their generalisability. What are the implications of what you have found out? Do you have any recommendations as a result?

# Evaluation, Reflections, and Conclusions

***Chapter 6*** − ***Evaluation, Reflections, and Conclusions***: This chapter should evaluate the project work as a whole. Here the original choice of objectives, the literature examined, the methods used, the planning, *etc.* are all reviewed to see what has been achieved by undertaking the project. There may be a summary of general conclusions drawn from the work done, highlighting the particular contribution of your project. You should also consider the implications of these conclusions. Discuss any proposals that you might make for further work, having discovered what you now know*.* It is also important to include a reflective section covering what you have learned from the project process. What would you do differently if you were to start again, knowing what you now know? **Your report MUST include adequate Evaluation, Reflections and Conclusions to gain a passing grade.**

***Glossary:*** If required, the glossary defines specialist terms that are not likely to be known by your intended audience.

***References****:* A full list of all references cited in the project report. Citations **must** follow the Harvard Referencing Style.

***Appendices and additional files:*** Any material that would interrupt reading the report must be presented here.

All evidence that allows the markers to assess your project **must** be available to the markers, in the appendices or additional files, clearly identified for the markers. Material that must be submitted if produced in your project includes:

• interview records; questionnaires and questionnaire replies;

• for any software developed: • routine design documentation

• source code, in a structured and readable format, properly commented;

• test inputs and results; output listings; displays, *etc*.;

• whenever feasible, what is needed for makers not only to examine the code but also build the software and/or test it, as the may consider necessary: ready-to-install software, installation guide & user guide; for web software, a URL where markers can test it

• any data collected or produced;

• wireframes and prototypes

• for any experiments or data analyses performed, details of method that were not provided in the main text, input data sets, documentation of the analyses performed on the data, processed data and outputs;

• anything else that may be important for your examiners to assess your project: evaluate complexity, challenge, work done, results and your contribution in producing all this.

Some kinds of appendix material can or should be made available online without submission: e.g. massive datasets, audio/video recordings, long source code listings. As you write up your report, your supervisor (referring if necessary to the module leaders) can answer specific doubts about what should be included and formats.

The principle behind these rules is that you **must** give your markers full access to all the information behind your dissertation. There may be special cases in which there are better ways to grant this access than large appendices. For instance, your supervisor may have advised you to store the data on City's permanent repository, https://city.figshare.com/, and you will then indicate in your report the URL where the markers have access to the data. If you think that for some appendices you should grant access outside an online or offline submission, seek your supervisor's advice; always make sure that (1) the markers can find and examine such material easily and quickly, at any time until your degree is awarded, and (2) that there is proof that the material was produced before the submission deadline.

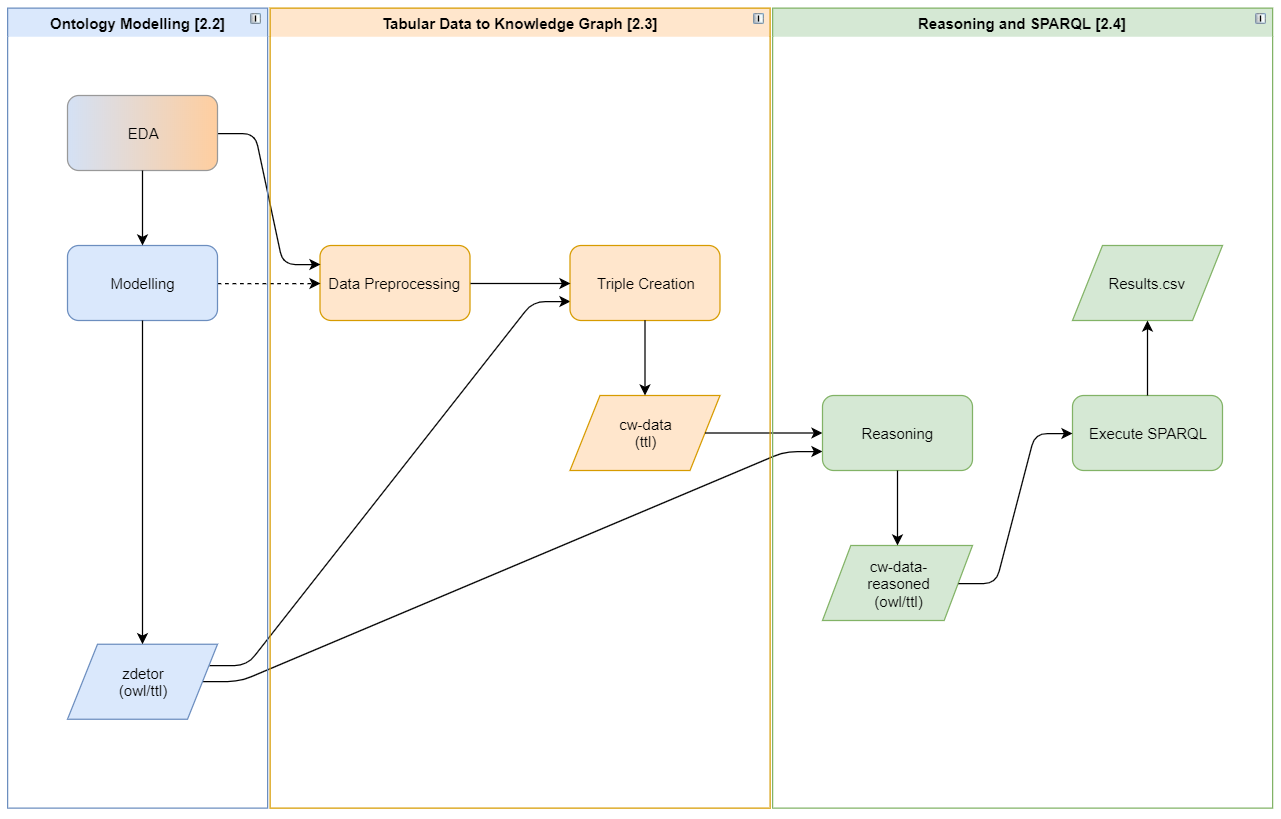


Figure 1. Process of modelling and converting tabular data to a KG

|  |  |  |
| --- | --- | --- |
| **name** | **address** | **Restaurant\_name** |
| Bertucci's | 2929 Berlin Tpke | Bertucci's\_\_1 |
| Bertucci's | 194 Buckland Hills Dr | Bertucci's\_\_2 |
| Bertucci's | 2847 State Route 35 | Bertucci's\_\_3 |

Table 1. Examples or restaurant names reused for different restaurants