Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Dissertation Title

John Carbeck
16309095

March 30, 2020

A Final Year Project submitted in partial fulfilment
of the requirements for the degree of
BA (Computer Science)

Declaration

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`. I

have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at

`http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: _____ Date: _____

Abstract

A short summary of the problem investigated, the approach taken and the key findings. This should be around 400 words, or less.

This should be on a separate page.

Acknowledgements

Thanks Mum!

You should acknowledge any help that you have received (for example from technical staff), or input provided by, for example, a company.

Contents

List of Figures

List of Tables

Nomenclature

A	Area of the wing	m^2
B		
C	Roman letters first, with capitals…	
a	then lower case.	
b		
c		
Γ	Followed by Greek capitals…	
α	then lower case greek symbols.	
β		
ϵ		
TLA	Finally, three letter acronyms and other abbreviations arranged alphabetically	

If a parameter has a typical unit that is used throughout your report, then it should be included here on the right hand side.

If you have a very mathematical report, then you may wish to divide the nomenclature list into functions and variables, and then sub- and super-scripts.

Note that Roman mathematical symbols are typically in a serif font in italics.

1 Introduction

The internet has brought with it an explosion of online content. Huge sets of textual based content now live on the World Wide Web. This ever growing volume of content has led to the development of information retrieval (IR) systems. These systems help guide users to relevant content from their expressed need. IR systems have been implemented to serve specific domains performing retrieval on smaller sets of documents. These domain specific systems aim to improve the form of the retrieved information as well as relevance of documents to the specific information needs of the user. These systems often use domain specific models to expand queries or perform summarization tasks. These domain models take the form of ontologies or knowledge bases.

Wikipedia offers over 6 million articles in english, a user faces the problem of two much information due to the plethora of relevant and related documents for a given topic in a larger domain. World War II as an example contains 26,388 directly related pages. Even within a small domain such as the Watergate Scandal there are 32 relevant pages and while some pages provide a general overview of a topic, when reading or learning on a sub-topic of a particular topic many of the pages that are related lay latent. Many of these topics in Wikipedia lack formal domain specific models that are commonly used in specific domain IR.

The creation of domain specific models for IR requires human intervention and experts in the given domain, and semantic based knowledge bases often fail to contain specific domain relations. Both general and domain specific information systems are permissive and rely on user requests of information. User permissive IR systems cause greater cognitive load then a system that anticipates information need and preforms preemptive retrieval.

The creation of domain specific models for IR requires human intervention and experts in the given domain. Semantic based knowledge bases often are too generalised and fail to contain specific domain relations. Both general and domain specific information systems are permissive and rely on user requests of information. User permissive IR systems cause greater cognitive load then preemptive IR systems

that anticipates user's information needs. A preemptive IR system allows for greater guidance to undiscovered content.

In this paper, a summarization system is proposed that uses Latent Dirichlet Allocation to generate unsupervised topic models for a domain specific set of documents, Wikipedia pages related to The Watergate Scandal. The generated topic models are used with a user knowledge model to generate queries relating to gaps in knowledge as represented in the model. Documents are retrieved based on the generated query and a summarisation of these latent topics are presented to guide the user to unknown topics in a domain. This system offers:

- No reliance on expert create ontologies or specific domain models
- Organic Topic generation
- Preemptive query generation based on user knowledge modeling
- Multidocument Extractive summarisation
- User interoperability of the systems summarisation generation

The proposed summarisation system is constructed from existing IR methods and tested against these systems to assess the implementation of these techniques. The use of different user knowledge models results in tailored summarization based on the information needs of that specific user.

1.1 The Conclusions chapter

The final chapter should give a short summary of the key methods, results and findings in your project. You should also briefly identify what, if any, future work might be executed to resolve unanswered questions or to advance the study beyond the scope that you identified in Chapter 1.

2 Background

2.1 Automatic Text Summarisation

Automatic text summarisation can be approached in many different ways. Generally the aim is to produce a summary, defined as "a text that is produced from one or more texts, that conveys important information in the original text(s)"(Radef et al. 2002). Allahyari et al. (2017) define Automatic text summarisation as "the task of producing a concise and fluent summary while reserving key information content and overall meaning". Automatic text summarisation has many forms as each summarisation task uses different types of source documents, representation of content, and reasoning in producing a summary. The many forms of automatic text summarisation are discussed in this section.

