Literature Review

CM4105 – Honours project

Cameron Urquhart - 1906014

2020

Contents

1. [Introduction 2](#_Toc57987777)

[1.1 Background 2](#_Toc57987778)

[1.2 Project Aims 2](#_Toc57987779)

[1.3 Definitions 2](#_Toc57987780)

[1.4 Target Audience 3](#_Toc57987781)

[2. Background Research 4](#_Toc57987782)

[2.1 Systematic Reviews 4](#_Toc57987783)

[2.1.1 Steps Involved in a Systematic Review 4](#_Toc57987784)

[2.1.2 Meta-Analysis 5](#_Toc57987785)

[2.2 Short Text Clustering 5](#_Toc57987786)

[2.3 Citation Graphs 5](#_Toc57987787)

[2.4 Document Similarity Comparisons 6](#_Toc57987788)

[2.4.1 Jaccard Similarity 6](#_Toc57987789)

[2.4.2 Cosine Similarity 6](#_Toc57987790)

[3. Existing Work 7](#_Toc57987791)

[3.1 Graph Representations to Analyse Publications Connectivity in Childhood Obesity Interventions 7](#_Toc57987792)

[4. Technologies 8](#_Toc57987793)

[4.1 Data Collection 8](#_Toc57987794)

[4.2 Database 8](#_Toc57987795)

[4.3 Graph Database 8](#_Toc57987796)

[4.3.1 Neo4J 8](#_Toc57987797)

[4.4 Natural Language Processing 8](#_Toc57987798)

[4.4.1 Natural Language Toolkit 9](#_Toc57987799)

[4.4.2 spaCy 9](#_Toc57987800)

[4.5 User Interface 9](#_Toc57987801)

[5. Conclusion 10](#_Toc57987802)

[5.1 Plan 10](#_Toc57987803)

[5.2 Issues 10](#_Toc57987804)

[5.2.1 Legal issues 10](#_Toc57987805)

[5.2.2 Ethical & Social Issues 10](#_Toc57987806)

[6. References 11](#_Toc57987807)

# 1. Introduction

## 1.1 Background

In the developed world, childhood obesity has progressively become a more significant issue. Consequently, there has been a significant increase in research interest regarding preventing and dealing with childhood obesity cases. Obesity is a condition defined by the WHO as a state of “abnormal or excessive fat accumulation that presents a risk to health”. Obesity is a significant risk factor for a wide variety of conditions such as heart disease, stroke, diabetes, and some cancers. Childhood obesity brings a risk of premature onset of obesity related ailments, and a likelihood of the obese state being maintained into adulthood. [1]

Health workers and researchers aim to better help those affected by childhood obesity by staying up to date on the latest research in the field. However, with such a wealth of information available on the subject, it can be restrictively time consuming for individuals or groups to try to parse through existing research. Data visualisation technologies could be highly beneficial in allowing workers in the health domain gain an understanding of the relationships between a wide volume of information, and in allowing them to more efficiently review the information present in research papers.

## 1.2 Project Aims

This project aims to serve as an investigation into the organisation, visualisation, and analysis of childhood obesity interventions to automate systematic reviews. The aims are:

* To utilise Natural Language Processing technologies to establish relationships between a dataset of scientific papers and to store this data in a graph database.
* To utilise the data in the graph database to generate visualisations to enable a broad understanding of the data in the dataset.
* To create a unified tool, which uses an underlying database to catalogue the dataset, allowing end users to add and remove papers, search for specific papers, and to generate the visualisations to better understand the data organised by the tool.

## 1.3 Definitions

Here, various terms used through the document will be defined.

Natural Language Processing (NLP) – a subset of machine learning where algorithms are used to manipulate and understand natural language. [2]

Database – a structure for the storage and organisation of data, usually containing several tables of data with relationships to one another.

Graph Database – a program to store, query and modify graphs.

Web Scraping – the use of programs to collect files and data from web pages.

Evidence Based Synthesis – a process of combining data from several studies to come to an overall conclusion about the research.

User Interface – The part of a program that the end user interacts with, with buttons, search bars, etc.

## 1.4 Target Audience

The target audience for this project are those working in the health sciences domain. Researchers and medical staff working in health sciences will seek to be as informed as possible on all new research and developments in the field. However, with so much research being done, it becomes very inefficient for individuals or groups to manually go through existing papers to try and draw broad conclusions. This can result in huge time wasting by researchers, potentially cutting into tight budgets and reducing the time that can be spent on the research itself.

This project, should it work as intended, could go a long way in reducing time spent by researchers. If they’re able to enter a set of research papers into a system, and have it perform analysis and show them relationships and connections between paper, this could significantly increase the speed of their work, lifting the burden that is manually going through papers to perform reviews.

