



Master Thesis  
Defense

# **Systematic Review and Meta-Analysis – Processes Towards Selection Automation**

Randy Ambrósio Quindai João  
raqj@ic.ufal.br

**Advisors:**

Prof. Dr. André Luiz Lins de Aquino  
Prof. Dr<sup>a</sup>. Fabiane Queiroz

Maceió, August 28, 2021

Randy Ambrósio Quindai João

# **Systematic Review and Meta-Analysis – Processes Towards Selection Automation**

Dissertation presented as a partial requirement for defense to obtain the degree of Master by the Informatics Master Course of Institute of Computing of Universidade Federal de Alagoas.

Advisors:

Prof. Dr. André Luiz Lins de Aquino

Prof. Dr<sup>a</sup>. Fabiane Queiroz

**Catálogo na fonte**  
**Universidade Federal de Alagoas**  
**Biblioteca Central**

Bibliotecário: Cláudio César Temóteo Galvino – CRB4/1459

Q7s      Quindai João, Randy Ambrósio.  
            Systematic review and meta-analysis: processes towards selection automation  
            / Lucas Moura Nutels. – 2018.  
            47 f.

Orientador: André Luiz Lins de Aquino.  
Co-orientador: Fabiane Queiroz.  
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas,  
Instituto de Computação, Maceió, 2021.

Bibliografia: f. 38-47.

1. Processamento de Linguagem Natural (PLN). 2. Alocação de Dirichlet  
Latente tradicional (LDA). 3. Revisão Sistemática da Literatura. 4. Sistemas de  
Informação. 5. Automação. I. Título.

CDU: 681.5



UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL  
**Programa de Pós-Graduação em Informática – PPGI**  
**Instituto de Computação/UFAL**  
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins  
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401



## Folha de Aprovação

RANDY AMBROSIO QUINDAI JOÃO

### SYSTEMATIC REVIEW AND META-ANALYSIS – PROCESSES TOWARDS SELECTION AUTOMATION


Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 28 de agosto de 2021.

#### Banca Examinadora:

  
\_\_\_\_\_  
Prof. Dr. ANDRÉ LUIZ LINS DE AQUINO  
UFAL – Instituto de Computação  
**Orientador**

  
\_\_\_\_\_  
Professora Dra. FABIANE DA SILVA QUEIROZ  
UFAL – Centro de Ciências Agrárias  
**Coorientadora**

  
\_\_\_\_\_  
Prof. Dr. RIAN GABRIEL SANTOS PINHEIRO  
UFAL – Instituto de Computação  
**Examinador Interno**

  
\_\_\_\_\_  
Prof. Dr. JORGE ARTUR PECANHA DE MIRANDA COELHO  
UFAL – Faculdade de Medicina  
**Examinador Externo**

# Acknowledgements

First and foremost, I thank God.

To my daughters Lana and Briana Quindai, which are the Sun of my life.

Apart from the efforts of me, the success of this thesis depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this thesis.

I would like to show my greatest appreciation to Dr. Andre Aquino and Dra. Fabiane Queiroz for the tremendous support and help. Specially Dr. Andre Aquino for encouragement and guidance through all my career.

Last but not least, I thank my wife Dayane Mota, my Mom Elsa Ambrósio and my father Matheus Quindai.

The list of important people in my life is not a big one, but I owe my success to the people that unconditionally supports me, my brothers and sisters: Délcio Carraco, Elsa Carraco, Mara Quindai, Milton Quindai and Soriano Quindai.

Amazingly, my world revolves around these people.

May God bless us!

# Abstract

Public health evidence and other disciplines are being produced and published at an unprecedented rate and scale. This evidence can take many forms, the more common and conventional types and sources of published evidence include journal articles, conference abstracts/proceedings, technical reports, and clinical trial records/registries. To keep pace with the rapidly evolving public health landscape, and to respond to the critical needs, issues, and public health crises of today in a timely manner, there is a growing need to explore, leverage and integrate insights from more novel sources of evidence. A common thread across all this evidence is that such data is, at large, stored in a noisy, unstructured format, which makes secondary research-led activities in data extraction, synthesis, and reporting incredibly challenging. Secondary public health research methods, such as evidence synthesis and systematic reviewing, are spreading across all research fields. The aim of this research project was to establish an evidence-based framework for an optimal Natural Language Processing (NLP) solution (including a working prototype) to support public health evidence extraction and synthesis research activity. The latest innovations in artificial intelligence (AI), machine learning, and NLP tools and techniques offer the ability to rapidly extract, analyze, synthesize, and understand unstructured textual data, at scale. Recent breakthroughs in these technologies have led to vastly improved NLP models, which are able to capture and model more complex linguistic relationships than ever before. By providing the ability to assess and analyze large quantities of this data, NLP has opened up vast opportunities. The aim of this research project was to establish an evidence-based framework for an optimal NLP solution (including a working prototype) to support public health evidence extraction and synthesis research activity. The traditional systematic review framework is a feasible starting point, where all steps are predicted and standardized. In order to reduce systematic reviewer burden while maintaining the high standards of systematic review validity and comprehensiveness we, developed a semi-automation screening approach using the reviewer's criteria written in natural language. We offer a simplified topic's extraction too and compare it to the traditional Latent Dirichlet Allocation (LDA). For clustering of studies, we transformed the title, abstract and keywords, into a *wordcloud* for each study, and grouped using a NLP technique called Sentence Boundary Detection for finding and segmenting meaningful individual sentences, studies with same sentences are put together, organized, and clustered by sentences frequency. We achieve the generation of summary for clustered studies using natural language generation. We perform a comparison of Markov Chain Generation with Recurrent Neural Network generation

for quality assessment of the generated text. We obtain data graphics by exploring BIBTEX data already available, and mining relations of semantic changes or author's groups of collaboration. The results methodology follows the best practices for conducting and reporting reviews, thus solving a practical problem effectively with reproducible and repeatable results. These results show that the desired tool is feasible with the current state of the art technology. This work resulted in a startup that delivers products to explore and analyze scientific documents in large scale, and it has been validated by the end user.

**Keywords:** SLR, NLG, Bibliometrics, Information Systems, Quantitative methods, R, NLP.

# Resumo

Evidências científicas na área médica e outras disciplinas estão sendo produzidas e publicadas em uma escala e taxa sem precedentes. Essas evidências podem assumir muitas formas, os tipos e fontes mais comuns e, convencionais de evidência publicada incluem artigos de periódicos, resumos / artigos de conferências, relatórios técnicos e registros de ensaios clínicos. Para acompanhar o cenário da medicina em rápida evolução e para responder às necessidades críticas das crises de saúde pública de hoje em tempo hábil, há uma necessidade crescente de explorar, alavancar e integrar os resultados das novas evidências. Uma linha comum em todas essas evidências é que tais dados são, em geral, armazenados em um formato ruidoso e não estruturado, o que torna incrivelmente desafiador conduzir atividades de pesquisa, síntese e geração de relatórios de dados. Métodos secundários de pesquisa, como a síntese de evidências e revisão sistemática, estão se espalhando por todos os campos de pesquisa. O objetivo deste projeto de pesquisa foi estabelecer uma estrutura baseada em evidências para uma solução ótima de Processamento de Linguagem Natural (PLN) (incluindo um protótipo funcional) para apoiar a extração de informação de artigos científicos na forma de texto de forma automática. As mais recentes inovações em inteligência artificial (IA), aprendizado de máquina, ferramentas e técnicas de PLN oferecem a capacidade de extrair, analisar, sintetizar e compreender rapidamente dados textuais não estruturados em escala. Avanços recentes nessas tecnologias levaram a modelos de PLN amplamente aprimorados, que são capazes de capturar e modelar relacionamentos linguísticos mais complexos do que nunca. Ao fornecer a capacidade de avaliar e analisar grandes quantidades desses dados, o PLN abriu vastas oportunidades. Sendo assim, nossa maior meta foi estabelecer uma estrutura baseada em evidências para uma solução de PLN ideal, com a estrutura da revisão sistemática tradicional, onde todas as etapas são previstas e padronizadas. Procuramos desse modo, reduzir a carga do especialista de revisão, mantendo os altos padrões de qualidade e abrangência disponíveis numa revisão sistemática, desenvolvemos uma abordagem de triagem semiautomatizada usando os critérios definidos pelo revisor escritos em linguagem comum. Também oferecemos uma extração de tópicos simplificada e comparamos com a Alocação de Dirichlet Latente tradicional (LDA). Para o agrupamento dos estudos, transformamos o título, o resumo e as palavras-chaves em uma nuvem de palavras para cada estudo e agrupamos usando uma técnica de PLN chamada Sentence Boundary Detection (Detecção de limite de sentença) para encontrar e segmentar sentenças individuais significativas, assim, estudos com as mesmas sentenças são



colocados juntos, organizados e agrupados por frequência de sentenças. Alcançamos a geração de resumo para estudos agrupados usando geração de linguagem natural. Realizamos uma comparação da Geração de Cadeia de Markov com a geração de Rede Neural Recorrente para avaliação da qualidade do texto gerado. Disponibilizamos gráficos de dados explorando os dados BibTeX e minerando relações de mudanças semânticas ou grupos de colaboração do autor. A metodologia de resultados segue as melhores práticas para a realização e relato de revisões, resolvendo um problema prático de forma eficaz e com resultados reproduzíveis e repetíveis. Esses resultados mostram que a ferramenta desejada é viável com o atual estado da arte da tecnologia. Esse trabalho resultou em uma startup que entrega produtos para explorar e analisar documentos científicos em larga escala, e foi validado pelo usuário final.

**Palavras-chave:** SLR, Revisão Sistemática da Literatura, Geração de Texto, NLG, Bibliometrics, Sistemas de Informação, Métodos Quantitativos, R, NLP, Automação.

# Contents

List Of Figures . . . . .	vii
<b>1 Introduction</b>	<b>1</b>
1.1 General Objectives . . . . .	3
1.2 Specific Objectives . . . . .	4
<b>2 Related Literature</b>	<b>5</b>
<b>3 Concepts and tools</b>	<b>10</b>
3.1 Machine Learning and Algorithms . . . . .	12
3.2 Metrics and Indicators . . . . .	14
3.3 Formal Verification . . . . .	16
<b>4 The proposal</b>	<b>21</b>
4.1 Proposal outline . . . . .	21
4.2 Proposed tool . . . . .	23
4.2.1 Question Answering for Assisted Selection . . . . .	26
4.2.2 Topics Extraction - LDA vs Our Proposal . . . . .	29
4.2.3 Assistant 2 . . . . .	33
<b>5 Final Considerations</b>	<b>36</b>
<b>Bibliography</b>	<b>38</b>

# List of Figures

1.1	Systematic Literature Review phases, macro detailing of the phases and average completion time. . . . .	2
3.1	Pyramids of evidence quality. Comparison of bias in medicine studies (Left) to level of acceptable generalized evidence (Right). . . . .	11
3.2	Recurrent Neural Network - Encoder → Decoder. Example for summarization of a <i>corpus</i> manipulated by a language model (RNN-LM). Sentence generated given an input of sentences (bag of words). . . . .	13
3.3	Process of input data to support conventional scientometrics vs Broadening multiple productivity indicators (Mugnaini et al., 2017, p. 82) . . . . .	15
3.4	The topic <i>simplex</i> for three topics embedded in the word simplex for three words. The corners of the word simplex correspond to the three distributions where each word respectively has probability 1. The three points of the topic <i>simplex</i> correspond to three different distributions over words. The mixture of <i>unigrams</i> places each document at one of the corners of the topic <i>simplex</i> . LDA places a smooth distribution on the topic simplex denoted by the contour lines. Source: (Blei et al., 2003). . . . .	18
4.1	Systematic review phases selected for automation (orange). . . . .	21
4.2	The flow of systematic literature review presented in the view of the proposed tool. . . . .	23
4.3	Pipeline of data preparation and our proposal language model. . . . .	26
4.4	Assisted selection in action . . . . .	27
4.5	Pages of the generated document after automation . . . . .	28
4.6	M' classes from resulting trimmed clusters. Sentences generated by Spacy's named entities. . . . .	29
4.7	LDA topic modeling for each k' topic suggested for the fog data, see Figure 4.8 . . . . .	30
4.8	Best k-topics suggested using ldatuning. Two maximization metrics and two minimization functions. . . . .	30
4.9	LDA Perplexity tests - Steps to minimize perplexity by maximizing probability . . . . .	31
4.10	<b>Result 1</b> - Generated text for the group "embedded system", Markov chain method only. . . . .	32
4.11	<b>Result 2</b> - Generated text for the group "embedded system", Spacy's Named Entities combined with Markov chain. . . . .	32
4.12	<b>Result 3</b> - Generated text for the group "embedded system", model GPT2-Tensorflow trained in the corpus of selected group with Attention. . . . .	32
4.13	<b>Result 4</b> - Generated text for the group "embedded system", language model 774M GPT2-Tensorflow Zero Shot. . . . .	33
4.14	Bibliometrics extracted from studies. . . . .	34
4.15	Bibliometrics extracted from studies . . . . .	35

# 1

## Introduction

**W**e consider science a systematic way of discovering how the universe behaves, or the body of accumulated knowledge of discoveries of all existing things. Nowadays, we widely define science as a knowledge-based in data with reproducible methods.

