UNIVERSITAT OBERTA DE CATALUNYA
UNIVERSITAT POLITÈCNICA DE CATALUNYA

MASTER THESIS

Automated similarity detection: identifying duplicated requirements

Author:
Quim MOTGER

Supervisor:
Cristina PALOMARES

A thesis submitted in fulfillment of the requirements for the Master in Informatics Engineering

in the

Universitat Oberta de Catalunya
Artificial Intelligence Department

November 4, 2019

Declaration of Authorship

I, Quim MOTGER, declare that this thesis titled, "Automated similarity detection: identifying duplicated requirements" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

UNIVERSITAT OBERTA DE CATALUNYA UNIVERSITAT POLITÈCNICA DE CATALUNYA

Abstract

Faculty Name
Artificial Intelligence Department

Master in Informatics Engineering

Automated similarity detection: identifying duplicated requirements

by Quim MOTGER

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

List of Abbreviations

LAH	**L**ist **A**bbreviations **H**ere
WSF	**W**hat (it) **S**tands **F**or

Chapter 1

Chapter Title Here

1.1 Welcome and Thank You

Welcome to this LaTeX Thesis Template, a beautiful and easy to use template for writing a thesis using the LaTeX typesetting system.

If you are writing a thesis (or will be in the future) and its subject is technical or mathematical (though it doesn't have to be), then creating it in LaTeX is highly recommended as a way to make sure you can just get down to the essential writing without having to worry over formatting or wasting time arguing with your word processor.

LaTeX is easily able to professionally typeset documents that run to hundreds or thousands of pages long. With simple mark-up commands, it automatically sets out the table of contents, margins, page headers and footers and keeps the formatting consistent and beautiful. One of its main strengths is the way it can easily typeset mathematics, even *heavy* mathematics. Even if those equations are the most horribly twisted and most difficult mathematical problems that can only be solved on a super-computer, you can at least count on LaTeX to make them look stunning.

1.2 Learning LaTeX

LaTeX is not a WYSIWYG (What You See is What You Get) program, unlike word processors such as Microsoft Word or Apple's Pages. Instead, a document written for LaTeX is actually a simple, plain text file that contains *no formatting*. You tell LaTeX how you want the formatting in the finished document by writing in simple commands amongst the text, for example, if I want to use *italic text for emphasis*, I write the `\emph{text}` command and put the text I want in italics in between the curly braces. This means that LaTeX is a "mark-up" language, very much like HTML.

1.2.1 A (not so short) Introduction to LaTeX

If you are new to LaTeX, there is a very good eBook – freely available online as a PDF file – called, "The Not So Short Introduction to LaTeX". The book's title is typically shortened to just *lshort*. You can download the latest version (as it is occasionally updated) from here: http://www.ctan.org/tex-archive/info/lshort/english/lshort.pdf

It is also available in several other languages. Find yours from the list on this page: http://www.ctan.org/tex-archive/info/lshort/

It is recommended to take a little time out to learn how to use LaTeX by creating several, small 'test' documents, or having a close look at several templates on: http://www.LaTeXTemplates.com

Making the effort now means you're not stuck learning the system when what you *really* need to be doing is writing your thesis.

1.2.2 A Short Math Guide for LaTeX

If you are writing a technical or mathematical thesis, then you may want to read the document by the AMS (American Mathematical Society) called, "A Short Math Guide for LaTeX". It can be found online here: http://www.ams.org/tex/amslatex.html under the "Additional Documentation" section towards the bottom of the page.

1.2.3 Common LaTeX Math Symbols

There are a multitude of mathematical symbols available for LaTeX and it would take a great effort to learn the commands for them all. The most common ones you are likely to use are shown on this page: http://www.sunilpatel.co.uk/latex-type/latex-math-symbols/

You can use this page as a reference or crib sheet, the symbols are rendered as large, high quality images so you can quickly find the LaTeX command for the symbol you need.

1.2.4 LaTeX on a Mac

The LaTeX distribution is available for many systems including Windows, Linux and Mac OS X. The package for OS X is called MacTeX and it contains all the applications you need – bundled together and pre-customized – for a fully working LaTeX environment and work flow.

MacTeX includes a custom dedicated LaTeX editor called TeXShop for writing your '`.tex`' files and BibDesk: a program to manage your references and create your bibliography section just as easily as managing songs and creating playlists in iTunes.