2.1.1 Summary Characteristics

The context of the summarisation task must be addressed in order to best perform automatic text summarisation. Spärck-Jones (1999) in her taxonomy writes: "It is important to recognize the role of context factors because the idea of a general-purpose summary is manifestly an ignis factus". The three context factors she identifies are input factors, purpose factors, and output factors. Input factors are classification of the representation of input document(s) in terms of structure, genre, format, and unit. The purpose factors are the relationship between the source and the output of summarisation and are described as dealing with situation, audience, and use. The output factors define the form of output of summary and are largely driven by the input and purpose factors of the system.

2.1.2 Types of Summaries

Gambhir and Gupta (2017) as well as Orăsan (2019) present a recent taxonomy to classify types of summaries. These classifications are important to consider when selecting an existing summarization method for a specific task or when creating a new automatic summarisation method.

Single document and Multi-document Summarisation

The input to a summarisation system can either be a single document or a set of multiple documents. Single document summarization addresses the content of a single document and produces a summary of that single document. Multi-document summarization considers content from multiple documents and produces a summary of the discussed topics across all given documents.

Many of the techniques of single document summarisation can be used in multi-document summarisation. Goldstein et al. (2000) identifies that: the redundancy of information of topically-related documents is much greater than in a single document making anti-redundancy methods crucial, the compression ratio (i.e. summary length with respect to document set length) is much smaller adding more difficulty to summarisation as compression demand increases, as well as the increased amount of co-referencing in a set of multiple documents than single documents. Many recent approaches attempt to deal with these issues. Methods to handle these issues will be discussed in the Redundancy Reduction section in this chapter.

Extractive and Abstractive Summarisation

The output of an automatic summarisation is either extractive or abstractive. An extractive summary is created from a subset of sentences from the source document(s). The sentences selected are those that the summarisation method finds most salient in the original text, using a similarity or centrality metric. An abstractive summary uses semantic models to generate a new piece of text that covers the themes, concepts or terms of the examined document or documents. Abstractive summarization requires natural language processing to extract concepts from the source material and to create an abstract summary from concept and word semantic relationships. Extractive summarization is simpler than abstractive but is limited because not all information in a sentence might not relate for a summary.

Generic and Query-focused Summaries

The purpose of summarisation is either generic or query-focused. Generic summaries attempt to summarize the content of all the material in the document or documents. This is the most common form of summary and is often used with single document summarisation. Query-focused, also referred to as topic-focused or user-focused, provides a summarisation based on a described need. These are commonly used with multi-document summarisation as multiple documents often contain a variety of topics. In this form of automatic text summarization a query is used both for the retrieval of documents as well as for the generation of the summary.

Personalised summaries are a type of user-focused summary. Personalised summaries aim to produce a tailored summary based on a model of the user. Díaz and Gervás (2007) personalised summaries of newswire texts using a model of user interests based on keywords, domain-specific factors and user feedback. Li, Liu and Zhao (2015) suggest an update summarisation system which considers the novelty of the sentence by adding novelty as a variable to traditional integer linear programming approaches of summarisation.

Supervised and Unsupervised Automatic Summarization

Another distinction of summarization methods is supervised and unsupervised methods. Supervised methods require training from a pre-labeled data set. Supervised methods use two class classification algorithms, trained on labeled data, for selection of important content in source documents. Unsupervised methods are able to generate summaries from only using the source documents, and can therefore operate on new documents without the need for training. Unsupervised summarisation identifies relevant sentences using a set of heuristic rules and uses those sentences to generate a summary.

2.2 Tasks of Summarisation

2.2.1 Intermediate Representation

2.2.2 Sentence Scoring

2.2.3 Sentence Selection

2.3 Summarisation Methods

2.3.1 Topic Representation

2.3.2 Indicator Representation and Machine Learning

2.3.3 Comparison of Methods

2.4 Query focused Summarisation

2.5 Figures

Graphs, pictures and other images should be included in your report as a numbered, captioned figure. An example is given in Figure 2.1.

The figure and caption should be centred. The figure numbering starts at 1 at the beginning of each chapter. The caption should provide a brief description of what is being shown. The figure should appear in the document after it is referred to in the text. No figure should be included which is not referred to in the text. Ensure that the size and resolution of images imported from software are sufficient to read any text.