One of the most brilliant advances in humankind's knowledge, such as the evolution of digital computers, occurred in healthcare on methodologies to fight infectious diseases, which in the last century victimized millions of people (Short et al., 2018; Nii-Trebi, 2017). A methodology that is mature enough to systematically assess the efficacy of a treatment or drug with extremely high population variability is the Systematic Literature Review (SLR) (Garfield, 1987; Mulrow, 1987; Morgan, 1986).

A process based on a formulated question that uses systematic methods and reproducibility to identify, select, and critically evaluate all relevant research is called SLR. Usually, it presents three distinct phases:

- **Phase 1** - Planning - It is the first step before undertaking a systematic review. The authors evaluate the need for a systematic review. They perform an exploratory search for an existing review on the same subject; identify a knowledge gap, and formulate a review question.
- **Phase 2** - Conducting the Review - this is the most extensive phase. It starts with a protocol registration, followed by the selection of peer-reviewed articles (stratification by title and abstracts), extraction (full studies assessments) and the review writing;
- **Phase 3** - Dissemination - this final phase is as important as previous phases. Reporting the review allows the community to share the findings. It enables others to replicate, interpret, and evaluate the applied methods.

Conducting the review comprises protocol registration, selection, extraction, and writing. Figure 1.1 depicts macro steps for each phase. Selection is the manual screening of titles and abstracts, include/exclude studies based on inclusion and exclusion criteria defined in the protocol,

appraises only titles and abstracts. Extraction is the manual full reading of studies. Exclusions are applied if the study does not align with the review question. The authors, usually, perform the writing backed on checklists and flow diagrams: PRISMA (Moher et al., 2009), GRADES (Guyatt et al., 2008), and others that could benefit from automation where could apply. Chapter 4 presents the detailing of the techniques used to contribute for: time reduction of SLR writing, quantitative analysis (bibliometrics or meta-analysis), avoid data management nightmare mitigating the process inefficiencies for all kinds of review types.

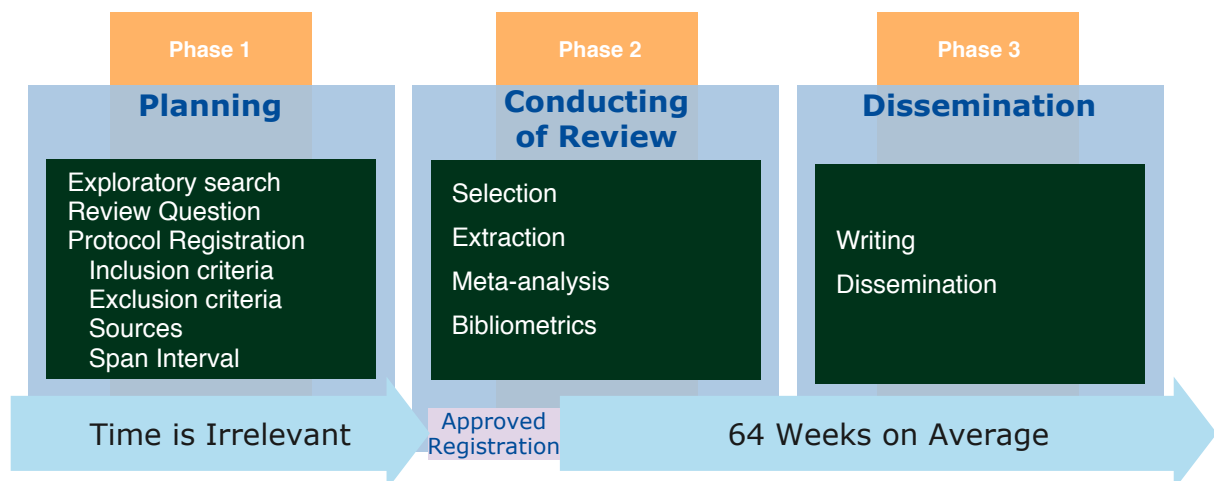


Figure 1.1: Systematic Literature Review phases, macro detailing of the phases and average completion time.

In particular, Kitchenham (2004) had quite a success applying SLR methods to summarize software engineering evidence, according to the following three guidances in health care for SLR: Cochrane (Higgins et al., 2008), National Health & Medical Research Council Australia (Glasziou et al., 2000) and CRD (Khan et al., 2001). Those guidances agree that the level of evidence offered by SLR is the utmost method to reduce bias at all levels in evidence-based knowledge. The systematic review is a secondary study that has the main purpose of mapping primary studies, and such, all kinds of related studies as well.

The process of conducting SLR, especially for new authors, will prove to be a worthwhile endeavor. Systematic reviews are a complicated, multi-step research method that requires a lot of time and statistics skills, Peričić and Tanveer (2019); Wormald and Evans (2018). It is important to note that the literature review is quite different from SLR. A systematic review is an analysis of all primary literature that exists on a specific topic. Primary literature includes only original research articles. Systematic reviews use original research articles to perform the Meta-analyses and qualitative assessment, and hence, we consider them secondary sources.

We can enrich the critical appraisal and synthesis by *meta-analysis*, a phrase coined by Glass and Smith (1979). In general, meta-analysis is the process of collecting and evaluating the data used in studies, thus, conducted in broader areas of human knowledge Kitchenham (2004); Kraus et al. (2020); Mallett et al. (2012).

The concern to speed up production has already raised the attention of systematic review practitioners (Marshall and Wallace, 2019). They estimate that conducting a single review requires more than 1000 h of (highly skilled) manual labor. On average, 67 weeks from registration to publication (Borah et al., 2017). This work explores qualitative analysis (Natural Language Processing) combined with quantitative analysis (bibliometrics).

Given the powerful effects of scientific studies in society, many research fields adopt SLR as a methodology to evaluate all reported breakthroughs. In health care, the lack of standards in the evaluation of studies and poor meta-analysis led Moher et al. (2009) to propose the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), which is a guideline to address several conceptual and practical advances in the science of systematic reviews. Several research fields benefit from review reports, Borrego et al. (2014) used SLR to map reviews in engineering research, Kitchenham et al. (2009) used SLR to map the growing published reviews in the software engineering field.

According to the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR) O'Connor et al. (2018), the tailoring of tools is a challenge for the community, screening via a pipeline of multiple tools, accurate data extraction, translation of available technology to tools with user interfaces, data extraction tools and amongst other broader challenges, extracting data from full texts is one of the open challenges too. Automating even small steps in the systematic review process will shorten the time before reviews are published and increase the number of questions about the reviews created. With time and trust, we will delegate the process to automation (Tsafnat et al., 2013; Thomas et al., 2017).

## 1.1 General Objectives

Systematic Literature Review (SLR) traditionally is a mapping process based on manual extraction of studies. The selection depends entirely on the reading of title and abstract, revealing itself as a tiresome process, mainly on areas with a high volume of publications per year. Therefore, we propose partial automation of this process using text mining by recommending primary studies and statistics extraction.

Systematic review processes have a well-known set of problems for conducting rapid, accurate, and efficient scientific evidence reviews. This work addresses some of (O'Connor et al., 2018, p.3) challenges: i) Designing an application programming interface that meets the needs of multiple scientific domains and goals for different systematic reviews; ii) Meeting review-specific/data-specific challenges; iii) Accurate data extraction of study characteristics, this last one is a challenge by algorithm developers.

## 1.2 Specific Objectives

The main contribution of this work is to assist in the systematic review process. We use natural language processing and text generation of natural language to evaluate the fittest options. Our main focus is in Phase 2, see Figure 1.1. We are interested in reducing the manual workload by processing the data available in *bibtex* files, aiming to evaluate those following experimental units:

1. Reduction of time spent to write conclusions and future breakthroughs using our approach.
2. Quality of generated text.
3. Compare our language model for NLP categorization to the LDA model.
4. Compare our approach for text generation to RNN text generation.

The combination of those experimental units is the key to reduce time without losing the review quality. Chapter 4 will depict in details each unit. We also will explore productivity metrics.

We organize this document as follows. Chapter 1 presents an introduction to the opportunities and challenges. Chapter 3 shows the main concepts and related works. Chapter 4 presents the pretended approach for the challenges and the plan to develop the intended tool. Chapter 5 final considerations and future directions.

This chapter introduces the field of systematic reviews, issues associated, and tools. We also present the proposal and contributions to this work.

# 2

## Related Literature

Several studies in different areas have been using NLP to find hidden relations in *corpus* of interest.

Some works related to machine learning applied to analyze science:

- [Chen and Friedman \(2004\)](#) presents a system called BioMedLEE that extracts a broad variety of biomedical phenotypic data. It uses NLP to extract expert terms specific to biomedical from textual titles. The article raises the problem of massive online literature and manual assessment of biomedical literature. BioMedLEE had 64.0% precision and 77.1% recall, respectively, according to the author's agreements.
- [Cohen et al. \(2006\)](#) developed a classifier algorithm based on a voting perceptron, that showed a significant savings of reviewer effort at the 95% recall level. TREC 2004 Genomics Track document corpus, a general static collection of MEDLINE<sup>1</sup> subset widely used in other experimental systems. It uses only words from the title and abstract. This article shows that automated document classification can provide some value in reducing the labor of manual review, and for about 20% of topics, the reduction is considerable, approaching  $\leq 50\%$ .
- [Blei et al. \(2007\)](#) applies correlated topic modeling (CTM) to analyze articles from *Science* published from 1990–1999. The objective is to classify those studies by topics, to facilitate the catalog of the digital library. CTM explicitly models the correlation between the latent topics in the collection and enables the construction of topic graphs and document browsers that allow users to navigate the collection in a topic-guided manner. They compare the results of CTM and LDA models in the same dataset. CTM presents a better prediction of the remaining words of a document after observing a portion of it. This model provided an analysis of the JSTOR<sup>2</sup> archive for the journal *Science*.

---

<sup>1</sup><https://ebSCO.com>

<sup>2</sup>[www.jstor.org](http://www.jstor.org)



- [Rosen-Zvi et al. \(2010\)](#) shows an unsupervised machine learning technique based on a Markov chain to extract data about authors and topics in a collection. They use experiments based on perplexity scores for test documents, and precision-recall for document retrieval to illustrate systematic differences between the proposed author-topic model and many alternatives. This article describes a generative model for document collections, the author-topic (AT) model, which simultaneously models the content of documents and the interests of authors. They discuss detecting papers that were written by different people with the same name. It extends probabilistic topic models to include authorship information. This model provides significantly improved predictive power in terms of perplexity. Applies the methodology to three large text corpora: 150,000 abstracts from the CiteSeer digital library ([Lawrence et al., 1999](#)), 1740 papers from the Neural Information Processing Systems (NIPS) Conferences
- [Thomas et al. \(2011\)](#) evaluates strengths and weaknesses in the application of text mining technologies, automatic term recognition, document clustering, classification, and summarization to support the identification of relevant studies within systematic reviews. They explore four text mining technologies, automatic term recognition, document clustering, classification, and summarization. The article outlines how text mining techniques could aid in the systematic review process. It uses a pipeline of text mining tools, ASSERT<sup>3</sup> project and TerMine<sup>4</sup>.
- [Wang and Blei \(2011\)](#); [Li et al. \(2013\)](#) proposes a scientific article recommendation system based on matrix factorization and LDA applied in a social media environment. They explore the topic regression Matrix Factorization (tr-MF), to solve the problem for recommending scientific articles, and recommendation for a specific field. CiteULike<sup>5</sup> and Mendeley<sup>6</sup> data were used in both studies.
- [Tian and Jing \(2013\)](#) present a graph-based system for articles recommendation, such that each node represents a researcher connected to a similarity network with other researchers on the same interests. They propose a Bi-Relational Graph model to combine article content and researcher-article readership information in a unified framework for scientific article recommendation system. It is an iterative random walk with restarts technique to predict both article-researcher relevances and researcher-researcher correlations. The solution includes three parts: i) the article content similarity, ii) researcher interest correlation, and iii) researcher-article readership. They use CiteULike<sup>5</sup> data in this study.