1.3 Getting Started with this Template

If you are familiar with LaTeX, then you should explore the directory structure of the template and then proceed to place your own information into the *THESIS INFORMATION* block of the `main.tex` file. You can then modify the rest of this file to your unique specifications based on your degree/university. Section 1.5 on page 4 will help you do this. Make sure you also read section 1.7 about thesis conventions to get the most out of this template.

If you are new to LaTeX it is recommended that you carry on reading through the rest of the information in this document.

Before you begin using this template you should ensure that its style complies with the thesis style guidelines imposed by your institution. In most cases this template style and layout will be suitable. If it is not, it may only require a small change to bring the template in line with your institution's recommendations. These modifications will need to be done on the `MastersDoctoralThesis.cls` file.

1.3.1 About this Template

This LaTeX Thesis Template is originally based and created around a LaTeX style file created by Steve R. Gunn from the University of Southampton (UK), department

of Electronics and Computer Science. You can find his original thesis style file at his site, here: http://www.ecs.soton.ac.uk/~srg/softwaretools/document/templates/

Steve's ecsthesis.cls was then taken by Sunil Patel who modified it by creating a skeleton framework and folder structure to place the thesis files in. The resulting template can be found on Sunil's site here: http://www.sunilpatel.co.uk/thesis-template

Sunil's template was made available through http://www.LaTeXTemplates.com where it was modified many times based on user requests and questions. Version 2.0 and onwards of this template represents a major modification to Sunil's template and is, in fact, hardly recognisable. The work to make version 2.0 possible was carried out by Vel and Johannes Böttcher.

1.4 What this Template Includes

1.4.1 Folders

This template comes as a single zip file that expands out to several files and folders. The folder names are mostly self-explanatory:

Appendices – this is the folder where you put the appendices. Each appendix should go into its own separate .tex file. An example and template are included in the directory.

Chapters – this is the folder where you put the thesis chapters. A thesis usually has about six chapters, though there is no hard rule on this. Each chapter should go in its own separate .tex file and they can be split as:

- Chapter 1: Introduction to the thesis topic
- Chapter 2: Background information and theory
- Chapter 3: (Laboratory) experimental setup
- Chapter 4: Details of experiment 1
- Chapter 5: Details of experiment 2
- Chapter 6: Discussion of the experimental results
- Chapter 7: Conclusion and future directions

This chapter layout is specialised for the experimental sciences, your discipline may be different.

Figures – this folder contains all figures for the thesis. These are the final images that will go into the thesis document.

1.4.2 Files

Included are also several files, most of them are plain text and you can see their contents in a text editor. After initial compilation, you will see that more auxiliary files are created by LaTeX or BibTeX and which you don't need to delete or worry about:

example.bib – this is an important file that contains all the bibliographic information and references that you will be citing in the thesis for use with BibTeX. You can write it manually, but there are reference manager programs available that will create and manage it for you. Bibliographies in LaTeX are a large subject and you

may need to read about BibTeX before starting with this. Many modern reference managers will allow you to export your references in BibTeX format which greatly eases the amount of work you have to do.

MastersDoctoralThesis.cls – this is an important file. It is the class file that tells LaTeX how to format the thesis.

main.pdf – this is your beautifully typeset thesis (in the PDF file format) created by LaTeX. It is supplied in the PDF with the template and after you compile the template you should get an identical version.

main.tex – this is an important file. This is the file that you tell LaTeX to compile to produce your thesis as a PDF file. It contains the framework and constructs that tell LaTeX how to layout the thesis. It is heavily commented so you can read exactly what each line of code does and why it is there. After you put your own information into the *THESIS INFORMATION* block – you have now started your thesis!

Files that are *not* included, but are created by LaTeX as auxiliary files include:

main.aux – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file.

main.bbl – this is an auxiliary file generated by BibTeX, if it is deleted, BibTeX simply regenerates it when you run the `main.aux` file. Whereas the `.bib` file contains all the references you have, this `.bbl` file contains the references you have actually cited in the thesis and is used to build the bibliography section of the thesis.

main.blg – this is an auxiliary file generated by BibTeX, if it is deleted BibTeX simply regenerates it when you run the main `.aux` file.

main.lof – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It tells LaTeX how to build the *List of Figures* section.

main.log – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It contains messages from LaTeX, if you receive errors and warnings from LaTeX, they will be in this `.log` file.

main.lot – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file. It tells LaTeX how to build the *List of Tables* section.

main.out – this is an auxiliary file generated by LaTeX, if it is deleted LaTeX simply regenerates it when you run the main `.tex` file.