2.6 Tables

Tables are an important way of displaying your results; Table 2.1 is a sample table, adapted from the Master/Doctoral Thesis template at `http://www.latextemplates.com/cat/theses`, which was generated with this code:

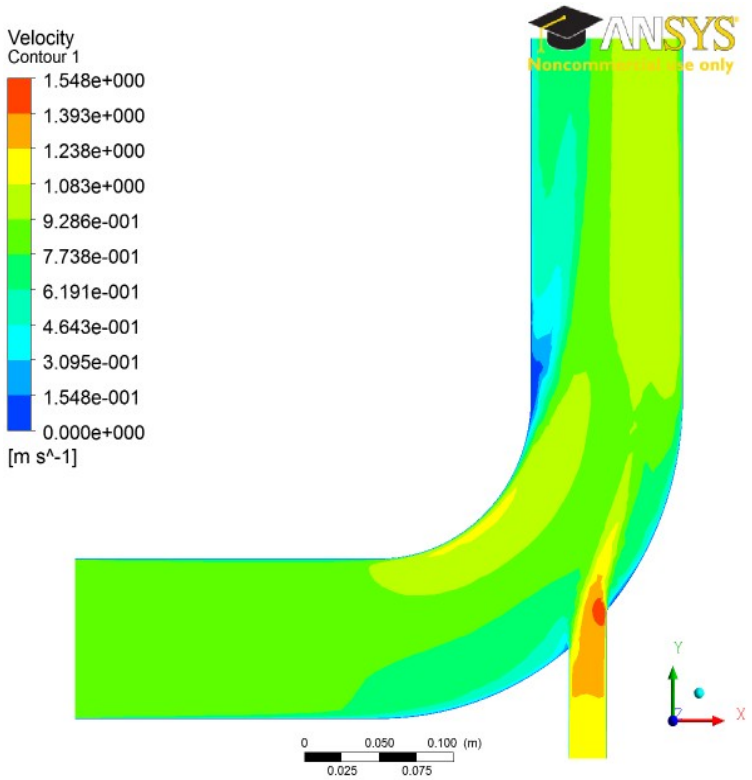

Figure 2.1: Velocity distribution on the mid-plane for an inlet velocity for case 1.

```
\begin{table}[b]
\caption{The effects of treatments X and Y on the four groups studied.}
\label{tab:treatments}
\centering
\begin{tabular}{l l l}
\toprule
\textbf{Groups} & \textbf{Treatment X} & \textbf{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}
```

Table 2.1: 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

Tables are numbered in the same way as figures. Typically tables also have a short caption, but this is not universally true. The number and caption appear above the table, not below as with figures. Again, no table should appear in the report which has not been referred to in the text. Tables should come after they are discussed in the text. The exact formatting of the table depends somewhat on the content of the table, but in general, the text in the table should be the same font and size as the main text.

2.7 Equations

All equations should be numbered sequentially. Do not restart the numbering at the beginning of each chapter. Unlike figures and tables, you may not need to refer to every equation in the text. You should take care to format equations properly. Do no simply try to use plain text. Use the equation layout facilities. An example of how equations should appear is shown in Equation 1. Here is the code for it:

```
\begin{equation}
\textrm{div}(\underline{u}) = \frac{\delta u}{\delta x} + \frac{\delta v}{\delta y} +
        \frac{\delta w}{\delta z} = 0
\label{sampleequation}
\end{equation}
```

$$\mathrm{div}(\underline{u}) = \frac{\delta u}{\delta x} + \frac{\delta v}{\delta y} + \frac{\delta w}{\delta z} = 0 \tag{1}$$

2.8 Referencing published work

It is important to give appropriate credit to other people for the work that they have shared through publications. In fact, you must sign a declaration in your report stating that you understand the nature of plagiarism. As well as avoiding plagiarism, citing results or data from the literature can strengthen your argument, provide a favourable comparison for your results, or even demonstrate how superior your work is.

There are many styles to reference published work. For example, the parenthetical style (which is also called the Harvard style) uses the author and date of publication (e.g. “Smith and Jones, 2001”). There is also the Vancouver (or the citation sequence) style, which is shown in this document. In the Vancouver style, the publications are cited using a bracket number which refers to the list in the References section at the

end of the report. The references are listed in order that they are cited in the report. A variant is name sequence style in which the publications are referenced by number, but the list is arranged alphabetically. For example, the text might say: several studies have examined the sound field around tandem cylinders generated by flow(1, 2), while other investigations have focused on the effect of an applied sound field on the flow(3). Papers from conference proceedings(4), books(5) and technical reports(6, 7) can be dealt with in the same style.

The Vancouver style has the advantage that it is a little more compact in the text and does not distract from the flow of the sentence if there are a lot of citations. However, it has the disadvantage that it is not immediately clear to the reader what particular work has been referenced.