# 2. Background Research

## 2.1 Systematic Reviews

Systematic reviews use analytical methods to collect data about sets of research papers, allowing for broad analysis of the research papers in question. These are essential in the health science domain, as there is such a massive volume of research available and are used by researchers to gain a consensus on what results exist.

Traditionally, when researchers wanted to summarise research evidence from a set of research papers, they would use *narrative reviews* [3,4]. A narrative review is when researchers collect a number of research papers on the same topic, the topic being their review question, and parse through the abstracts of the papers to ensure the papers selected are relevant to the topic they are examining, removing the ones that appear unrelated. They then create a summary of the documents they selected for the review. [5]

Systematic reviews are reviews wherein repeatable analytical methods are utilised to collect data and analyse it. Systematic reviews aim to identify all relevant evidence, select studies for inclusion, analyse the quality of the studies, synthesise the findings, and produce a balanced summary of the data found across the set of papers. [3]

Narrative reviews are often inconsistent in their results. A review paper of Dr Cipriani and Dr Geddess contained 7 narrative reviews, and these narrative reviews reached varying conclusions despite being based on the same set of articles. The conclusion to be drawn here is that while narrative reviews are based on evidence, they are not valuable as scientific evidence. Due to true systematic reviews being conducted using the same thorough steps each time, their results are more useful in terms of understanding the research that has been done on a given topic[6].

### 2.1.1 Steps Involved in a Systematic Review

The steps involved in carrying out a systematic review, as outlined in the book *What is a systematic review?* [3] are as follows:

1. The creation of relevant review questions

This is the creation of one or more review questions relevant to the topic being reviewed. It’s important that these are concise as they will be used as the basis for colleting the literature for review in the next step.

1. Collection of relevant literature

Using scholarly search engines, such as Google Scholar and Microsoft Academic, with keywords relating to the review questions, to retrieve a set of research papers. It’s important that a variety of sources are used to ensure as many relevant documents are collected as possible.

1. Assessing the literature

The literature collected in the previous step must then be analysed to ensure they are appropriate for the review. The reviewer(s) must exclude papers which do not have a high degree of relevance to the review questions, but also not be extremely strict with this process, as it could result in too few papers for the review.

1. Compiling the results

In this step, the results from the set of publications are extracted and analysed to determine how similar the results are across the documents.

1. Contextualising the findings

In this step, the results are summarised and contextualised to grant an understanding of the results of the review to those reading it.

### 2.1.2 Meta-Analysis

Meta-Analysis is something that often comes up in systematic reviews but is not required for a systematic review to be carried out [7]. Meta-analysis is a statistical technique for developing conclusions from the set of documents being reviewed. Meta-analyses are based on the systematic reviews performed on the documents, being used to provide a summary of the results.

## 2.2 Short Text Clustering

Text clustering is the use of cluster analysis to analyse text-based documents [8]. Short text clustering is when text clustering is applied to texts with a small volume of words. Short text clustering can prove challenging as most words in small bodies of text will appear infrequently. Consequently, it would be preferable to apply text clustering to a large volume of documents if the texts being clustered are short in size.

## 2.3 Citation Graphs

Citation graphs are a means of graphically representing the relationships between documents in terms of their citations. Citation graphs utilise vertices, representing the documents being analysed, and lines connecting them, which represent citations between the documents. These can be valuable tools in helping to understand visually the relationships between a variety of documents.

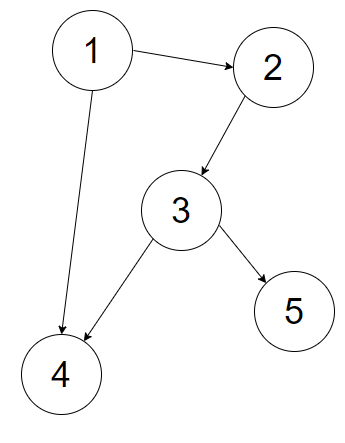


Figure 1 - Citation Graph Example

In the example above, *figure 1*, the arrow heads represent which document cited which. For example, document 4 cited documents 1 and 3.

Problems can arise when trying to programmatically generate citation graphs. Due to possible variations in how references are written between the documents, a citation graph generated automatically may not recognise that two papers had cited the same document, due possibly to the references being written to different standards, punctuation differences, or other small errors.

## 2.4 Document Similarity Comparisons

### 2.4.1 Jaccard Similarity

Jaccard similarity is one approach for measuring the similarity between documents. Firstly, the words in the documents are lemmatized, meaning they’re reduced to their root words (e.g. “dancing” becomes “dance”). Then, the count of the intersecting words is divided by the size of the union of texts [15]. This exists as a potentially useful measure of similarity, but it has issues. Jaccard similarity has no way to consider the meaning of a body of text, and as such two bodies of text could convey the exact same thing but have 0 intersection of words. It also has the issue that as the sizes of the documents increase, the number of common words is likely to expand, regardless of whether the documents are focused on the same topic or not.