So from this long list, only the files with the `.bib`, `.cls` and `.tex` extensions are the most important ones. The other auxiliary files can be ignored or deleted as LaTeX and BibTeX will regenerate them.

1.5 Filling in Your Information in the `main.tex` File

You will need to personalise the thesis template and make it your own by filling in your own information. This is done by editing the `main.tex` file in a text editor or your favourite LaTeX environment.

Open the file and scroll down to the third large block titled *THESIS INFORMATION* where you can see the entries for *University Name*, *Department Name*, etc ...

Fill out the information about yourself, your group and institution. You can also insert web links, if you do, make sure you use the full URL, including the `http://` for this. If you don't want these to be linked, simply remove the `\href{url}{name}` and only leave the name.

When you have done this, save the file and recompile `main.tex`. All the information you filled in should now be in the PDF, complete with web links. You can now begin your thesis proper!

1.6 The `main.tex` File Explained

The `main.tex` file contains the structure of the thesis. There are plenty of written comments that explain what pages, sections and formatting the LaTeX code is creating. Each major document element is divided into commented blocks with titles in all capitals to make it obvious what the following bit of code is doing. Initially there seems to be a lot of LaTeX code, but this is all formatting, and it has all been taken care of so you don't have to do it.

Begin by checking that your information on the title page is correct. For the thesis declaration, your institution may insist on something different than the text given. If this is the case, just replace what you see with what is required in the *DECLARATION PAGE* block.

Then comes a page which contains a funny quote. You can put your own, or quote your favourite scientist, author, person, and so on. Make sure to put the name of the person who you took the quote from.

Following this is the abstract page which summarises your work in a condensed way and can almost be used as a standalone document to describe what you have done. The text you write will cause the heading to move up so don't worry about running out of space.

Next come the acknowledgements. On this page, write about all the people who you wish to thank (not forgetting parents, partners and your advisor/supervisor).

The contents pages, list of figures and tables are all taken care of for you and do not need to be manually created or edited. The next set of pages are more likely to be optional and can be deleted since they are for a more technical thesis: insert a list of abbreviations you have used in the thesis, then a list of the physical constants and numbers you refer to and finally, a list of mathematical symbols used in any formulae. Making the effort to fill these tables means the reader has a one-stop place to refer to instead of searching the internet and references to try and find out what you meant by certain abbreviations or symbols.

The list of symbols is split into the Roman and Greek alphabets. Whereas the abbreviations and symbols ought to be listed in alphabetical order (and this is *not* done automatically for you) the list of physical constants should be grouped into similar themes.

The next page contains a one line dedication. Who will you dedicate your thesis to?

Finally, there is the block where the chapters are included. Uncomment the lines (delete the % character) as you write the chapters. Each chapter should be written in its own file and put into the *Chapters* folder and named `Chapter1`, `Chapter2`, etc...Similarly for the appendices, uncomment the lines as you need them. Each appendix should go into its own file and placed in the *Appendices* folder.

After the preamble, chapters and appendices finally comes the bibliography. The bibliography style (called *authoryear*) is used for the bibliography and is a fully featured style that will even include links to where the referenced paper can be found online. Do not underestimate how grateful your reader will be to find that a reference to a paper is just a click away. Of course, this relies on you putting the URL information into the BibTeX file in the first place.

1.7 Thesis Features and Conventions

To get the best out of this template, there are a few conventions that you may want to follow.

One of the most important (and most difficult) things to keep track of in such a long document as a thesis is consistency. Using certain conventions and ways of doing things (such as using a Todo list) makes the job easier. Of course, all of these are optional and you can adopt your own method.

1.7.1 Printing Format

This thesis template is designed for double sided printing (i.e. content on the front and back of pages) as most theses are printed and bound this way. Switching to one sided printing is as simple as uncommenting the *oneside* option of the `documentclass` command at the top of the `main.tex` file. You may then wish to adjust the margins to suit specifications from your institution.

The headers for the pages contain the page number on the outer side (so it is easy to flick through to the page you want) and the chapter name on the inner side.

The text is set to 11 point by default with single line spacing, again, you can tune the text size and spacing should you want or need to using the options at the very start of `main.tex`. The spacing can be changed similarly by replacing the *singlespacing* with *onehalfspacing* or *doublespacing*.

1.7.2 Using US Letter Paper

The paper size used in the template is A4, which is the standard size in Europe. If you are using this thesis template elsewhere and particularly in the United States, then you may have to change the A4 paper size to the US Letter size. This can be done in the margins settings section in `main.tex`.

