Identifying text phrases containing similar semantic meaning with search keywords using NLP

Y. T. L Somarathna

Index No : 16001402

Supervisor: Dr M. G. N. A. S. Fernando

<February 2021>

Submitted in partial fulfillment of the requirements of the

B.Sc in Computer Science Final Year Project (SCS4124)

UCSC

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: Y. T. L. Somarathna

..………………………………………..

Signature of Candidate   Date:

This is to certify that this dissertation is based on the work of

Dr. M. G. N. A. S. Fernando

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Principle/Co- Supervisor's Name: Mr W.V Welgama

..………………………………………..

Signature of Supervisor   Date:

Abstract

A core strategy of content SEO is placing the right amount of keywords in a webpage to archive the optimal keyword density with related quality keywords. The current approach for performing this task is manually going through the content and checking for text phrases that can be replaced by keywords and replacing them. In this study, we focus on a specific domain of websites (Android) and try to develop a model that can automatically recognize text phrases that are semantically similar to keywords. Distributional semantics is used to capture the meaning of words and a un-supervised classifier is used to classify phrases related to a particular keyword with the help of words semantics captured by distributional semantics.

Even with the latest NLP advancement capturing and comparing the semantics of text phrases seems to be some problems as text phrases have less context to work with. This study shows that still, we can produce useful results in recognizing text phrases with similar semantic meaning with search keywords.

Preface

The purpose of this is for you to state explicitly the extent to which your dissertation relies on the work of others, and highlight the portion that you claim to be your own original work. For example: you might say: "The results of chapter 3 rely upon a simulation provided by the research group. The analysis of the data is entirely my own work. I carried out the analytical calculation of chapter 4 in conjunction with my supervisor..." and so on. <u>Without this statement, it will be assumed that no work is original and that your thesis is a review article</u>. If you merely claim that the thesis is all your own work, you should be aware that any evidence to the contrary may leave you susceptible to charges of plagiarism.

Acknowledgement

General thanks that you may want to give.

Table of Contents

List of Figures

Note the figure numbering: the chapter number and a sequence number within that chapter. The caption should be short & meaningful.

List of Tables

Note the table numbering: the chapter number and a sequence number within that chapter. The caption should be short & meaningful.

List of Acronyms

Chapter 1 - Introduction

Length of dissertation. This depends somewhat on the project, but 3,000 - 8,000 words will suffice in most cases. Survey reports, or reports on projects which have required extensive reviews of the literature, will usually be longer than reports of a mathematical or experimental nature. A balanced presentation is more important than the precise length, but the length must not exceed 20,000 words.

You can assume that your reader has the same general background knowledge of computer science (but not necessarily of IoT) as one of your fellow-students who is doing a different project. You then need to put in enough specific background information in your Introduction to enable such a student to understand what your project is about.

The order and titles of the sections/chapters in this will, in most cases, follow broadly the typical pattern of a substantial paper in a journal or of a MPhil thesis. **You should consult with your supervisor about the precise layout most relevant to your project.** Such a consultation will be most profitable if you present your supervisor with an initial draft list of the section headings you think you should use.

Note that the chapter headings, sections may differ from dissertation to dissertation according to your preference of presenting your research work. However, try to maintain uniformity in having the basic sections and most importantly the format so that we have uniform dissertations coming out of your batch.

Formatting guide prepared by Dr. H.E.M.H.B. Ekanayake | V1.0_27082017 – This document is under development and it is based on several external sources such as "A

Structured Approach to Presenting Theses: Notes for Students and Their Supervisors"
by Chad Perry.

Few other considerations:

- Each chapter should start in a new page;

- Try to avoid the use of active voice (use of "I" or "we"), however, sometimes you have to use it in places like Preface and Conclusions;

- Each table and figure should be referred in main text before they appear;

- Figure captions should be put below the figure;

- Try to convert any diagram, equation, etc. to a picture (jpeg, gif, etc. with sufficient resolution or dpi) when you insert them into this document as sometimes such diagrams or equations make problems during formatting or printing;

- Table captions should be put above the table;

- For the draft report, you can print on both sides;

- For the hard-bound report, you should print on single side of paper;

- Fig. 1.1 shows the formatting guide used by previous batches; however, you can use this document itself as your formatting guide.