Some works related to systematic review tools:

---

<sup>3</sup><http://www.nactem.ac.uk/assert/>

<sup>4</sup><http://www.nactem.ac.uk/software/termine/>

<sup>5</sup><http://www.citeulike.org>

<sup>6</sup><http://www.mendeley.com>

- [Thomas et al. \(2011, 2010\)](#) the Cochrane EPPI-Reviewer4 initiative for text mining techniques for research synthesis. They present this tool as software for all types of literature review, including systematic reviews, meta-analyses, 'narrative' reviews, and meta-ethnographies. Fee-based offers a one-month free trial.
- [Ouzzani et al. \(2016\)](#) is a free web, and a mobile tool for systematic reviews works offline and then syncs back to servers when online. Was explicitly developed to expedite the initial screening of abstracts and titles using a process of semi-automation. It uses a support vector machine (SVM) classifier to compute MeSH terms to labeling and output scores for each study, suggestions for labels based on your pattern of selection, and it learns from your include/exclude decisions. The authors say that their ultimate goal is to support the entire systematic review process, but initially, the focus is on facilitating abstract/title screening and collaboration. Regardless of the semi-automation, Rayyan lacks some features that would benefit from several improvements, including better handling of duplicates, automatic data extraction from full text, automatic risk of bias analysis, and seamless integration with Review Manager (RevMan), the Cochrane software used for preparing and maintaining Cochrane reviews. Rayyan does not support any additional phases of the SR workflow past the screening. It is available for free<sup>7</sup> and funded by Qatar Foundation, a non-profit organization in the State of Qatar.
- [Torres and Adams \(2017\)](#) present RevManHAL, open-source software created in Java, which assists reviewers and produces XML-structured files. Uses editable phrase banks to envelop text/numbers from a prepared readable text for RevMan format. In this way, they create a considerable part of the review's: 'abstract', 'results', 'discussion' sections, and a phrase added to 'acknowledgments'. The Cochrane Collaboration employs Review Manager (RevMan) produced by the Informatics and Knowledge Management Department of the Cochrane Collaboration.
- [Kohl et al. \(2018\)](#) maps a lot of commercial and open-source tools for the systematic review. A critical appraisal on 22 software packages on setting up, scoping/pilot, literature searching, duplicate checking, article screening, data coding, critical, synthesis, and documentation. This article introduces the open-access online tool CADIMA, a tool for data extraction, critical appraisal, and evidence synthesis. It shows a comparison of state-of-the-art tools available for systematic reviews, SESRA and StArt([Molléri and Benitti, 2015](#); [Fabbri et al., 2016](#)) mirrors the stages of systematic reviews in [Kitchenham \(2004\)](#), SLuRp ([Bowes et al., 2012](#)), SLR-Tool ([Fernández-Sáez et al., 2010](#)), DistillerSR <sup>8</sup> ([Matwin et al., 2010](#)), which encompasses all systematic review phases with AI support, fee-based, offers special pricing for students and Cochrane Review Groups. Available in two versions

---

<sup>7</sup><http://rayyan.qcri.org>

<sup>8</sup><https://www.evidencepartners.com/>

(DistillerSR and DistillerCER) with varying features and many others described broadly in the study. Covidence<sup>9</sup> (Couban, 2016) provides support for title and abstract screening, it offers tools for quality assessment and data extraction optimized for Cochrane Reviews, has a free trial option and is free for use in Cochrane. It produces a PRISMA flow diagram that can export additional information to RevMan; it is the result of the Cochrane Collaboration; Australia's Monash University, Alfred Hospital and, National ICT; England's University College; and Argentina's Instituto de Efectividad Clinica y Sanitaria.

- Scells and Zuccon (2018) present a search refiner, an open-source tool to assist in formulating, visualizing, and understanding Boolean queries in a systematic literature review search. This tool is to both experts and novices, as a tool for query formulation and refinement, and as a tool for training users to search for literature to compile systematic reviews. The authors are interested in automatic query transformation and query formulation for systematic reviews<sup>10</sup>. This tool comprises three core components: i) a query interface, ii) a query visualizer, and iii) a query transformation tool. A service similar to this tool is offered by PubMed<sup>11</sup> and Ovid MEDLINE<sup>12</sup> portals.
- Munn et al. (2019) present the JBI System for the Unified Management, Assessment, and Review of Information (SUMARI). It is a word processor, reference management program, statistical and qualitative data analysis program accessible to use web applications for systematic reviews. Fee-based.
- Marshall and Brereton (2015)<sup>13</sup> is a website that brings together a lot of open source tools for a systematic literature review. This reference lists many tools concerning many subjects of a systematic review, encompassing from simple flow diagrams to text mining techniques. Officially there are 187 tools available to support the systematic review process.

This proposal thoroughly compares to those developed solutions presented above. Our main goal is to simplify the complex logistical process of systematic reviews (SR). The massive number of registered solutions shows that there is a gap in the industry; some of them we address in this proposal. Mirroring the manual stages of SR is highly explored by those tools, support of writing, and resume are not. The use of NLP aims to explore the corpus in such a way that humans cannot explore. Our main strengths in comparison to existing tools are:

**Clustering the studies:** SR as a process that uses systematic methods to collect primary data, critically appraise research studies, and synthesize findings qualitatively or quantitatively, is a manual process. Critical appraisal starts by the selection phase, where authors select

---

<sup>9</sup><https://www.covidence.org/>

<sup>10</sup><http://ielab.io/projects/systematic-reviews.html>

<sup>11</sup><https://pubmed.ncbi.nlm.nih.gov>

<sup>12</sup><https://www.ovid.com/product-details.901.html>

<sup>13</sup><http://systematicreviewtools.com/>

studies reading titles and abstracts, a dull and diligent process for people. This proposal supports this phase by offering an automatic selection by topics using sentiment analysis, the sentences to rate positive or negative sentiment, are given by authors in inclusion and exclusion criteria. Furthermore, we perform the coverage of the extraction phase by clustering the selected studies explained in detail in chapter 4.

**Generation of summary for clustered studies:** We provide a sentence representing each group of articles generated at the previous step. It generates a resume for each group using natural language generation (NLG). The practical use of NLG is relatively new. This proposal brings to reality this technology to assist the writing of systematic reviews.

**Data graphics:** It provides data and graphics for further analyses. It is mandatory to support SR with data. Therefore, we provide all data in many formats (CSV, Excel, XML).

We value the role of specialists in the process of qualitative analysis. We integrate those steps into a generated document, editable, and according to the structure of known reporting needs (PRISMA, GRADES, AMSTAR) in an analysis-ready format, with all graphics and generated sections. We provide an online interface to facilitate the process. Our daily living is full of technology. Professionals should have to take advantage of this to perform their activities. We aim to facilitate the use of systematic review methods to conduct studies with better efficiency, easy access through devices, and reduction of time for publication. We also seek to provide customization of structured document's output.

This chapter presents all related literature, from NLP used in different areas of studies to systematic review tools.

# 3

## Concepts and tools

This chapter introduces all concepts used in this work, techniques, and their formal verification. You will find the concepts of systematic literature review (SLR), bibliometry, neural network and metrics in NLP, and text generation techniques.

The scientific literature has many authors that define a systematic literature review (SLR). [Greenhalgh \(1997\)](#) defines it as a broad vision of primary studies using systematic and explicit reproducible methods; [Brereton et al. \(2007\)](#) defines SLR as a formulated question that uses systematic and reproducible methods, to identify, select and critically evaluate all relevant research, by analyzing and collecting all included studies data in the review.

In general, we define SLR as a methodology that has the primary goal of collecting, mapping, and reporting new findings in a specific research field, using reproducible methods with reduced selective bias.

Meta-analysis is a quantitative, formal, epidemiological study design used to systematically assess previous research studies to derive conclusions about that body of research ([Haidich, 2010](#)). The benefits of meta-analysis encompass an examination of variability, quantitative analysis, and heterogeneity in study results. It plays a central role in evidence-based medicine, the strength of the freedom from various biases that beset medical research, meta-analyses are in the top, figure 3.1(a).

This study proposes a generalization of the use of the medical strongest evidence synthesis, in other areas of research, figure 3.1. Observe that systematic reviews and meta-analysis are at the top level of generalized studies. In summary, levels of figure 3.1(b) have followings meanings:

- Level IV refers to the single case study as the least essential evidence, usually offered to views or experiences of one person;
- Level III illustrates practical report of responses;
- Level II refers to theoretical concepts studies; and

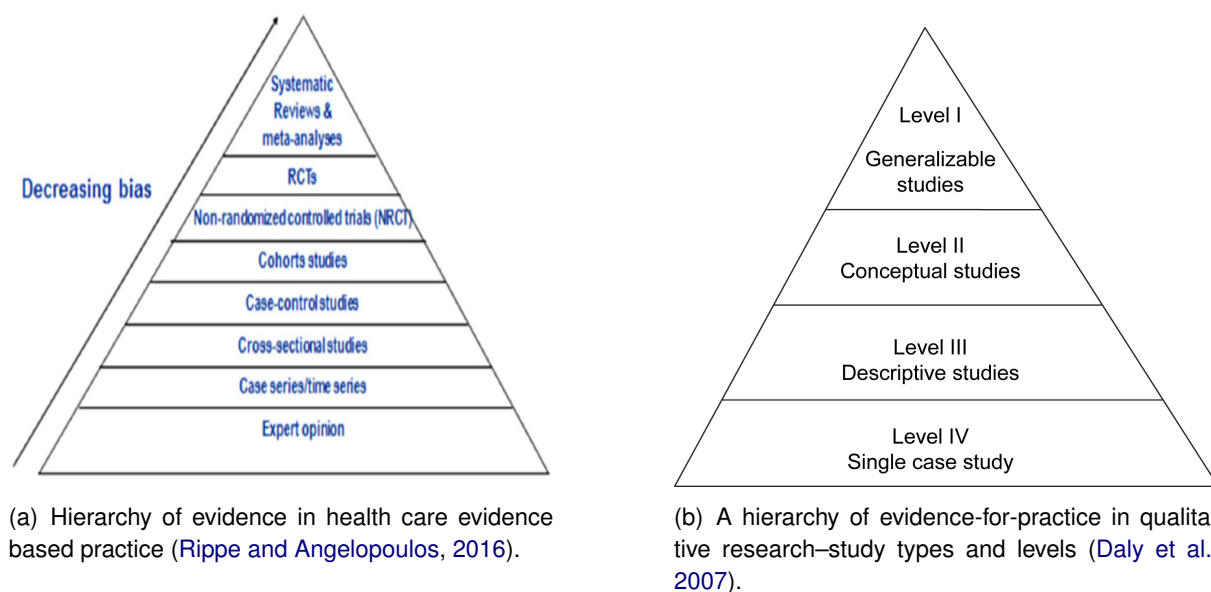


Figure 3.1: Pyramids of evidence quality. Comparison of bias in medicine studies (Left) to level of acceptable generalized evidence (Right).

- Level I focus on theory and the literature by assessing relevance to other settings from comprehensive, clear, and analytical procedures.

The level at the apex of the hierarchy is the ideal, well-developed qualitative studies. This level is the main object of interest in this work.

There are risks in reducing a complex set of professional research procedures to a simple code, but in common with professionals in evidence-based medicine, hierarchies of evidence-for-practice can be used and abused. A lack of understanding of social theory or the literature means that theories and concepts are not entirely used to frame the research process. Practitioners unfamiliar with the qualitative research method's intricacies lack a framework for judging which studies provide a secure basis for practice decisions.

The primary purpose of Level I studies is to indicate future directions, making evident the unavoidable limitations. We reach this proposal by improving the process of conducting qualitative studies. Many examples exist that lead to an inadequate level I reports, there are unavoidable impediments to the ideal research process. Researchers may have insufficient research funding for manual data saturation, lack of broad experience for judging which studies provide a secure basis for practice decisions. This demand opens wide to automation methods, artificial intelligence is evolving and presenting as a mature methodology to support practitioners on qualitative studies at the Level I (Daly et al., 2007).

Policy and practices of Level I evidence rigorously rely on three 'E' initiatives: economy, efficiency, and effectiveness, Tranfield et al. (2003). Evidence-based medicine has already migrated from medicine to other disciplines Kitchenham (2004); Moayed et al. (2006); Denyer and Tranfield (2009).

BIBTEX data, considered as Metadata, which contains the study data, such as year of

publication, funding institution, patents, number of citations, and others. It provides information about how the research fields are evolving, not only demographically, but also economically. The analysis of these data is defined as Scientometrics (Nalimov and Mulchenko, 1971; Callon et al., 1986).

Scientometrics is a term that historically has been overlapping interests with Bibliometrics and Informetrics (Hood and Wilson, 2001; Mingers and Leydesdorff, 2015). Chen et al. (2002) maps the evolution of these terms over time. Therefore, Scientometrics is the study of quantitative aspects of science, communication in science, and science policy.

Bibliometrics (Broadus, 1987) generally is defined as the statistical or quantitative description of a body of literature. Such a definition includes any quantitative measure, e.g., number of titles, number of volumes in a collection, and multi-volume sets (number of articles in a journal, the collaboration between authors). We can enrich such disciplines with Natural Language Processing (NLP) Manning et al. (1999), applied in this work for text processing of title, abstract, and all data presented in a group of selected studies in PDF format.

### 3.1 Machine Learning and Algorithms

Natural language generation (NLG) refers to any text generation for any context. Is commonly used in machine translation, predictive typing, speech recognition, summarization, dialog, creative writing and others (Chopra et al., 2016). The task to predict the next word, given the previous words and a condition  $x$ , is given by:

$$P(y_t|x, y_1, \dots, y_{t-1}) \sim RNN. \quad (3.1)$$

Radford et al. (2019); Sutskever et al. (2011), named Conditional Language Model (CLM), equation 3.1, a language model that assigns a probability to a sequence of words given some conditioning context  $x$ . One of CLM's top tasks used in this work is summarization, where ( $x$  = input text;  $y_1, \dots, y_{t-1}$  = given words;  $y_t$  = generated summary).