Due to the differences in the paper size, the resulting margins may be different to what you like or require (as it is common for institutions to dictate certain margin sizes). If this is the case, then the margin sizes can be tweaked by modifying the values in the same block as where you set the paper size. Now your document should be set up for US Letter paper size with suitable margins.

1.7.3 References

The `biblatex` package is used to format the bibliography and inserts references such as this one (**Reference1**). The options used in the `main.tex` file mean that the in-text citations of references are formatted with the author(s) listed with the date of the publication. Multiple references are separated by semicolons (e.g. (**Reference2**; **Reference1**)) and references with more than three authors only show the first author with *et al.* indicating there are more authors (e.g. (**Reference3**)). This is done automatically for you. To see how you use references, have a look at the `Chapter1.tex` source file. Many reference managers allow you to simply drag the reference into the document as you type.

Scientific references should come *before* the punctuation mark if there is one (such as a comma or period). The same goes for footnotes[1]. You can change this but the most important thing is to keep the convention consistent throughout the thesis. Footnotes themselves should be full, descriptive sentences (beginning with a capital

[1] Such as this footnote, here down at the bottom of the page.

letter and ending with a full stop). The APA6 states: "Footnote numbers should be superscripted, [...], following any punctuation mark except a dash." The Chicago manual of style states: "A note number should be placed at the end of a sentence or clause. The number follows any punctuation mark except the dash, which it precedes. It follows a closing parenthesis."

The bibliography is typeset with references listed in alphabetical order by the first author's last name. This is similar to the APA referencing style. To see how LaTeX typesets the bibliography, have a look at the very end of this document (or just click on the reference number links in in-text citations).

A Note on bibtex

The bibtex backend used in the template by default does not correctly handle unicode character encoding (i.e. "international" characters). You may see a warning about this in the compilation log and, if your references contain unicode characters, they may not show up correctly or at all. The solution to this is to use the biber backend instead of the outdated bibtex backend. This is done by finding this in `main.tex`: *`backend=bibtex`* and changing it to *`backend=biber`*. You will then need to delete all auxiliary BibTeX files and navigate to the template directory in your terminal (command prompt). Once there, simply type `biber main` and biber will compile your bibliography. You can then compile `main.tex` as normal and your bibliography will be updated. An alternative is to set up your LaTeX editor to compile with biber instead of bibtex, see here for how to do this for various editors.

1.7.4 Tables

Tables are an important way of displaying your results, below is an example table which was generated with this code:

```
\begin{table}
\caption{The effects of treatments X and Y on the four groups studied.}
\label{tab:treatments}
\centering
\begin{tabular}{l l l}
\toprule
\tabhead{Groups} & \tabhead{Treatment X} & \tabhead{Treatment Y} \\
\midrule
1 & 0.2 & 0.8\\
2 & 0.17 & 0.7\\
3 & 0.24 & 0.75\\
4 & 0.68 & 0.3\\
\bottomrule\\
\end{tabular}
\end{table}
```

You can reference tables with `\ref{<label>}` where the label is defined within the table environment. See `Chapter1.tex` for an example of the label and citation (e.g. Table 1.1).

1.7.5 Figures

There will hopefully be many figures in your thesis (that should be placed in the *Figures* folder). The way to insert figures into your thesis is to use a code template like this:

TABLE 1.1: The effects of treatments X and Y on the four groups studied.

Groups	Treatment X	Treatment Y
1	0.2	0.8
2	0.17	0.7
3	0.24	0.75
4	0.68	0.3

```
\begin{figure}
\centering
\includegraphics{Figures/Electron}
\decoRule
\caption[An Electron]{An electron (artist's impression).}
\label{fig:Electron}
\end{figure}
```

Also look in the source file. Putting this code into the source file produces the picture of the electron that you can see in the figure below.

FIGURE 1.1: An electron (artist's impression).

Sometimes figures don't always appear where you write them in the source. The placement depends on how much space there is on the page for the figure. Sometimes there is not enough room to fit a figure directly where it should go (in relation to the text) and so LaTeX puts it at the top of the next page. Positioning figures is the job of LaTeX and so you should only worry about making them look good!

Figures usually should have captions just in case you need to refer to them (such as in Figure 1.1). The `\caption` command contains two parts, the first part, inside the square brackets is the title that will appear in the *List of Figures*, and so should be short. The second part in the curly brackets should contain the longer and more descriptive caption text.