Description	Draft Report	Hard Bound
dissertation text Times New Roman	12pt	12pt
text in tables and code listing	11pt	11pt
line spacing (preface and main text)	1.5	1.5
line spacing appendices	1.0	1.0
left margin	37mm	37mm
top/bottom/right margins	25mm	25mm
chapter heading	24pt bold	24pt bold
section headings	16pt	16pt
subsection headings	14pt	14pt
other headings	12pt bold	12pt bold
tables headings font	11pt bold	11pt bold
printing	on both sides of paper use mirror image option	Single sides of paper

Table 5.1: Typesetting recommendations

Figure 1.1: Formatting guide

Number Figures with "Figure <Chapter#>.<Figure Number in Chapter>". The font size is 11 pt.

Follow it for Tables as well. See Table 1.1 below.

Table 1.1: Sample Table

1.1 Background to the Research

This section outlines the broad field of study and then leads into the focus of the research problem. This section is short and aims to orient the readers and grasp their attention.

A dissertation should be able to reference at least four or five writers in the first one or two paragraphs, to demonstrate from the start of the dissertation that care has been taken to acknowledge and chart the depth and breadth of the existing body of knowledge. Most of the material in this section is covered in more detail in later sections such as section 1.3, and so these sections will have to be referred to, and section 1.1 is usually only about one or two pages maximum. For these reasons, section 1.1 is often one of the last sections of chapters 1 and 2 to be written.

1.2 Research Problem and Research Questions

The research problem is one or two sentences that cannot be answered 'yes' or 'no'; it is the broad problem that the researcher will examine more precisely in the hypotheses (or using objectives).

1.3 Justification for the research

Examiners are concerned that the student has not addressed a trivial research area. It is not enough to show there are gaps in the body of knowledge, they must be important gaps. That is, the research problem should be important on several theoretical and practical grounds.

1.4 Methodology

This section gives an introductory overview of the methodology, and is placed here in chapter 1 to satisfy the initial curiosity of the examiner. This section should refer to sections in chapters 2 and 3 where the methodology is described and justified in far more detail. Max length = 2 pages.

1.5 Outline of the Dissertation

Each section and chapter is briefly described in this section.

1.6 Definitions

Definitions adopted by researchers are often not uniform, so key and controversial terms are defined here to establish positions taken by the research. Definitions should match the underlying assumptions of the research and students may need to justify some of their definitions. A definition of a core construct may be discussed in depth at the beginning of chapter 2, and defining the construct in this section can merely present the definition and refer to the discussion in chapter 2. Perhaps the student could make some minor changes to a standard definition to make it particularly appropriate to the dissertation.

1.7 Delimitations of Scope

This section 'builds a fence' around the research findings which are additional to the limitations and key assumptions established in the previous section about definitions. For example, the explicit boundaries of the research problem described in section 1.2 above should be noted again in this section and other implicit boundaries should be clearly expressed. Other delimitations could be the environmental factors, and variables that could not be controlled. In effect, the 'population' about which findings are to be made, is outlined here. In most dissertations, other limitations caused specifically by the methodological methods chosen are placed in chapter 3 rather than in this section.

In this section, the researcher is trying to forestall examiners' criticisms, so justifications for these delimitations must be provided in the section. It would be wise not to emphasize that 'time' and/or 'resources' were major influences on these delimitations of the research for an examiner may think that the student should have chosen a research project that was more appropriate for these obvious limitations of any research.

Note: limitations are beyond the researcher's control while delimitations are within his or her control. The first term is common in US dissertations and is suggested here as referring to the planned, justified scope of the study beyond which generalization of the results was not intended.

1.8 Conclusion

Example:

"This chapter laid the foundations for the dissertation. It introduced the research problem and research questions and hypotheses. Then the research was justified, definitions were presented, the methodology was briefly described and justified, the dissertation was outlined, and the limitations were given. On these foundations, the dissertation can proceed with a detailed description of the research."

Chapter 2 - Literature Review

The purpose of this chapter is to provide the <u>typical reader</u> with information that they cannot be expected to know, but which they will need to know in order to fully understand and appreciate the rest of the dissertation. It should explain why the project is addressing the problem described in the dissertation, indicate an awareness of other work relevant to this problem and show clearly that the problem has not been solved by anyone else. This section may describe such things as:

- the wider context of the project;
- the problem that has been identified;
- likely stakeholders within the problem area;
- any theory associated with the problem area;
- any constraints on the approach to be adopted;
- existing solutions (i.e. related work) relevant to the problem area, and why these are unsuitable or insufficient in this particular case;
- methods and tools that your solution may be based on or use to solve the problem.