It actually does not matter which particular referencing style is used as long as three important considerations are observed:

- the referencing style used throughout the document is consistent;
- all material used or discussed in the text is properly cited;
- nothing is included in the reference list that has not been cited.

This template has a suitable referencing style already set up – you should use it and use the built-in BibTeX system to manage your references. See above for examples of how to cite a reference and look in the `sample.bib` file to see BibTeX references. Remember Google Scholar and other search engines will give you BibTeX references for lots of academic publications. Otherwise, you can easily make up your own based on the examples in that file.

3 Method

seeing LaTeX, or more properly "LaTeX 2ε", is a very useful document processing program. It is very widely used, widely available, stable and free. Famously, TeX, upon which LaTeX is built, was originally developed by the eminent American mathematician Donald Knuth because he was tired of ugly mathematics books(8). Although it has a learning curve (made much less forbidding by online tools and resources – see below), it allows the writer to concentrate more fully on the content, and takes care of most everything else.

While it can be used as a word processor, it is a *typesetting* system, and Knuth's idea was that it could be used to produce beautiful looking books:

> *LaTeX is a macro package which enables authors to typeset and print their work at the highest typographical quality, using a predefined, professional layout.*[1]

LaTeX has great facilities for setting out equations and a powerful and very widely supported bibliographic system called BibTeX, which takes the pain out of referencing.

Three useful online resources make LaTeX much better:

(1) An excellent online LaTeX environment called "Overleaf" is available at `http://www.overleaf.com` that runs in a modern web browser. It's got this template available – search for a TCD template. Overleaf can work in conjunction with Dropbox, Google Drive and, in beta, GitHub.

(2) Google Scholar, at `http://scholar.google.com`, provides BibTeX entries for most of the academic references it finds.

(3) An indispensable and very fine introduction to using LaTeX called *"The not so short introduction to LATEX 2ε"* by Oetiker et al. (9) is online at `https://doi.org/10.3929/ethz-a-004398225`. Browse it before you use LaTeX for the first time and read it carefully when you get down to business.

[1] This is from Oetiker et al. (9). Did we mention that you should minimise your use of footnotes?

Other tools worth mentioning include:

- `Draw.io` – an online drawing package that can output PDFs to Google Drive – see `https://www.draw.io`.

4 Results

5 Conclusion

Bibliography

[1] JA Fitzpatrick. Flow/acoustic interactions of two cylinders in cross-flow. *Journal of Fluids and Structures*, 17(1):97–113, 2003.

[2] SL Finnegan, C Meskell, and S Ziada. Experimental investigation of the acoustic power around two tandem cylinders. *Journal of Pressure Vessel Technology*, 132(4): 041306, 2010.

[3] JW Hall, S Ziada, and DS Weaver. Vortex-shedding from single and tandem cylinders in the presence of applied sound. *Journal of Fluids and Structures*, 18(6): 741–758, 2003.

[4] Peter Jordan, John Fitzpatrick, and Craig Meskell. Array beam pattern control for measurement of propeller noise. In *AIAA/CEAS Aeroacoustics Conference and Exhibit, Maastricht, Netherlands*, 2001.

[5] Michael P Païdoussis, Stuart J Price, and Emmanuel De Langre. *Fluid-structure interactions: cross-flow-induced instabilities*. Cambridge University Press, 2010.

[6] L Reyes. Power uprate program status report-secy-07-0090. Technical report, Technical Report, US Nuclear Regulatory Commission, 2007.

[7] International Energy Agency. Technology Roadmap — Geothermal Heat and Power. `https://www.iea.org/publications/freepublications/publication/Geothermal_Roadmap.pdf`.

[8] Len Shustek and Donald Interviewee-Knuth. Interview donald knuth: A life's work interrupted. *Communications of the ACM*, 51(8):31–35, 2008.

[9] Tobias Oetiker, Hubert Partl, Irene Hyna, and Elisabeth Schlegl. The not so short introduction to latex 2ε. 2001. `https://doi.org/10.3929/ethz-a-004398225`.

A1 Appendix

You may use appendices to include relevant background information, such as calibration certificates, derivations of key equations or presentation of a particular data reduction method. You should not use the appendices to dump large amounts of additional results or data which are not properly discussed. If these results are really relevant, then they should appear in the main body of the report.

A1.1 Appendix numbering

Appendices are numbered sequentially, A1, A2, A3...The sections, figures and tables within appendices are numbered in the same way as in the main text. For example, the first figure in Appendix A1 would be Figure A1.1. Equations continue the numbering from the main text.