The `\decoRule` command is optional and simply puts an aesthetic horizontal line below the image. If you do this for one image, do it for all of them.

LaTeX is capable of using images in pdf, jpg and png format.

1.7.6 Typesetting mathematics

If your thesis is going to contain heavy mathematical content, be sure that LaTeX will make it look beautiful, even though it won't be able to solve the equations for you.

The "Not So Short Introduction to LaTeX" (available on CTAN) should tell you everything you need to know for most cases of typesetting mathematics. If you need more information, a much more thorough mathematical guide is available from the AMS called, "A Short Math Guide to LaTeX" and can be downloaded from: `ftp://ftp.ams.org/pub/tex/doc/amsmath/short-math-guide.pdf`

There are many different LaTeX symbols to remember, luckily you can find the most common symbols in The Comprehensive LaTeX Symbol List.

You can write an equation, which is automatically given an equation number by LaTeX like this:

```
\begin{equation}
E = mc^{2}
\label{eqn:Einstein}
\end{equation}
```

This will produce Einstein's famous energy-matter equivalence equation:

$$E = mc^2 \tag{1.1}$$

All equations you write (which are not in the middle of paragraph text) are automatically given equation numbers by LaTeX. If you don't want a particular equation numbered, use the unnumbered form:

```
\[ a^{2}=4 \]
```

1.8 Sectioning and Subsectioning

You should break your thesis up into nice, bite-sized sections and subsections. LaTeX automatically builds a table of Contents by looking at all the `\chapter{}`, `\section{}` and `\subsection{}` commands you write in the source.

The Table of Contents should only list the sections to three (3) levels. A `chapter{}` is level zero (0). A `\section{}` is level one (1) and so a `\subsection{}` is level two (2). In your thesis it is likely that you will even use a `subsubsection{}`, which is level three (3). The depth to which the Table of Contents is formatted is set within `MastersDoctoralThesis.cls`. If you need this changed, you can do it in `main.tex`.

1.9 In Closing

You have reached the end of this mini-guide. You can now rename or overwrite this pdf file and begin writing your own `Chapter1.tex` and the rest of your thesis. The easy work of setting up the structure and framework has been taken care of for you. It's now your job to fill it out!

Good luck and have lots of fun!

Guide written by —
Sunil Patel: www.sunilpatel.co.uk
Vel: LaTeXTemplates.com

Chapter 2

State of the art: systematic review

In the following sections of chapter 2 we introduce the details about the research method used to describe the state of the art of the similarity detection techniques in duplicated text detection.

2.1 Definition of the research method

First of all, it is necessary to defined a protocol to carry on the research. In order to avoid a vague research methodology that could lead to poor results, we aim to define a review method based on a systematic review. Using this guidance, we ensure a thorough literature review of the field of interest (i.e., similarity detection in texts) of significant scientific value, which is used in this thesis as the foundations for the main developed work.

For this purpose, it is proposed to follow the guidelines of B. Kitchenham's systematic review methodology **??**, which is focused on applying this review process in the software engineering research field. This guidelines include a series of well-defined stages and processes for both planning and conducting the review.

Therefore, we propose to design and implement the following steps for the systematic review, based on Kitchenham's proposal.

1. Planning the review (see section 2.2)
   (a) Identifying the need for a review
   (b) Specifying the review questions
   (c) Developing a review protocol
2. Conducting the review (see section 2.3)
   (a) Identification of the search
   (b) Selection of primary studies
   (c) Quality assessment study
   (d) Data extraction & data synthesis
3. Reporting the review (see section **??**)

2.2 Planning the review

The following subsections provide a description of the three main steps of the systematic review planning. In this stage, we refine the scope of the research based on the scope of our project by establishing the general requirements of this task (i.e., what do we want to research about or where do are we going to look for the required knowledge) and all related details about how to perform this review.

2.2.1 Identifying the need for a review

As a first step it is necessary to justify the need for a systematic review in the similarity detection field for this master thesis. If we go back to the general objectives described at section **??**, we can relate this to the objective *O1*:

O1. To research the state-of-the-art of the textual similarity detection field.

Although this state-of-the-art research could be executed following alternative methods to the systematic review, we justify this approach with two reasons.

The first one is related to the requirements and specific objectives of this project. The purpose of developing and evaluating specific similarity detection algorithm implementations is to provide empirical demonstrations of the most important proposals in identifying paraphrase textual units. Furthermore, and as a contribution to the field, we aim to evaluate how this approaches behave in the requirements similarity detection field. We need to ensure that our review is thorough enough to provide a general overview of the main proposals considered as suitable for solving this problematic. This output can then be used as an input to choose specific implementations to develop and to evaluate in the scope of this master thesis.