Recurrent Neural Network (RNN) Rumelhart et al. (1986), is a deep neural network. This kind of deep neural network is created by applying the same set of weights recursively over a structured input, to produce a structured prediction over variable-size input structures, or a scalar prediction on it, by traversing a given structure in topological order, particularly applied in acyclic directed graphs (Socher et al., 2011; Irsoy and Cardie, 2014). Represented in NLP by language modeling (LM) (Bengio et al., 2003; Bengio, 2000), that is, given a sequence of words  $x_1, x_2, \dots, x_t$ , compute the probability distribution of the next word  $x_{t+1}$ :

$$\prod_{t=1}^T P(x_{t+1}|x_t, \dots, x_1), \quad (3.2)$$

where  $x_{t+1}$  can be any word in the vocabulary  $V = \{w_1, \dots, w_{|V|}\}$ ,  $T = |V|$ . These chunks of  $n$  consecutive words are called  $n$ -gram,  $n$  is the number of words in the sentence's chunk. In literature these models are called RNN-LM.

The goal of an RNN implementation is to enable propagating context information through faraway time-steps. Figure 3.2 represents an RNN-LM for a set of source sentences as the input layer, and a target sentence result of summarization as the output layer. The black box represents the bag of words and data preparation steps, the orange box represents the input layer, and the box with green circles represents the output layers. Meanings of variables in Figure 3.2 are:  $W$  - Bag of Words of input sentences;  $V$  - vocabulary of  $W$  words;  $x_t$  - input word vector at time  $t$ ;  $\sigma$  - non-linearity function to compute the hidden layer output features at each time-step  $t$ ;  $s$  - the hidden state output probability distribution over the vocabulary at each time-step  $t$ . To calculate the next word multiply  $x_{t-1}$  and  $x_t$  by different weights.

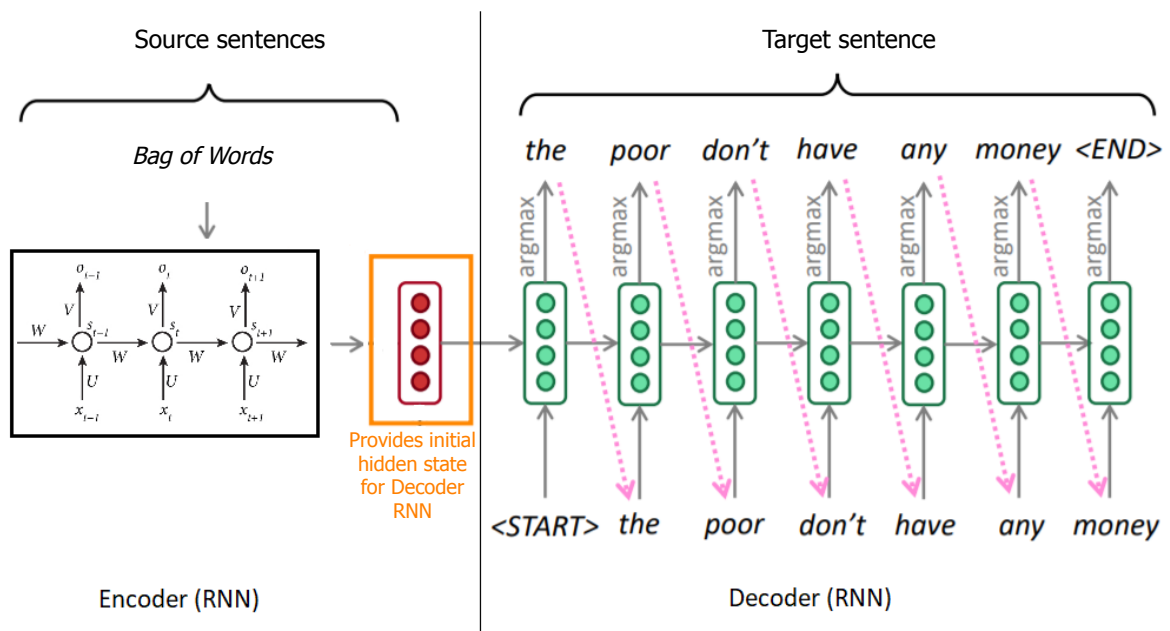


Figure 3.2: Recurrent Neural Network - Encoder → Decoder. Example for summarization of a corpus manipulated by a language model (RNN-LM). Sentence generated given an input of sentences (bag of words).

The state-of-the-art (SOTA) of RNN-LM shows that the task to generate text could lead to unexpected results. Incoherent sentences, repetition, sparsity problems, no symmetry in how the inputs are processed, and no guarantees about the accuracy, and can be offensively wrong (Bengio et al., 1994; Jurafsky, 2000; Goodfellow et al., 2016).

Neural Machine Translation (NMT) is the task of translating a sentence  $x$  from one language to  $y$  in another language (Bahdanau et al., 2014; Graham et al., 2014). NMT uses a single neural network comprised of two RNNs:

- i) Encoder RNN: Extracts all of the pertinent information from the source sentence to produce an encoding.



- ii) Decoder RNN: A language model that generates the target sentence conditioned with the encoding created by the encoder.

Basically, this neural network architecture is called sequence-to-sequence (seq2seq) [Sutskever et al. \(2014\)](#), and it is a CLM, see Figure 3.2. The decoder is predicting the next word of the target sentence conditioned to the source.

To evaluate these models [Papineni et al. \(2002\)](#) proposes the metric Bilingual Evaluation Understudy (BLEU), ROUGE ([Lin and Och, 2004](#)), METEOR ([Banerjee and Lavie, 2005](#)), F1-score and Perplexity ([Brown et al., 1993](#)). All of them are not ideal for machine translation tasks and are much worse for summarization, which is more open-ended than machine translation. Perplexity can capture how powerful the language model is, but unaffected on text generation tasks. We have no automatic metrics to capture overall quality, which is an open challenge adequately, but there are more focused topics to capture particular aspects of the generated text: fluency, diversity, similarity measures, repetition, and others. Though these do not measure overall quality, they can help track some important qualities. Later on, we will dive more on Perplexity details.

Markov Chain ([Geyer, 1992](#)) is a general method for the simulation of stochastic processes having probability densities known up to a constant of proportionality. It can be used to simulate a wide variety of random variables and stochastic processes and is useful in Bayesian, likelihood, and statistical inference. This strategy is well-studied with vast literature. This work is used as a text generation technique to predict the next word given the previous one.

## 3.2 Metrics and Indicators

Electronic databases facilitate the survey of new publications, even though the databases use different representations of the data. The researchers have highly discussed the growing use of bibliometric data as productivity indicators ([Persson et al., 2004](#); [Costas and Bordons, 2007](#); [Durieux and Gevenois, 2010](#)), as well as the unification of citation indexes at different databases ([Van Raan, 2005](#); [Aguillo, 2012](#); [Orduna-Malea et al., 2017](#)). Figure 3.3 describes a transition proposal from conventional scientometrics to a broadening multidimensional representation, exposing the freedom to explore these indicators by appraisal methods to describe the publication impact better.

Several metrics for productivity measure add value to scientometrics indicators, there are three types of indicators that are worth to take note ([Durieux and Gevenois, 2010](#)):

**Quantity Indicators:** intended to measure the productivity of a researcher or a group.

- According to [Burrel \(2001\)](#), given a  $\lambda$  mean in a *Poisson* distribution, if we assume a Gamma distribution for citations variability over time, the obsolescence of citations is

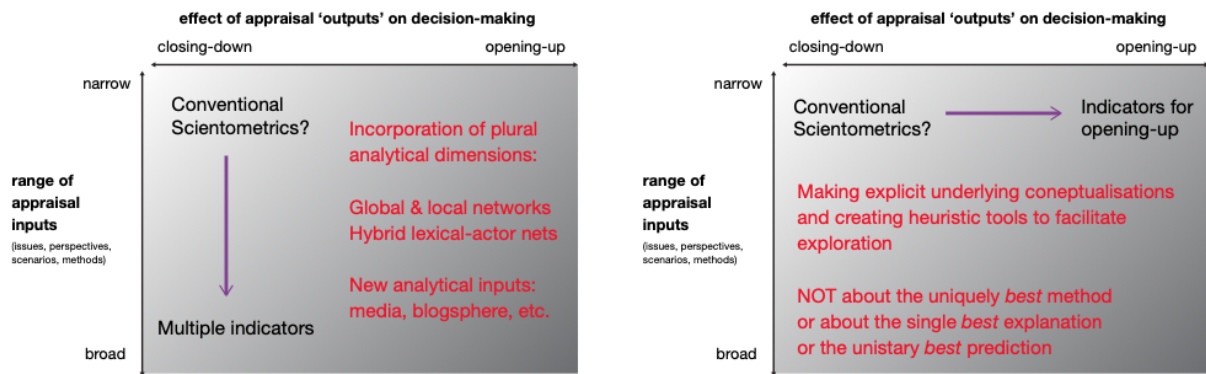


Figure 3.3: Process of input data to support conventional scientometrics vs Broadening multiple productivity indicators (Mugnaini et al., 2017, p. 82)

given by the binomial form:

$$P(X_t = r) = \binom{r+v-1}{v-1} \left(\frac{\alpha}{\alpha+t}\right)^v \left(1 - \frac{\alpha}{\alpha+t}\right)^r, \quad r = 0, 1, 2, \dots \quad (3.3)$$

where  $\lambda = vt/\alpha$ , variance =  $vt(t + \alpha)/\alpha^2$ ,  $v$  and  $\alpha$  are empirically determined parameters. This distribution is expected to be strongly asymmetric.

**Performance Indicators:** used to gauge the impact of the research on the scientific community, usually calculated by the number of citations locally.

- *Journal-to-field impact score* measures the average number of cited articles in a specific *journal* and compares it with others *journals* in the same research field category.
- *Eigenfactor* is the citing  $c$  quality measure. It weights the *journal* citations through its impact on the scientific community.
- *Crown* is calculated by dividing the average number of received citations (from a researcher or a research group) by the average number expected for publications of the same type, during the same year, and published in journals within the same field. A crown indicator of 0.9 indicates that the publications from this researcher or this research group are cited 10% less than the world average in their particular field; a crown indicator of 1.2 indicates that the publications from this researcher or research group have 20% more citations than the world average in that field.

**Structural Indicators:** This reflects the article’s quality by the citation frequency in other local articles, i.e., connections between publications, authors, and areas of research.

- **Zipf (1932)** postulated that the number of occurrences in a text is inversely related to frequency, that is, the most frequent word will occur twice as often as the second

most frequent word and three times more than the third:

$$f(r) = \frac{1/r^s}{\sum_1^N (1/n^2)} N, \quad (3.4)$$

where  $r$  is the classification,  $f(r)$  the frequency in that classification,  $N$  word count,  $s$  an adjustment parameter.

The community widely explores those metrics to calculate bibliometric indicators of studies, section 4.2.3 gets insights on the relations of them with the scientific community.

However, on structural indicators, [Leydesdorff \(2002\)](#) says that the evolutionary perspective changes the time horizon, that is, jargon and technical terms change their meaning over time, reinforcing the importance of productivity indicators for correct mapping of research impact over time.

A first step in identifying the content of a document is determining which topics that document addresses ([Griffiths and Steyvers, 2004](#)). The study of Natural Language Processing (NLP) allows enlarging the analysis of scientometrics indicators through text analysis. NLP refers to the way humans communicate with each other, speech and text. It is a field that has raised interest for more than half a century, with origins in the field of linguistics ([Chomsky, 1956](#); [Martinet and Palmer, 1966](#)), evolving later to computational linguistics ([Winograd, 1971](#); [Woods, 1970](#); [Gazdar et al., 1985](#)), and finally to NLP ([Harris, 1984](#); [Brownlee, 2017](#)).

### 3.3 Formal Verification

Linguistics is the scientific study of language, including grammar, semantics, and phonetics. Computational Linguistics is the set of computational tools with statistical or rule-based modeling of natural language. NLP is a machine learning technique that arises to improve the interaction between computers and humans ([Blei, 2012](#)). In particular, it defines how to program the computer to understand human speech and writing. NLP is difficult and complicated; some famous sayings in NLP describe better the challenges.

"It is hard from the standpoint of the child, who must spend many years acquiring a language [...], it is hard for the scientist who attempts to model the relevant phenomena, and it is hard for the engineer who attempts to build systems that deal with natural language input or output."

— ([Kornai, 2007](#), p. 248)

"Human language is highly ambiguous [...]. It is also ever-changing and evolving. People are great at producing language and understanding language and are capable of expressing, perceiving, and interpreting very elaborate and nuanced meanings. Simultaneously, while

we humans are great users of language, we are also very poor at formally understanding and describing the rules that govern language."

— (Goldberg, 2017, p. 1)

The urge of analyzing big volumes of text Blei et al. (2003) proposed the model *Latent Dirichlet Allocation* (LDA), which is a generative probabilistic model for collections of discrete data such as text corpora, which in practice, is a three-level hierarchical Bayesian model, in which it models each item of a collection as a finite mixture over an underlying set of topics. There are some terms that we must define before to proceed:

**word** is the basic unit of discrete data, defined as an item from a vocabulary indexed by  $V$ .

**Document** is a sequence of  $N$  words denoted by  $w = (w_1, w_2, \dots, w_N)$ , where  $w_N$  is the  $N$ th word in the sequence.