### 2.4.2 Cosine Similarity

Cosine similarity is another approach for measuring the similarity between documents. Cosine calculates the cosine of angle between two non-zero vectors to get the similarity. Cosine similarity has some benefits when compared to other measures such as Euclidean Distance. In Euclidean Distance, if one document has a word 5 times, and the other has it 100 times, the similarity would not be great, but under cosine similarity, there would still be a significant similarity figure.

# Existing Work

## Graph Representations to Analyse Publications Connectivity in Childhood Obesity Interventions

Vinod Kumar Karimilla undertook research into utilising graphical visualisations to analyse the relationships between research papers on childhood obesity interventions. Vinod successfully utilised frameworks such as NLTK (Natural Language Toolkit) and Neo4J (a graph database) to plot the relationships between the documents.

Vinod’s work however did have some issues. Firstly, his work did not result in a single unified system. The different elements of his research, such as the graph database and the NLTK elements were disconnected, meaning that if another researcher wanted to attempt to use his frameworks it would be awkward and complicated to achieve. As such, I have made it an aim of this project to have everything in a single unified tool.

The work done by Vinod also lacked in testing and evaluation. Testing performed on the systems he developed was limited and the evaluation overall on what he’d done was limited. Consequently, I aim to perform very thorough testing of every element of the unified tool, and to perform a deep evaluation of the performance of the individual elements, and of how they work together as a whole.

# 4. Technologies

## 4.1

In order to have a large set of research papers for use within the project, many will have to be collected from various internet sources. This can be done easiest by acquiring the research papers through scholarly search engines. There are many search engines focusing on research papers, such as Google Scholar, Microsoft Academic Search, and WorldWideScience [10]. According to *Science Mag*, as of 2014, Google Scholar alone had indexed around 160 million results (±10%) [11]. Access to such tools enables collection of a significant volume of documents related to the project aim. It is important that the documents are collected from a wide variety of sources, as this is important in ensuring systematic reviews can reach valid conclusions.

## 4.2 Database

Databases are data structures used to make large sets of data easy to browse and interact with. Most databases have data structured into tables. The columns represent different attributes and each row is a data item. For use in this project, a database could be essential for the cataloguing of the stored research papers, making it possible for papers to be browsed, added, and removed through the front end of the unified tool. Relational databases are interacted with using various variants of SQL (Structured Query Language). This can streamline the implementation of functionality such as keyword searching and sorting alphabetically, or by characteristics such as the dates the documents were published. As an example, the database for this may have a table named *papers*, and columns may include *paper\_name*, *researchers*, *date\_published*, *publisher*, etc. If an end user wanted to view all papers published by only a certain publisher, they could easily filter the results from the database.

MariaDB is a free to use, open source relational database. It’s a fork of MySQL, one of the most popular relational database solutions, and is developed by the original developers of MySQL. As MariaDB is designed for “speed, reliability, and easy of use”, it seems a fitting choice for this project [9].

## 4.3 Graph Database

Graph databases are a type of database used to store complex data structures for use in visualisations [11]. Due to the complex nature of the data produced by reviews, graph databases are an appropriate storage solution, and the data in a graph database can then be drawn on to produce visualisations. There exist a variety of graph database solutions.

### 4.3.1 Neo4J

Neo4J is an open-source, free to use graph database solution [12]. Neo4J graph databases are queried using Cypher, a graph query language originally designed for use with Neo4J. Cypher was designed to work like SQL, being simple and efficient, but being structured in a way to work well with the complex data structures found within graph databases.

## 4.4 Natural Language Processing

Natural Language Processing refers to any kind of computer manipulation of natural language samples. This could be something simple, like spell checking within Microsoft Word, or something more complex, like smart assistants such as Google Assistant interpreting spoken commands and executing them [13].

### 4.4.1 Natural Language Toolkit

Natural Language Toolkit (NLTK) is one of the longest running natural language processing frameworks which is still maintained and used. There exist a variety of modules covering many natural language processing tasks, such as string processing, collocation discovery, classification, and semantic interpretation [13]. NLTK was designed with 4 goals in mind:

* Simplicity – to provide a framework which is simple to use and avoids unnecessary convolution.
* Consistency – to provide a framework with consistent naming and structuring.
* Extensibility – to provide a framework within which new modules can be easily introduced.
* Modularity – to provide a framework wherein individual components can be used without a requirement of understanding of the rest of the toolkit.

Talk more specifically on the capabilities of NLTK.

### 4.4.2 spaCy

spaCy is an open source Python package for Natural Language Processing. It stands out among others due to its flexibility and multi-functionality [14]. spaCy exists as a viable alternative to NLTK, with it having much of the same functionality available, and a wide variety of documentation and instruction available online.