The second one is related to the available literature in textual duplicate detection. Kitchenham's methodology states the importance of looking for any systematic reviews available on the field, which in case of existing would undermine the need of performing a systematic review for this master thesis. Therefore as a first step we focus on looking for already available state-of-the-art reviews.

We focus on looking for any review related to the similarity detection field using Natural Language Processing, Machine Learning or general Artificial Intelligence techniques. To increase the results of the research, we do not focus on systematic reviews, but in any kind of research regardless the methodology.

The following databases are used to look for literature review of the field of study:

- **Scopus ??**
- **ACM Digital Library ??**
- **IEEE Xplorer ??**
- **Science Direct ??**

Based on the target of the literature review, we propose a search string composed by three blocks of data or information we are interested in capturing with the research:

1. **Similarity field.** We use any match with *similar**, *duplicat** or *paraphras**, which are three of the main synonyms used to refer to a pair of texts which may be considered as semantically equivalent.
2. **Artificial Intelligence technologies.** The search must be restricted to the detection of similar textual items using *AI* technologies, with a special emphasis in Natural Language and Machine Learning techniques, which represent the main approaches for this issue.

3. **State-of-the-art review.** In this stage of the systematic review, it is necessary to focus only on papers and publications which contribute providing a detailed state-of-the-art analysis. Therefore it is important to use terms such as *review* or *state of the art* as part of the search. This will guarantee that one of the main goals of the results in the search are focused on providing this state-of-the-art as an output of the publication.

Consequently, the following search string is proposed:

(similar* OR duplicat* OR paraphras*) AND ("natural language" OR "machine learning" OR "artificial intelligence" OR "AI" OR "NLP" OR "ML") AND (review OR "state of the art")

We apply this search to the title and the keywords of the publications - this will ensure a minimum noise on the results which are not specifically focused on these three main blocks.

These are the results:

- Scopus - 1 result.
- ACM Digital Library - 1 result.
- IEEE Xplorer- 0 results.
- Science Direct - 0 results.

After a general overview of the results provided by Scopus and ACM, it is concluded than none of the two results are related to the paraphrase detection field in natural language texts and hence they can be excluded. As a conclusion, it is stated that there are no publications providing a detailed, structured analysis of the state-of-the-art techniques for similarity detection between natural language text pairs. Therefore it is confirmed the need of a systematic review as a first step of this master thesis.

2.2.2 Specifying the review question

Kitchenham's methodology states the need of defining three elements of the research to help designing and defining the review scope. Applied to the software engineering field, these items are:

- **Population**: it refers to groups or agents (subjects) that are affected by the intervention of the review question. In this thesis, and due to the fact that natural language paraphrase detection is a wide application area, we generally focus on software Engineering teams (developers, team managers...), which are the agents interested in using this automated techniques for solving the problem addressed in this master thesis.
- **Intervention**: it applies to the technologies used to address a specific issue. We focus on the automated similarity detection techniques and the algorithmic approaches using machine learning, natural language processing and artificial intelligence techniques in general.

- **Outcome**: it relates to factors and output data which are relevant to evaluate the quality of a specific solution and to compare them. For evaluating the quality of the developed algorithms, we will focus on the accuracy and the performance of these solutions.

Using these three elements as an input, we focus our systematic review in solving two related research questions:

1. How does the software engineering community handles automated similarity detection between natural language text pairs using Artificial Intelligence (i.e., Natural Language Processing and Machine Learning)?
2. Which are the results of these general approaches in terms of accuracy and performance and which are considered as the best approaches from a qualitatively point of view using these indicators?

2.2.3 Developing a review protocol

Once the basis of the review has been established, the next step is to define a practical review protocol to be applied when conducting the search. This synthesizes which data to look for, the filters to apply to the result data, and how to extract and analyze the information of each publication from a practical point of view.

The search strategy: search & selection procedures

Analogously to the requirements of the systematic review need justification, we focus on answering the questions *'what'* (i.e., which data are we going to look for) and *'where'* (i.e., which databases are going to be used).