**Corpus** is a collection of  $M$  documents.

**Corpora** is the plural of the corpus, can represent a set of the corpus as well.

Observe in Figure 3.4 the illustration of LDA, assume that each word in documents, over the hidden and observed variables, are generated by a random topic drawn by a distribution with chosen hidden parameters. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for each document  $w$  in a *corpus*  $D$ :

1. Choose  $N \sim \text{Poisson}(\xi)$ .
2. Choose  $\theta \sim \text{Dir}(\alpha)$ .
3. For each of the  $N$  words  $w_n$ :
  - (a) Choose a topic  $z_n \sim \text{Multinomial}(\theta)$
  - (b) Choose a word  $w_n$  from  $p(w_n|z_n, \beta)$ , a multinomial probability conditioned on the topic  $z_n$ .

Thus, a  $k$ -dimensional Dirichlet random variable  $\theta$  can take values in the space  $k - 1$ -simplex of a topic, and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}, \quad (3.5)$$

where the parameter  $\alpha$  is a  $k$ -vector with components  $\alpha_i > 0$  and  $\Gamma(x)$  is the Gamma function.

Given  $\alpha$  and  $\beta$ , joint distribution of a topic mixture  $\theta$ , a set of  $N$  topics  $z$ , is given by the marginal distribution of documents:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left( \prod_{n=1}^N \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta. \quad (3.6)$$

Ultimately, taking the product of the marginal probabilities of single documents, we obtain the probability of a *corpus*:

$$p(D|\alpha, \beta) = \prod_{d=1}^M p(\theta_d|\alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d, \quad (3.7)$$

where the parameters  $\beta$  and  $\alpha$  are corpus-level parameters, assumed to be sampled once in the process of generating a corpus, the parameter  $\theta_d$  is a document-level variable, the parameters  $z_{dn}$  and  $w_{dn}$  are word-level variables. The equations 3.5, 3.6, 3.7, are used further in this text as one methodology for topic classification, chapter 4.

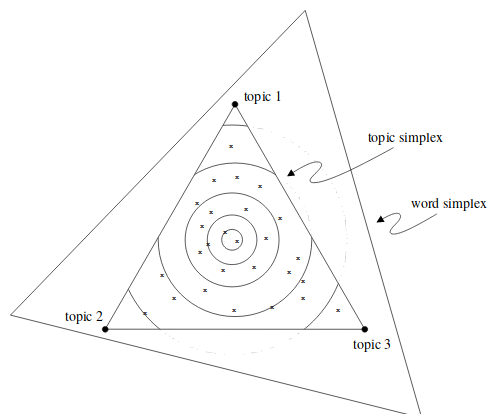


Figure 3.4: The topic *simplex* for three topics embedded in the word simplex for three words. The corners of the word simplex correspond to the three distributions where each word respectively has probability 1. The three points of the topic *simplex* correspond to three different distributions over words. The mixture of *unigrams* places each document at one of the corners of the topic *simplex*. LDA places a smooth distribution on the topic simplex denoted by the contour lines. Source: (Blei et al., 2003).

An estimate of an upper bound for the entropy of natural language, specifically English, is a combined view of the same metric, the aforementioned Perplexity. It is a statistical measure of how well a probability model predicts a sample. In NLP perplexity is a way of evaluating language models, can be measured at word per word level and estimating of optimal  $k$ -topics of a corpus, as applied to LDA. Language models addresses the problem of multiple introduction of unknown tokens, i.e. text generation, as Equation 3.8, observe the similarities with equations eqs. (3.2) and (3.7):

1. Generate a hidden string of *tokens* using a  $n$ -gram model.

2. Generate a *spelling* for each *token*.
3. Generate a case for each spelling.
4. Generate a spacing string to separate cased spellings from one another.

$$M_{token}(t_1 t_2 \cdots t_n) = M_{token}(t_1 t_2) \prod_{i=3}^n M_{token}(t_i | t_{i-2} t_{i-1}). \quad (3.8)$$

The conditional probabilities  $M_{token}(t_3 | t_1 t_2)$  are modeled as a weighted average of this four assumptions, where the weights  $\lambda_i$  satisfy  $\sum \lambda_i = 1$  and  $\lambda_i \geq 0$  (Brown et al., 1992; Shannon, 1951).

We consider written text as a stochastic process over an alphabet, including numbers and punctuation. Then, suppose  $\chi = \{\cdots X_{-1}, X_0, X_1 \cdots\}$  is a stationary stochastic process over a finite alphabet. Let  $P$  denote the probability distribution of  $\chi$  and let  $E_p$  denote expectations with respect to  $P$ . The entropy of  $\chi$  is defined by

$$H(\chi) \equiv H(P) \equiv -E_p \log P(X_0 | X_{-1}, X_{-2}, \cdots). \quad (3.9)$$

If the process is ergodic then the Shannon-McMillan-Breiman theorem (Shannon, 1948; McMillan et al., 1953; Breiman, 1957) states that almost surely

$$H(P) = \lim_{n \rightarrow \infty} -\frac{1}{n} \log P(X_1 X_2 \cdots X_n). \quad (3.10)$$

Under suitable regularity conditions, where  $M$  is a model for  $P$ , it can be shown that the cross-entropy of  $P$  as measured by  $M$  is defined by

$$H(P, M) = \lim_{n \rightarrow \infty} -\frac{1}{n} \log M(X_1 X_2 \cdots X_n), \quad (3.11)$$

where  $H(P) \leq H(P, M)$ . The difference between  $H(P, M)$  and  $H(P)$  is a measure of the inaccuracy of the model  $M$  (Brown et al., 1992).

By the perspective of text compression, the entropy and cross-entropy represents a fair approximation of topics extraction. For LDA, consider a series of approximations that successively take more and more statistics of the language into account and approach  $H$  as a limit, using logarithmic scales, as demonstrated in Zipft equation 3.4:

$$Perplexity_{LDA} = \exp \left\{ -\frac{\mathcal{L}(w)}{\text{count of tokens}} \right\} \quad (3.12)$$

$$\mathcal{L}(w) = \sum \log P(w_d | \Phi, \alpha), \quad (3.13)$$

$\mathcal{L}(w)$  is the log-likelihood of a set of unseen documents,  $w_d$  collection of unseen documents, the

model topic matrix  $\Phi$  and the hyperparameter  $\alpha$  for topic-distribution of documents.

Equations 3.12 and 3.14 represents how well the model represents the data set; the lower score is the best one. These equations will tell us which model provides the best results.

For language models, perplexity is the evaluation metric normalized by number of words  $T$ , represented by the inverse probability of corpus.

$$Perplexity_{LM} = \prod_{t=1}^T \left( \frac{1}{P_{LM}(x_{t+1}|x_t, \dots, x_1)} \right)^{1/T}. \quad (3.14)$$

Lower perplexity is a good result, but perplexity is not strongly correlated to human judgment (Chang et al., 2009), found that perplexity did not do a good job of conveying whether topics are coherent or not.

The primary purpose of this dissertation is to introduce a tool to assist all systematic review steps, supporting itself on text analysis/generation, bibliometrics, statistics, and artificial intelligence.

This chapter brings up related areas, meta-analysis, NLP, NLG, scientometrics, and bibliometrics. Bottlenecks pinpointed in the manual systematic review process are enlightened where automation could apply.



# The proposal

## 4.1 Proposal outline

Our main proposal is to automate as much as can be possible the traditional SLR process. Figure 4.1 depicts where we will hold the improvements. The traditional SLR process relates to 6 phases, where Meta-analysis is optional. The figure details the phase's flow. The proposed tool has the main goal to accelerate the selection, extraction, bibliometrics, and writing phases by clustering and generating text, supported by partial articles analyses, bibliometrics, and meta-analysis.

Recall that systematic reviews start by outlining the Review Plan, the filled document called Protocol has the inclusion and exclusion criteria, the query string, keywords, and objectives. The selection and extraction phases are directly dependent on inclusion and exclusion criteria, followed by a critical assessment of full studies. The selection phase selects studies by assessing titles and abstracts. If aligned with the review and inclusion criteria's objective, we consider the study for the extraction phase. The extraction phase excludes studies that cross the exclusion criteria. We perform the writing on top of the remainings studies.

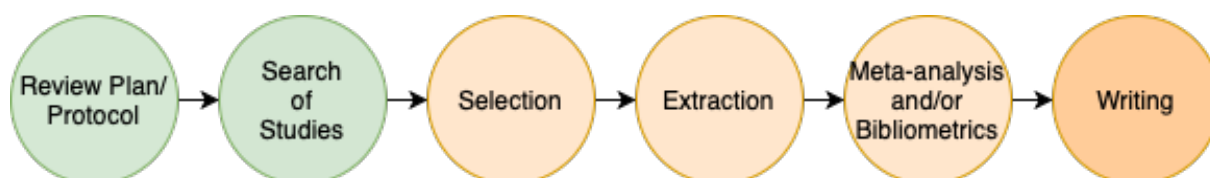


Figure 4.1: Systematic review phases selected for automation (orange).

This figure illustrates the general view where efforts will be applied, detailed in section 4.2, unfolding into two Assistants. The general strategy is explained in followings three steps:



1. The strategy for selection is to conduct manual selection on collected titles and abstracts, placed for revisor's manual assessment, as stated by the methodology.
2. On the Extraction phase, this is where our solution shows its potential, we group the studies in related topics, explained further below, section 4.2, and generate text to describe each group with a generated resume of included studies.
3. To support all these data, bibliometrics, statistics and graphs are added. Finally, a generated report in  $\text{\LaTeX}$  format will be available for detailed customization ready for dissemination.

In Figure 4.1, we provide a macro and simplified view of a systematic review. Green balloons have their own set of challenges, not covered in this work. Orange balloons are where most of the contributions were made, and the challenges related to each orange balloon are explained below.

**Selection:** the first step of a review, is to build a protocol, which inclusion/exclusion criteria are defined to narrow the findings. Those criteria are used by the authors to select which study will be excluded or included in the review, this is done by reading title and abstracts, and takes a lot of time to finish.

For us humans, the *Selection* is just deciding if the manuscript is adherent to an inclusion/exclusion criteria, the level of uncertainty is quite ambiguous to a computer, which is led by numbers. The challenge is how a computer decides if the manuscript should be accepted or rejected? We answer this question in 4.2 with a bidirectional encoder for question answering.

**Extraction:** after *Selection*, full text reading of selected studies is required. Extraction of quantitative data is held, and relations between studies are reported.

It is quite a challenge to extract quantitative data and uncover relations between studies. There are two main problems that arises in this phase, the lack of standard in data preparation for the existing tools and, the topics uncovered in documents that do not appear to be related. We developed an automatic (few or none intervention of a human) topics extraction.

**Meta-analysis:** is done to evaluate the quantitative data, sometimes, only bibliometric data is computed. Traditional meta-analysis is the extraction of quantitative data inside selected studies. It is used to verify if multiple scientific studies addressing the same question, with each individual study reporting measurements, if the degree of error expected is reasonable. We only cover the bibliometric analysis, see section 4.2.3

**Writing:** the automatic selection, clustering and topics extraction allows us to deliver an automatic report with reasonable results. Therefore, this step is the junction of all techniques presented in this proposal, thoroughly depicted in the next section, see Figure 4.2.

## 4.2 Proposed tool

For a better explanation of aimed goals, Figure 4.2 presents the new flow proposed as a means of automating the traditional systematic review process. A traditional review, after planning and feasibility discussion, always starts with a protocol, next, follows up with extract bibliography data from selected bases, from this step on, we propose two assistants for automation, underlined by number 1 and 2.

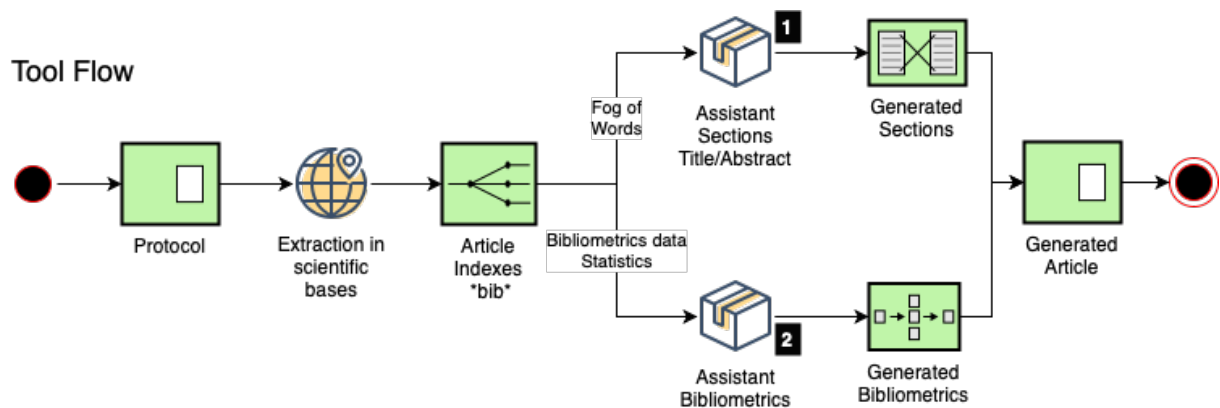


Figure 4.2: The flow of systematic literature review presented in the view of the proposed tool.

For a use case scenario, the data comes from a survey conducted, **Development Tools for IoT Applications: A Bibliometric Survey**. The selection process was manual, and we used the StArt tool in this process. This work aims to list all available research on the ISI Web of Science (WoS) about the IoT development solutions. The search string used is as follows. Other details of the SLR based protocol is available in the study.