Finally, you should provide a conclusion of the chapter.

Chapter 3 - Design

Recommended sections:

- Justification for the methodology in terms of the research problem and the literature review;
- The unit of analysis and subjects or sources of data;
- Instruments or procedures used to collect data;
- Limitations of the methodology if they were not explicitly discussed in section 1.7;
- Any special or unusual treatments of data before it was analyzed;
- Computer programs used to analyze the data, with justifications for their use; and
- Ethical issues.

Make sure to give a detailed high-level architecture or a conceptual model of your project here or in the next chapter.

Chapter 4 - Implementation

This chapter describes the software solution at a finer level of detail, down to the code level. This chapter is about the realization of the concepts and ideas developed earlier. It can also describe any problems that may have arisen during implementation and how you dealt with them.

Do not attempt to describe all the code in the solution, and do not include large pieces of code in this chapter (detailed source code can be provided in Appendix C and also be included in the CD). Just pick out and describe the pieces of code which, for example:

- are especially critical to the operation of the solution;
- you feel might be of particular interest to the reader for some reason;
- illustrate a non-standard or innovative way of implementing an algorithm, data structure, etc.

You should also mention any unforeseen problems you encountered when implementing the solution and how and to what extent you overcame them. Common problems are:

Chapter 5 - Results and Evaluation

In this chapter, you should describe to what extent you achieved your goals/objectives based on the results (or evidence).

Chapter 6 - Conclusions

This chapter is the most important chapter of the dissertation, for after ensuring the methodology and research processes are sound, the examiners will spend much time studying this chapter.

6.1 Introduction

6.2 Conclusions about research questions (aims/objectives)

Findings for each research question or hypothesis are summarized from chapter 5 and explained within the context of this and prior research examined in chapter 2; for example, with which of the researchers discussed in chapter 2 does this research agree or disagree, and why? Disagreement suggests the research is making a contribution to knowledge and this contribution of the research should be clearly developed.

6.3 Conclusions about research problem

Based on section 6.2, implications of the research for furthering understanding of the research problem is explored. You are warned that examiners are careful that conclusions are based on findings alone, and will dispute conclusions not clearly based on the research results. This section concludes with a summary listing of the contributions of the research together with justifications for calling them 'contributions'.

6.4 Limitations

Section 1.7 has previously outlined major limitations of the research that were a deliberate part of the research. This section discusses other limitations that became apparent during the progress of the research.

6.5 Implications for further research

This final section is written to help other researchers in selection and design of future research. Further research could refer to both topics and to methodologies or to both. Removing some limitations mentioned in section 1.7 usually provides opportunities for further research.

References

Wherever possible, references should be specific and to the primary published source material, i.e. to a journal article or conference proceedings or to a book, and not to a web page containing reference to the primary source. (The reasons for this should be obvious – web pages often have a limited lifetime. In addition, web material is not subject to any quality control in the form of peer review.)

The reference style should follow either the "Harvard" (or APA) style or be numbered strictly (e.g. IEEE) in the order in which they appear in the text. In the former, reference is made in the text by giving the name of the author and the year of publication in parentheses e.g. "...the large polarization (Moffat 1971a) is accounted for...." or "....Jones (1990) found that....". In the case of more than two authors, use the form "....Budd et al (1971)....." At the end of the text, the references should be listed in alphabetical order of the first author (and chronological order for several papers by one author). All authors of a paper should be given in this listing. The style should be:

Budd W F, Jones M and Radol C, 1971 *Rep Prog Phys*, **31** 1-70.

Moffat, P.H., 1971a *Mon Not R Astr Soc*, **153** 401-418.

If you use reference numbers, this should appear in the text either as a superscript, Budd et al[27], or in square brackets, Moffat [32]. Some word-processing or editing packages (e.g. Word and Latex) will take care of the numbering for you. Without such a tool, renumbering the references each time you modify the text becomes difficult and it is hence often easier to use the Harvard style.

Appendix A: Publications

Your publications if any.

Appendix B: Diagrams

Additional diagrams related to your research project.

Appendix C: Code Listings

Detailed code of software implementations.