1. The selected databases are the same ones used in section 2.2.1: Scopus, ACM Digital Library, IEEE Xplorer and Science Direct
2. The search string is composed of two of the three main blocks of the one used in section 2.2.1, removing the part related to the state-of-the-art review:

(similar* OR duplicat* OR paraphras*) AND ("natural language" OR "machine learning" OR "artificial intelligence" OR "AI" OR "NLP" OR "ML")

Once the required input data for the research has been selected, we can define a review protocol to look for and to filter the research results. This protocol is proposed to follow three steps:

1. To search in the databases using the search string defined above.
2. To merge the documents into a single repository and to remove duplicates (i.e., publications found in more than one database). For this purpose, and for future management tasks related to the literature documentation, we propose to use the Mendeley tool **??**, a research documentation reference manager.
3. To filter the found documents following a study selection criteria. This procedure is proposed to followed the next stages.
   (a) First, to filter publications by title, removing all results that are clearly out of the scope of our research

(b) Second, to filter by reading the abstract of the publication

(c) Third, to filter by giving a general overview to the document, also known as *skimming*. This include reading some of the main important sections (i.e., introduction and conclusions) or by evaluating the general structure of the paper.

(d) Fourth, to give a full reading to the document.

Data extraction & synthesis strategies

We propose a template (see table **??**) to fulfill for each document of the systematic review. The purpose of these template is to provide a general, structured analysis of the most relevant issues of each publication.

Topic	**Details**
Domain of the proposal	Is a domain-specific proposal? If it is, which domain?
Objects of similarity detection	Which are the subjects of the similarity detection (full documents, short texts, sentences...)? Does the proposed methodology apply to a specific textual entity?
Similarity algorithm description	Does it include a sequential description of the technical process? Is it detailed enough to reproduce?
NLP preprocessing	Does the algorithm include a NLP preprocess step? Which are the tasks related to natural language preprocessing of the data?
Similarity functions	Are any word-to-word similarity functions used? If so, which ones? Does it include a specific aggregated function proposal?
Machine learning classification process	Does the algorithm include a classification process? Which kind of features are used (semantical, lexical, synctactical...)
Output results	Is a similarity score available for each pair of compared text items? Is there a implicit/explicit classification result (i.e. are duplicates retrieved by the algorithm itself, or a threshold score must be used?)
Frameworks and external tools	Is there a list/reference to NLP/ML frameworks and third-party tools used for the similarity detection process? Are they free-to-use?
Experimentation	Is the algorithm tested with real-data experiments? Is data available? Which kind of data is used (type, volume...)?
Results	Is there any reference to results in terms of accuracy? Reliability? Execution time? Used hardware resources?
Evaluable tools	Is the proposal distributed as a tool or piece of code which can be tested?

2.3 Conducting the review

This sections describes the process of the conducted review and all the decisions that have been made along this process, in correlation with the previous systematic review plan. It also collects the problems and difficulties found during the systematic review process.

2.3.1 Conducting the search

This section relates to steps 1 and 2 of the review protocol: *To search in the databases using the search string defined above* and *To merge the documents into a single repository and to remove duplicates*.

Two main problems are faced during the search:

- Due to the complexity of the search string, which uses a logical combination of union and intersection logical operations between terms, it was required to adapt the search string to each consulted database. Each document search engine uses its own syntax to define searches and these logical operation between strings and hence it was necessary to adapt and test each one to guarantee the results were applying to our criteria.
- As introduced in section 2.2, it is necessary to identify and remove duplicated results from the search in order to avoid redundancies. Sometimes publications metadata or document variations present some anomalies making it difficult to automatically detect these duplicates. However, using the previously mentioned Mendeley tool, it is also possible to detect documents which are not exact duplicates but have a high probability of being replicates. These feature was used to solve this issue.

After facing the first issue, the search is conducted, retrieving the following results.

- Scopus - 193 results
- ACM Digital Library - 121 results
- IEEE Xplorer- 38 results
- Science Direct - 6 results

The previous list only indexes individual results for each database. The second step of the review protocol is to use the reference manager to remove duplicates and nearly duplicates, which relates to the second problem faced during this step.

- **Total nº documents = 358**
- **Duplicates = 48**
- **Almost duplicates = 11**
- **Final nº documents = 358 - (48 + 11) = 299**

At the end of step 2, the research has achieved a total number of 299 publications to be reviewed in the following stages.

2.3.2 Selection of primary studies

This section relates to step 3 of the review protocol: *To filter the found documents following a study selection criteria.*. This phase is composed by 4 *sub-steps* where each one acts as a filter to reduce the number of documents based on non-relevant research studies identification.