*((“visual programming” OR “graphical programming” OR “generative programming” OR “code generator” OR “gamification” OR “component-based”) AND (“embedded” OR “iot” OR “cyber-physical” OR “sensor network” OR “smart building” OR “smart home” OR “smart city” OR “smart grid”))*

Web Of Science returned a total of 1760 records, available here <sup>1</sup>.

i Inclusion criteria

<sup>1</sup><https://www.webofscience.com/wos/woscc/summary/0b136258-eb19-4de2-8241-e6e2d4b8ecb0-046a14c6/relevance/1>

- Code generator
- Graphical component based for IoT

ii Exclusion criteria

- Systematic review
- Framework building
- Mobile games
- Simulator
- Non graphical component based

Following are detailed information about how these assistants work:

It is the assistant for NLP and text generation. Given a fog of words collected on the titles and abstracts of selected studies, we generate titles for group sections and the resume of each body of each group.

This assistant consists of two phases: clustering and text generation. The clustering consists of:

- Extraction of chunks from sentences
- Removal of Stopwords
- Custom removal of undesirable expressions
- Removal of expressions with only one word
- Stemming of expressions
- Extraction of clusters

Let the ideal *automation* be denoted by

$$O \xrightarrow{N} H^* \xrightarrow{S} A,$$

where  $O$  is the set of *corpus* extracted from all documents  $D$ .  $N$  is the vector of sentences,  $H^*$  is the bag of words built with title, abstracts, and keywords. After performing data cleaning, and feature engineering, the outcome is the dimensionality reduction  $S$ , which lead to generated sections  $A$ .

To represent the pipeline of data preparation ( $H^*$ ), chapter 3, equations 3.9 to 3.14, consider this example:  $P(\text{pencil} \mid \text{For dinner I'm making}) < P(\text{tapioca} \mid \text{For dinner I'm making})$ . Usually, the interest is in the probability that the model assigns to a full sentence  $W$  made of the sequence of words  $w_d$ .

$$P(W) = P(w_1, w_2, \dots, w_N) = P(w_1)P(w_2) \dots P(w_N).$$

This is particularly true for unigram models, which only works at the level of individual words. N-gram models, instead, looks at the previous  $w_{N-1}$  words to estimate the next one. Given that, the individual probabilities could be estimated based on the frequency of the words in the training corpus.

With a zero shot approach,  $H^*$  lead to sentences of particularly  $\hat{S}$  chunks across  $O$ .

$$O \xrightarrow{N} H^* \xrightarrow{\hat{S}} A.$$

The frequency of sentences gives the exact number that a given sentence occurs. When the sample is large, the frequency and probability distributions are similar in shape (Aaron and Spivey, 1998). What if we can reduce the sample dimensionality, considering sentences, instead of words?

$$P(w|\alpha, \beta) \text{ vs } Freq(\hat{S}) = k \in \mathbb{N}.$$

---

**Algorithm 4.1** Dimensionality reduction  $S \rightarrow Freq(\hat{S})$

---

```

1: for each  $\hat{S} \in D$  do
2:    $A \leftarrow D$ 
3:    $O \leftarrow \{O - D\}$ 
4:   if  $O$  is null then break
5:   end if
6: end for

```

---

Figure 4.3 depicts in detail all steps considered for the assistant 1. A bag of words built with title, abstracts, and keywords for each study (feature selection) and followed by data cleaning and data transformation of meaningful chunks of sentences. Next, feature engineering extracts the most frequent chunks, eliminating the less frequent in the process (dimensionality reduction). Data cleaning and data transform are all typical text processing steps. Usually, the steps are removing articles, punctuation, and standardizing whitespace.

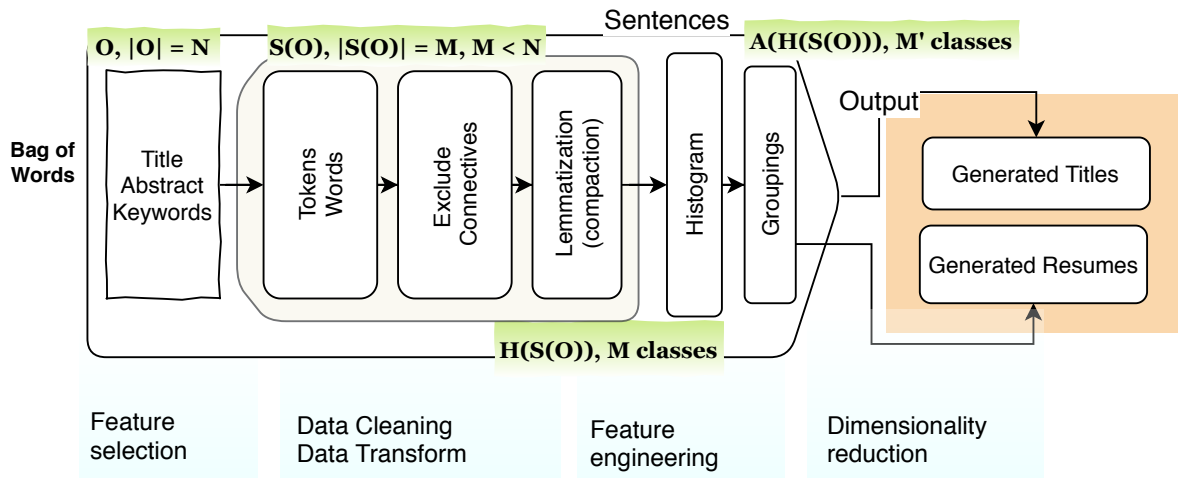


Figure 4.3: Pipeline of data preparation and our proposal language model.

#### 4.2.1 Question Answering for Assisted Selection

Systematic review selection by inclusion and exclusion criteria, is one of the most time consuming step. BERT (Devlin et al., 2018) which stands for Bidirectional Encoder Representations from Transformers, provides the Stanford Question Answering Dataset (SQuAD) Test F1 to 93.2.

While word embeddings are learnt from large corpora, their use in neural models to solve specific tasks is limited to the input layer. Recent advances in neural language models, have shown evidence that task specific architectures are not longer necessary and transferring some internal representations (attention blocks) along with shallow feed forward networks is enough, see Figure 3.2.

Let the selection with custom criteria be denoted by

$$O \xrightarrow{N} B \xrightarrow{\hat{N}} H^* \xrightarrow{\hat{S}} A,$$

where  $B$  represents the Bert tokenizer batch of sentences with tensor values.  $\hat{N}$  is the accepted vector of sentences, representing each document  $D$ . This step pad the batch to the length of the maximum sentence and truncate to the maximum length.

The assisted selection depends on a given question (criterion), a context( $O$ ) and returns an answer,  $\hat{N} \leftarrow \text{pair}(\text{criterion}, \text{context})$ . Pairwise each criterion with each document's bag of words, builds a vector with 12.320 entries/outputs.

For the aforementioned criteria, applied to our 1760 sample, BERT<sup>2</sup> result is available in table 4.1. The table has 3 columns, **Criteria** listing separated in two categories (inclusion, exclusion), **(%)Adherent** is the proportion of articles adherent to the category and **Doc.Count** the number of articles. The duplicates are just 186 articles, not representing a major role at this step.

<sup>2</sup>[https://huggingface.co/transformers/model\\_doc/bert.html](https://huggingface.co/transformers/model_doc/bert.html)

Criteria	(%)Adherent	Doc. Count
Inclusion	24,44	430
Code generator	–	140
Graphical component based for IoT	–	290
Exclusion	65,56	1144
Systematic review	–	120
Framework building	–	300
Mobile games	–	200
Simulator	–	400
Non graphical component based	–	124
Total	100%	1760 (-186)

Table 4.1: Assisted automation for systematic review’s selection - Available here

Figure 4.4 shows the feature in action in a beta environment for the end user. The **Status** column is the actual status of the article, the **Sug Selection** shows the resulting  $\hat{N}$  process after transformation.

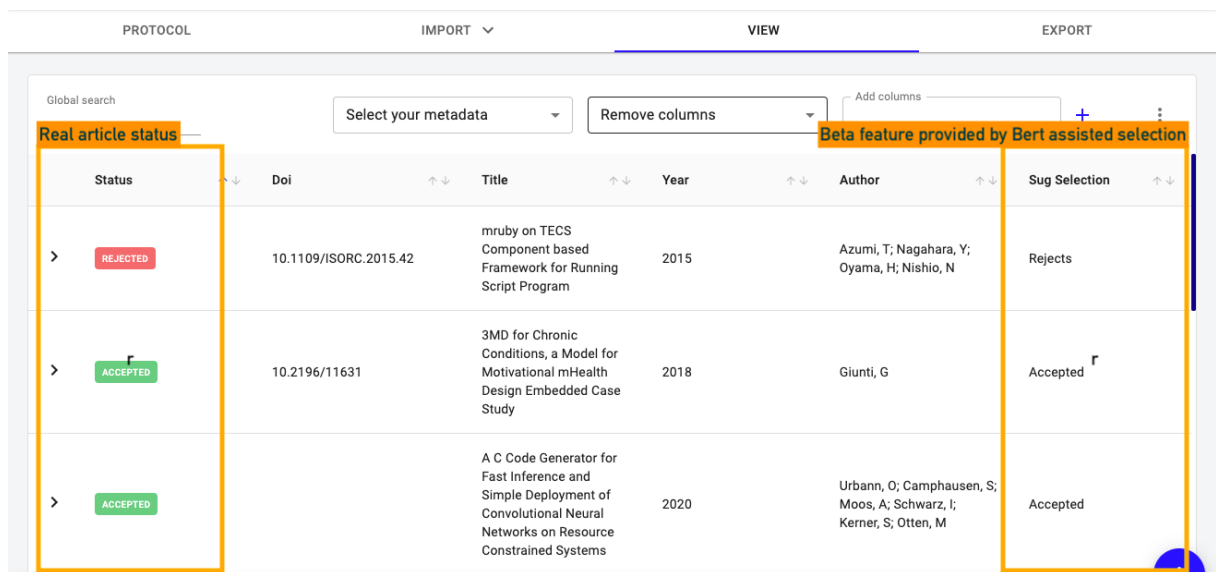


Figure 4.4: Assisted selection in action

Figure 4.5 shows pages of the generated document after all automation steps. For each session, a wordcloud is given with a listing of selected articles.



Fig. 16: Wordcloud of most frequent terms within section contiki operating system.

Table 15: Articles adherent to section Copyright ( C ) 2001 Ifac

Title	DOI*
A new compact programmable logic controller with integrated programming equipment Anomcompact-programmablelogiccontrollerwithintegratedprogrammingequipment	

Table 16: Articles adherent to section Developed Tool

Title	DOI*
Low cost implementation of mathematical functions using piecewise interpolation Locationpiecewise-interpolationofmathematicalfunctionsandpiecewise-interpolation	



Fig. 48: Wordcloud of most frequent terms within section present approach.

Table 46: Articles adherent to section Present Approach

Title	DOI*
Fostering Reuse in Scientific Computing With Embedded Components Application in High Performance Bayesian Inference for Bioinformatics FosteringReuseinScientificComputingWithEmbeddedComponentsApplicationinHighPerformanceBayesianInferenceforBioinformatics	10.1109/MCSE.2018.2883688
Adapting computational independent modules for derivation of architectural requirements of software product lines Adaptingcomputationalindependentmodulesfor derivationof architectural requirements of software product lines	10.1109/MOMPES.2007.2

\* Clickable item



Fig. 17: Wordcloud of most frequent terms within section copyright ( c ) 2001 Ifac.

Table 17: Articles adherent to section Earlier Work

Title	DOI*
Interactive scheduling for clustered VLSI DSPs in stream-based architectures VLSI DSPs in stream-based architectures	10.1109/PACT.2000.888203
Formal ModelBog and Verification of IEEE1489 Function Blocks with Abstract State Machines and SMV Execution Semantics FormalModelBogandVerificationofIEEE1489FunctionBlockswithAbstractStateMachinesandSMVExecutionSemantics	10.1002/9781118283942.ch10

\* Clickable item



Fig. 49: Wordcloud of most frequent terms within section present model.

Table 47: Articles adherent to section Present Model

Title	DOI*
An engineering approach to determining sampling rates for switches and sensors in real time systems Anengineeringapproachtodeterminingsampling rates for switches and sensors in real time systems	10.1109/RTAS.2009.852448

\* Clickable item

Table 48: Articles adherent to section Previous Research

Title	DOI*
XMM A high performance automatic memory management system with memory constrained design XMM A high performance automatic memory management system with memory constrained design	

\* Clickable item

Figure 4.5: Pages of the generated document after automation

### 4.2.2 Topics Extraction - LDA vs Our Proposal

To generate the results presented in the next sections were used Markov chain techniques for text generation jointly with Spacy<sup>3</sup> framework, and compared with the language model OpenAI's GPT-2 with TensorFlow. Spacy is a framework for NLP that uses named entities on neural network pre-trained models, created for industrial applications.

GPT-2 Radford et al. (2019) is an encode/decode model for NLP problems with zero shot training. It uses a heat parameter defined as

$$P_t(w) = \frac{\exp(s_w/\tau)}{\sum_{w'} s'_w/\tau} \tag{4.1}$$

to control the diversity.

