(temp) Using Word Embedding for Citation Recommandation in FinTech Scientific Articles

by Ting-Chun Chen, BSc

Submitted to The University of Nottingham

September 2023

in partial fulfilment of the conditions for the award of the degree of

Master of Science in Data Science

I declare that this dissertation is all my own work, except as indicated in the text

Contents

Abstract

According to the guidelines the abstract should not exceed 300 words. However, it is unlikely that anyone will be counting when you submit. If you still wish to do so, then better use Emacs `count-words` command.

Acknowledgements

It is good to thank here your supervisors and any sponsoring bodies, as well as any family, friends, cats, dogs etc. that have been supportive during your time at University.

Chapter 1

Introduction

1.1 (About Citation)

Citations are used to demonstrate research background, existing techniques and pieces of evidence for a statement, playing a critical role in scientific writing. Authors can properly acknowledge the source of information of a statement and avoid plagiarism with an accurate citation. It also serves as a verification that the idea of this statement is provided and supported by previous studies. These days, references and citations also include additional information for search engines and citation recommendation systems to recognise the subfields and similar research of this article.

There has been a noticeable increase in the number of scientific articles being published. 3.8 million scientific articles are being published in 2022, according to the Web of Science database. It is expected that there would be more scientific articles available on the internet. It can be harder for academics to absorb all the new perspectives with their exact sources.

When looking for reference papers, keywords are commonly used in most academic databases and search engines. The authors would then review the results of papers from search engines to evaluate their relevance to our research and reliability. It could cost a huge amount of time and effort to go through full papers, not to mention that the academic database and search engines might miss some relevant papers.

Some academic databases come with citation recommendation tools, which provide relevant articles that share the same category with or are similar to our research. Citation recommendation tools would have a bias towards published journals, impact factors, times being referred to, and the availability of the journal if it's open to everyone or subscription-only.

1.2 Natural Language Processing

Natural language processing (NLP) refers to the approach of giving computers the ability to understand the meaning and thoughts of words, sentences and articles just like human beings do, building a way of communication between computer and human language. However, human languages have several characteristics that make it tricky to process them. The variability and informality, such as different vocabularies, syntax and phrases in different cultures or individuals representing a similar meaning, bring difficulties to applying a general transformations pipeline for information extraction. A large amount of noise such as mis-spellings, irrelevant contexts and grammar errors confuse not only human readers but also machines and affect the performance of language processing algorithms, not to mention that people might have different understandings of the same sentence which yield even more inaccuracies. Some advanced circumstances such as sarcasm, irony and the speaker's intention, whose implied meanings are beyond plain text, are more difficult for algorithms to detect and extract straightforwardly. Problems in normal vector data such as missing values and lack of labelling occur in natural language as well.

NLP manages to overcome the challenges of processing human language mentioned above and enables machines to analyse, cluster or classify not only vector instances but also texts. Advanced applications of NLP, such as sentiment analysis, text clustering, text summarisation, human language generation and improving automatic translation, provide solutions and assistance to a wide range of tasks. It reduces our workloads and time-spending and helps improve the interaction between humans and computers by mimicking human language to express in an easy-understanding way for most o the people not only programmers.

1.2.1 Second Level Heading

Sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text.

Third Level Headings Do Not Appear in the Table of Contents

Sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text (Lamport, 1994)

References

Lamport, L. (1994). *LaTeX : A document preparation system* (Second ed.). Addison-Wesley.

Appendix A

Supplementary Materials I

Sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text.

Appendix B

Supplementary Materials II

Sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text sample text.