The first step is to **filter by the title** of the publication. In this step the number of documents is reduced from 299 to **74**, which means a total of 225 papers are deleted from the manager tool. The main reasons for discarding these documents are covered in the following list:

- An important number of papers are focused on the medical field - specifically, they address detecting similarities and replication patterns using artificial intelligence techniques, such as parallel patients, or disease diagnosis by medical recognition. All papers focused on different areas than natural language similarity are rejected.
- Some of the publications are focused on grammar and syntax knowledge of non-romance languages like Hindi or Chinese. The rules and techniques of natural language are highly coupled to the main characteristics of the language itself. Therefore, we exclude these publications and focus on those using techniques for the English language.
- Although matching the search string and non of the above restrictions, some titles suggested that some of the works were focused on solving different problems that are out of scope of this master thesis - for instance, language translation. These works are also eliminated from our article repository.

We pass all works which title does seem to focus on the research field and do not satisfy any of the restriction criteria commented above.

The second step is to **filter by the abstract** by reading it carefully and acquiring a deeper understanding on the subject that the information provided by the title, which sometimes might be vague or imprecise. In this second step we reduce the number of works from 74 to **34**, discarding a total of 40 works due to reasons like the ones listed below:

- Some abstracts give us a deeper knowledge on some of the filtering criteria used on the previous step, like out of field subjects (i.e., an abstract introduced that the work was focused on web-service similartiy) or the use of different languages (i.e., Turkish).
- Some works that study the text similarity are not focused on the semantic dimension, but on other criteria such as the authorship of a piece of text. Therefore they target the study of properties and other criteria which is of no interest for the scope of this thesis

The third step is based on applying a **skimming** or general overview evaluation of the whole work, reading some of the main parts and studying the general structure and main features of its contributions. In this step we reduce the number of documents from 34 to 15. Listed below are some of the filter criteria used in this stage:

- Non-relevant contributions or out-of-date approaches that are better approached and discarded in other more recent works are removed. By applying this simplification we decrease redundancy and focus on evaluating actually relevant contributions.
- Works claiming poor results or a lack of achievement of their main goals

- Some works are focused on studying and analyzing the word-to-word similarity evaluation. This is an important step in any natural language paraphrase detection process, but we consider that the purpose of this thesis must be focused on text similarity detection using requirements and hence relatively small pieces of texts as the elements to be compared. Technical details and the study of word-to-word similarity are assumed to be already acknowledged and this master thesis does not focus on the details of these area.
- Some of the general approaches for an algorithmic process for similarity detection provide are ontology based, which means they require of an explicit, modeled knowledge of the domain of the input data to detect similarities. Although this is an interesting solution with a high potential, we decide to discard this kind of solutions of the scope of this thesis for two reasons.
- Missing detailed-specific process
- More language (not detected before)
- I.e. Semilar TOOL description (too general) Initially we keep those about plagiarism, with a skeptical mind. Also those related to sentences (i.e. subject vs subject, predicate vs predicate, etc.)

Full reading From ?? to ??. Reasons:

- A paper was too focused on plagiarism and it was more an index of tools and frameworks than a proposal or a technical description of a similarity process.
- A paper was too focused on a tool portfolio and although it provided results and an overview of different algorithms and tools, it did not introduce further details that were necessary to the research.
- A paper required a first process in which training data required chunk identification. Moreover, it did not include enough detail for all 247 features used.

2.3.3 Study quality assessment

2.3.4 Data extraction and synthesis

General approaches:

- VECTOR-SPACE APPROACH: NLP preprocessing of data + customized vector based method (combining TFIDF) + optimization process [2][4]
- ALIGNMENT APPROACH: NLP preprocessing of data + lexical & syntactic basic module score + average score against threshold [3][5]
- MACHINE LEARNING APPROACH: NLP preprocessing of data + Feature Extraction process (lexical & syntactic) + ML classification (supervised) [1][6][8][10]

Other approaches: -

2.4 Algorithm selection and technical analysis

2.5 Algorithm comparative analysis

Appendix A

Frequently Asked Questions

A.1 How do I change the colors of links?

The color of links can be changed to your liking using:

`\hypersetup{urlcolor=red}`, or
`\hypersetup{citecolor=green}`, or
`\hypersetup{allcolor=blue}`.

If you want to completely hide the links, you can use:

`\hypersetup{allcolors=.}`, or even better:
`\hypersetup{hidelinks}`.

If you want to have obvious links in the PDF but not the printed text, use:

`\hypersetup{colorlinks=false}`.