We use the two existing summarization techniques: i) Abstractive Summarization and ii) Extractive Summarization.

**Extractive Summarization** selects chunks of original text to create a resume.

**Abstractive Summarization** generates new text using natural language generation techniques.

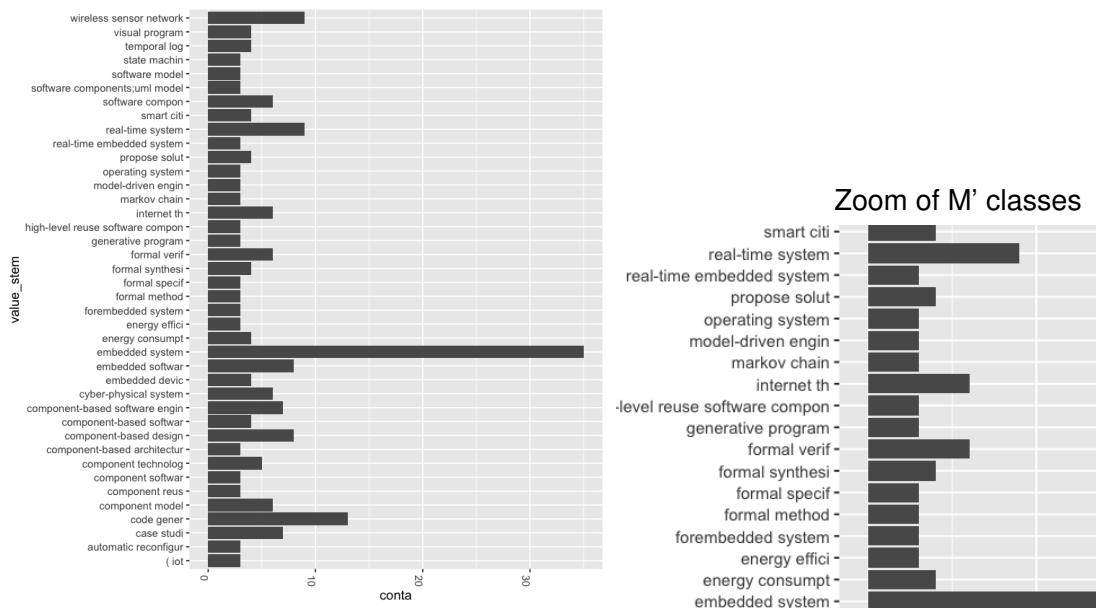


Figure 4.6: M' classes from resulting trimmed clusters. Sentences generated by Spacy's named entities.

<sup>3</sup><https://spacy.io>



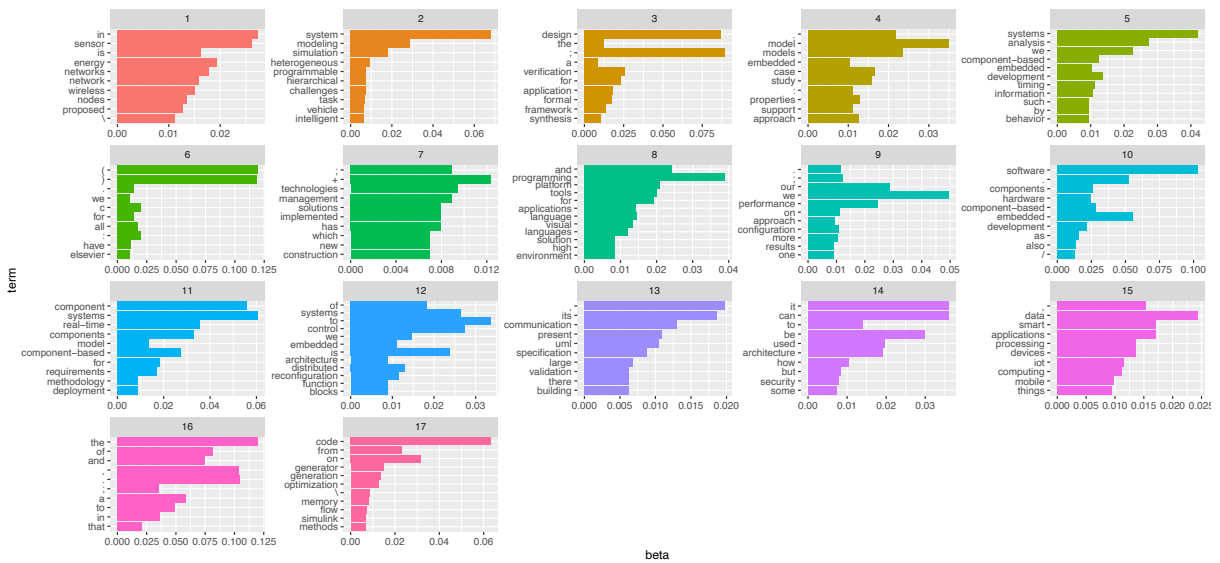


Figure 4.7: LDA topic modeling for each k' topic suggested for the fog data, see Figure 4.8

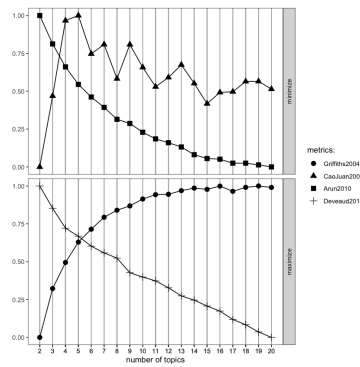


Figure 4.8: Best k-topics suggested using ldatuning. Two maximization metrics and two minimization functions.

For data cleaning steps, we set a small set of conditions that needed to be satisfied: i) All extracted sentences must be different; ii) By sampling, we elect four random sentences to represent each group; iii) We discard the set of sentences with length zero. The Figure 4.6 shows the sentences generated after clustering using the model *en-core* from Spacy, we removed the stopwords and extracted the sentences using named entities.

For categorization of similar studies, Figures 4.7 and 4.8 details the two methodologies used for comparison purposes. Figure 4.6 shows our proposal for categorization, thoroughly explained in section ???. Figure 4.8 shows LDA topic modeling on the same data set, figure 4.8 depicts best *k*-topics that LDA estimates for this data. The vectorization and extraction of relevant studies for each technique are very similar and generally discussed in section ???.

$$P(w|\alpha, \beta) \text{ vs } Freq(\hat{S}) = k \in \mathbb{N}$$

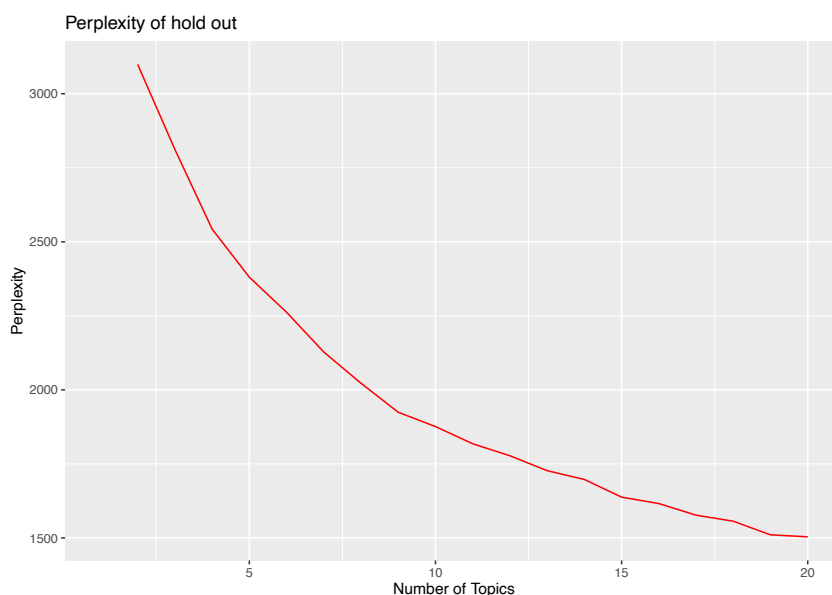


Figure 4.9: LDA Perplexity tests - Steps to minimize perplexity by maximizing probability

Observe that LDA must calculate the best  $k$  by sampling across all data, our approach, instead, use the frequency of sentences mapped across all data, making it faster and unsupervised. Figure 4.9 shows each step taken of correctly guessing the smallest value of the distribution to represent the data. Which lead to

$$Perplexity_{LDA} = 0.0006424961 \text{ for this dataset, where } k = 17.$$

Results 1-4 depicted in figs. 4.10 to 4.13 show the generated text for the same group using different techniques. The leverage of GPT2 is the structured generated text. Observe Result 1, essentially, Markov chain does an extractive summarization and generates sentences with existing chunks of the fog. Result 2 is Spacy's named entities combined with Markov chain, its better than result 1, it uses extractive chunks, but it seems more coherent for us humans. We can train an RNN-LM on any text, then generate text in that style. Result 3 is GPT2-Tensorflow trained in the corpus of the selected group, observe the repetition, it seems like the model chooses an antagonist theme and lecture against it, and it seems surprisingly reasonable. It takes time to transfer the knowledge to the model, about 15 minutes of training in this set, *et voilà*. Result 4 is the result of the 774M GPT2-Tensorflow zero-shot language model, interestingly is the more repetitive result, it tends to focus on an idea deeply, but could be the wrong idea. Thoroughly discussion on perplexity should be presented for comparison reasons, and other factors like human perception still need appraisal too.

[1] "The disadvantage of this narrow targeting is that the proposed solution exploits an environment software platform with a high-level graphical programming language is proposed to replace part of the component in response."

[1] "Once verified, the conditions allow concluding the correctness of the method through discussion of the framework proposes revolutionary architecture solutions in contrast to human architects, who tend to propose solutions that allow to reduce the periods of drives inoperability."

[1] "BALBOA: A component-based design environment for system models This paper presents Unison as a supplement for these students."

[1] "Methods and tools are needed to assist the architect during system design. The goal of this paper applies modeling techniques to embedded software development faces additional challenges."

Figure 4.10: **Result 1** - Generated text for the group "embedded system", Markov chain method only.

[1] "Nevertheless, the efficiency of JIT compilation phases by reducing their occurrences, this paper presents the BALBOA component composition framework for system-level systems software Component-based design concepts in mind."

[1] "Differently than standard approaches for FPGA programming, which hinders its reusability, analyzability and maintainability."

[1] "Correct concurrent codes achieving maximum performance on the model and advanced solving techniques using constraint programming."

[1] "At the same time it demonstrates that it has very low overhead and is suitable for implementing both system-level systems software."

[1] "The tool support relies on an automotive sub-system. Due to the everlasting consumer demand for new innovative user functions."

[1] "Although the case study shows savings in manual effort to the problem."

Figure 4.11: **Result 2** - Generated text for the group "embedded system", Spacy's Named Entities combined with Markov chain.

[1] Embedded System

A Generic Component-Based Approach for Programming, Composing and Tuning Sensor Software Wireless sensor networks (WSNs) are being extensively deployed today in various monitoring and control applications by enabling rapid deployments at low cost and with high flexibility. However, high-level software development is still one of the major challenges to wide-spread WSN adoption. The success of high-level programming approaches in WSNs is heavily dependent on factors such as ease of programming, code well-structuring, degree of code reusability, required software development effort and the ability to tune the sensor software for a particular application. Component-based programming has been recognized as an effective approach to satisfy such requirements. However, most of the componentization efforts in WSNs were ineffective due to various reasons, such as high resource demand or limited scope of use. In this article, we present Remora, a novel component-based approach to overcome the hurdles of WSN software implementation and configuration. Remora offers a well-structured programming paradigm that fits very well with resource limitations of embedded systems, including WSNs. Furthermore, the special attention to event handling in Remora makes our proposal more practical for embedded

[2] Embedded System

The design of a complex embedded control system involves integration of a large number of components. These components need to interact in a timely fashion to achieve the system level end-to-end requirements. In practice, the component level timing specification consists of design attributes like component task mapping, task period and scheduled definition but often lack details on their real-time (functional) requirements. As we observe, there is no systematic methodology in place for decomposing the feature level timing requirements into component level timing

Figure 4.12: **Result 3** - Generated text for the group "embedded system", model GPT2-Tensorflow trained in the corpus of selected group with Attention.

```

gpt2.generate(sess,
               model_name=model_name,
               prefix="Embedded System",
               length=100,
               temperature=0.8,
               top_p=0.9,
               nsamples=1,
               batch_size=1
              )

```

[1] Embedded System Components and the Kernel

The embedded systems are generally built with a kernel module that provides the access to the system hardware. A kernel module is a program that is loaded when the system boots, or when it starts up. The kernel module is the only component of a computer that is made available to the user. The kernel module may be either in the form of an executable program or a library that is loaded from disk. The kernel module is usually loaded into the system in the same place.

Figure 4.13: **Result 4** - Generated text for the group "embedded system", language model 774M GPT2-Tensorflow Zero Shot.