## 4.5 User Interface

The work done previously by Vinod Karimilla was a valuable piece of research in using graphs to model the relationships between documents, however Vinod’s work did not result in a unified, user-friendly tool at the end. The creation of such a unified tool could allow researchers to, in one unified interface, browse documents in the system, add and remove documents, and generate visualisations to model the relationships between the documents in the system.

A good approach for such an interface could be to make it browser based. By using HTML, CSS, and JavaScript, an interface system can be created which has the capability to run in modern web browsers. By creating a system which resembles a modern website, most end users will intuitively understand how to navigate the interface, through interactions such as buttons and text fields.

# 5. Conclusion

Overall, it seems that there is much that can be utilised for this project. Through gaining an understanding of systematic reviews and meta-analyses, I can see how important these are and how beneficial it could be to develop a framework for automating them. It can be seen how measures such as Jaccard similarity can be used to measure the relationships between documents and how visualisations such as citation graphs can be used to create a visual representation of the relationships between the documents.

## 5.1 Plan

The plan for the development in this project will be as follows

* To obtain the set of research papers from a variety of sources.
* To review and select which tools and frameworks will be utilised to complete this project.
* To develop the automation of systematic review and meta-analysis utilising the selected Natural Language Processing framework(s).
* To develop the framework for the visualisations of the results to be generated using the selected graph database solution.
* To unify the aforementioned aspects into a singular tool, which utilises a database to hold information about the papers in the system, allowing users to browse, add, and delete documents.
* To thoroughly test each individual aspect of the project and the unified tool to remove as many issues as possible.
* To perform a thorough evaluation of the tools developed to examine their performance and effectiveness.

## 5.2 Issues

### 5.2.1 Legal issues

Throughout this project, any work done by others will be properly referenced. Where required, appropriate permission will be acquired for the use of a work.

### 5.2.2 Ethical & Social Issues

To ensure that any researchers who could end up using such a tool in the future do not end up with inaccurate or misleading information, the tool will be thoroughly evaluated in all aspects to ensure that the information being generated and shown by the tool is accurate and representative.

# 6. References

[1] - World Health Organisation. *Obesity.* Available: https://www.who.int/health-topics/obesity#tab=tab\_1. Last accessed 28th Nov 2020.

[2] - Jason Brownlee. (2017). *What is Natural Language Processing?.* Available: https://machinelearningmastery.com/natural-language-processing/. Last accessed 28th Nov 2020.

[3] - Hemingway, P. & Brereton, N (2009). *What is a systematic review?*. Hayward Medical Communications.

[4] - Uman, L. (2011). Systematic Reviews and Meta-Analyses. *J Can Acad Child Adolesc Psychiatry*.

[5] - Demiris, G & Oliver, D & Washington, K (2019). *Behavioral Intervention Research in Hospice and Palliative Care*. Academic Press. Cambridge, Massachusetts.

[6] - Pae, C. (2015). Why Systematic Review rather than Narrative Review?. *Psychiatry Investig*.

[7] - Crombie, I (2009). *What is meta-analysis?*. Hayward Medical Communications.

[8] - Editorial Team. (2018). *What is Text Clustering?.* Available: https://insidebigdata.com/2018/07/26/what-is-text-clustering/. Last accessed 1st Dec 2020.

[9] - Saive, R. (2020). *What is MariaDB? How Does MariaDB Work?.* Available: https://www.tecmint.com/what-is-mariadb-how-does-mariadb-work/. Last accessed 1st Dec 2020.

[10] - Heinrich, A. (2017). *15 Educational Search Engines College Students Should Know About.* Available: https://www.rasmussen.edu/student-experience/college-life/15-educational-search-engines/. Last accessed 1st Dec 2020.

[11] - Heddings, A. (2020). *What Is a Graph Database, and What Is It Used for?.* Available: https://www.cloudsavvyit.com/2185/what-is-a-graph-database-and-what-is-it-used-for/. Last accessed 2nd Dec 2020.

[12] -Bruggen, R (2014). *Learning Neo4j*. Birmingham: Packt Publishing Ltd.

[13] - Bird, S; Klein, E; Loper, E (2009). *Natural Language Processing with Python*. Sebastopol: O'Reiley Media Inc.

[14] - Joshi, P. (2020). spaCy Tutorial to Learn and Master Natural Language Processing (NLP). Available: https://www.analyticsvidhya.com/blog/2020/03/spacy-tutorial-learn-natural-language-processing/. Last accessed 2nd Dec 2020.

[15] - Sieg, A. (2018). *Text Similarities : Estimate the degree of similarity between two texts.* Available: https://medium.com/@adriensieg/text-similarities-da019229c894. Last accessed 3rd Dec 2020.