### 4.2.3 Assistant 2

It is the assistant for statistics, bibliometrics, and graphs generation. It gets bibliography data as input and outputs graphics and statistics. Figure 4.14 shows some self explanatory graphs extracted from data. In particular, observe figure 4.14(a) and table 4.2, Lotka ([Lotka, 1926](#)) function estimates the Beta coefficient of our bibliographic collection and assess, through a statistical test, the similarity of this empirical distribution with the theoretical one. That is for every 100 scientists who produce one paper there are approximately  $100/2^2$ , or 25, who produce two papers,  $100/3^2$ , or 11. We do not reject this behavior in our data,  $P - value > 0.05$ , non-significant result. Maybe it is because it is a 'hot' area, and authors publish more often with less impact, perhaps.

N.Articles	N.Authors	Freq
1	691	0.957063712
2	26	0.036011080
3	4	0.005540166
5	1	0.001385042
$\beta = 4.133$		$P_{value} = 0.699$

Table 4.2: Lotka distribution

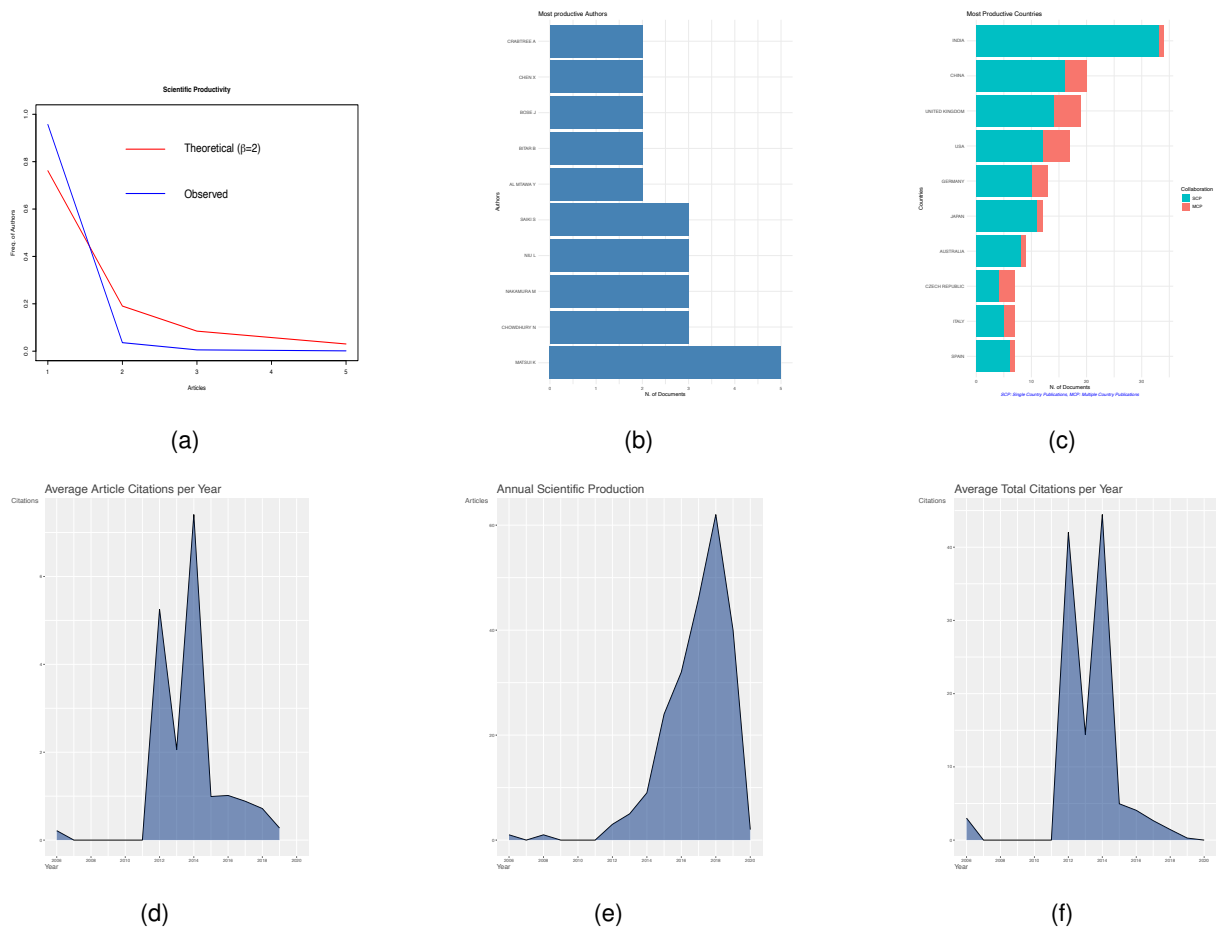
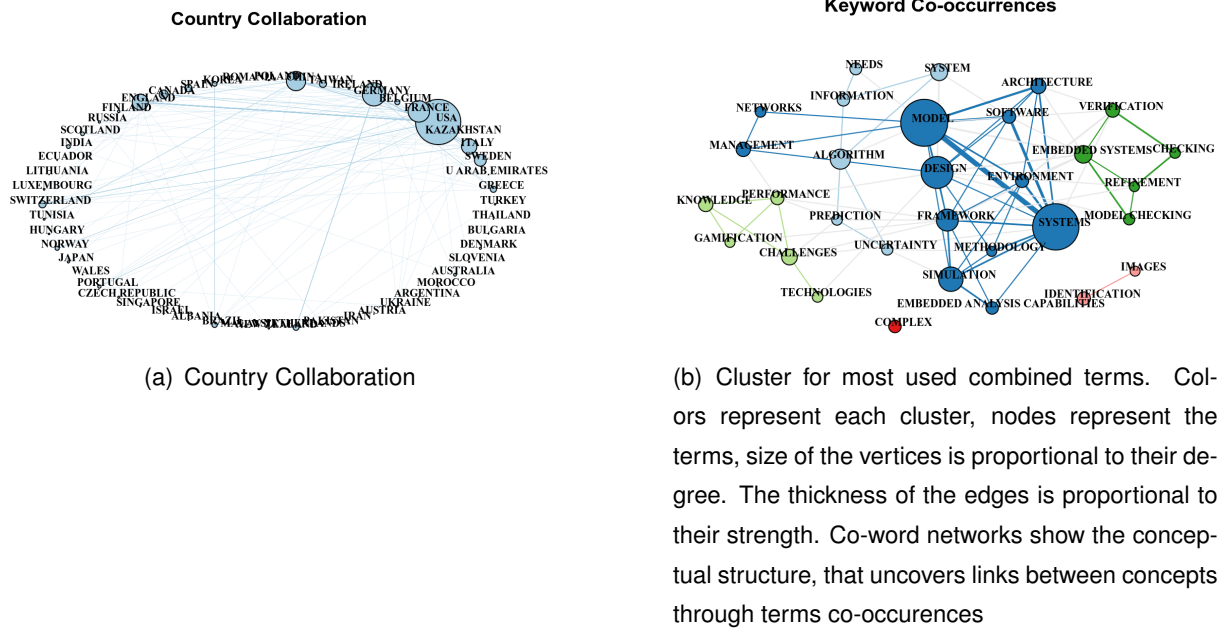


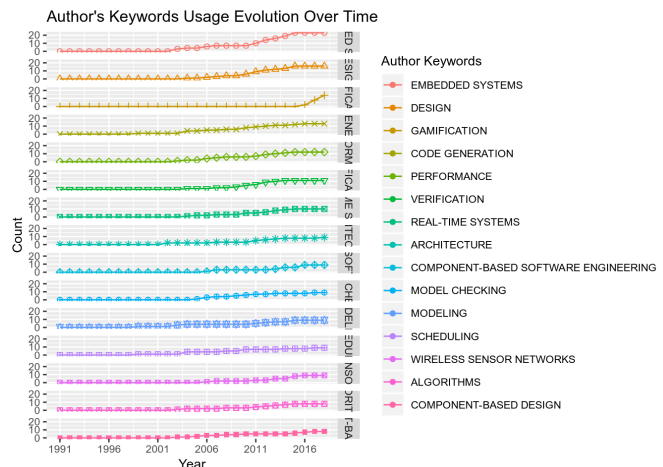
Figure 4.14: Bibliometrics extracted from studies.

A keen look at the data enlightens insights into this particular research field. Observe the country collaboration. Edges are thicker to/from the USA, Germany, and France. Meaning that most advances are coming from the collaboration of authors residing in these countries, Figure 4.15(a). Figure 4.15(b) shows n-gram of most used keywords in *bibtex* file, which is how often a term appears next to others. The evolution of a language comes with the birth of new terms, extinction of others, and change of meaning. Figure 4.15(c) the usage evolution of most popular author keywords.



DOCUMENT CONTENTS	
Keywords Plus (ID)	189
Author's Keywords (DE)	792
AUTHORS	
Authors	722
Author Appearances	760
Authors of single-authored documents	13
Authors of multi-authored documents	709
AUTHORS COLLABORATION	
Single-authored documents	16
Documents per Author	0.312
Authors per Document	3.21
Co-Authors per Documents	3.38
Collaboration Index	3.39

Table 4.3: Important information about data



(c) Author's keywords usage evolution over time. Observe that our data set shows that a lot of them was first used midst 2004. Those are the most used keywords extracted from *bib* file.

Figure 4.15: Bibliometrics extracted from studies

The outputs of assistants 1 and 2 are combined to generate a  $\text{\LaTeX}$  document, ready for revisors inputs.

This chapter presents our proposal for systematic review automation. Techniques and planning are widely discussed.

# 5

## Final Considerations

This work addresses the systematic review methodology by the point of view of computational challenges. The work presented focuses on the human task conducted by a specialist assisted by computational techniques, namely, text mining, text generation, bibliometry, and nested graphics.

This solution is less costly computationally and provides substantial gains in terms of quality to the specialists conducting the review.

We showed that AI research development could aid in better automation of manual practices. NLP advances to our daily living because it is a technology maturing at a fast pace. Real applications of text generation are becoming more viable and acceptable to human perceptions. The selection, extraction, and writing of systematic reviews benefit significantly in this work.

Be aware that most of the tools we encountered were written by academic groups involved in research into evidence synthesis and machine learning. Very often, these groups have prototyped a software to demonstrate a method. However, such prototypes do not perform well: we commonly encountered broken web links, challenging to understand and slow user interfaces, and server errors. Moving from the research prototypes to professionally maintained platforms remains a significant problem to overcome.

Eventually, the notion of a review becoming almost immediately out of date at the time of publication will disappear as autonomous agents sift the evidence continuously and use their protocols to provide updated reviews on demand. In such a way that practitioners will have access to the best evidence at a reasonable time.

Still, there is much work to do to improve the presented work. We aspire to continue making further progress: improve the usability on top of the pipeline, better text generation techniques, addition of more bibliometrics indicators, expand to a more extensive health care standards, full-text analysis, and automatic meta-analysis, just to mention a few essential improvements. Nevertheless, we understand that reproducibility is a big challenge in NLG, the lack of metrics and well-established standards to measure the quality of the generated text is an open challenge

in literature, for now, we lean towards human validation. This work is available at repository <https://github.com/SensorNet-UFAL/rnatlp.git>.



# Bibliography

Eric Aaron and Michael Spivey. Frequency vs. probability formats: Framing the three doors problem. In *Proceedings of the twentieth annual conference of the Cognitive Science Society*, pages 13–18, 1998.

Isidro F Aguillo. Is google scholar useful for bibliometrics? a webometric analysis. *Scientometrics*, 91(2):343–351, 2012.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Yoshua Bengio. Probabilistic neural network models for sequential data. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 5, pages 79–84. IEEE, 2000.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012. ISSN 0001-0782. DOI [10.1145/2133806.2133826](https://doi.org/10.1145/2133806.2133826). URL <http://doi.acm.org/10.1145/2133806.2133826>.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

David M Blei, John D Lafferty, et al. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

- Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545, 2017.
- Maura Borrego, Margaret J Foster, and Jeffrey E Froyd. Systematic literature reviews in engineering education and other developing interdisciplinary fields. *Journal of Engineering Education*, 103(1):45–76, 2014.
- David Bowes, Tracy Hall, and Sarah Beecham. Slurp: a tool to help large complex systematic literature reviews deliver valid and rigorous results. In *Proceedings of the 2nd international workshop on Evidential assessment of software technologies*, pages 33–36, 2012.
- Leo Breiman. The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.
- Pearl Brereton, Barbara A Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software*, 80(4):571–583, 2007.
- Robert Broadus. Toward a definition of “bibliometrics”. *Scientometrics*, 12(5-6):373–379, 1987.
- Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, 1992.
- Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.
- Jason Brownlee. Deep learning for natural language processing. *Machine Learning Mystery, Vermont, Australia*, 322, 2017.
- Quentin Burrel. Stochastic modelling of the first-citation distribution. *Scientometrics*, 52(1): 3–12, 2001.
- Michel Callon, John Law, and Arie Rip. Qualitative scientometrics. In *Mapping the dynamics of science and technology*, pages 103–123. Springer, 1986.
- Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.
- Chaomei Chen, Katherine McCain, Howard White, and Xia Lin. Mapping scientometrics (1981–2001). *Proceedings of the American Society for Information Science and Technology*, 39(1):25–34, 2002.

- Lifeng Chen and Carol Friedman. Extracting phenotypic information from the literature via natural language processing. In *Medinfo*, pages 758–762. Citeseer, 2004.
- Noam Chomsky. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124, 1956.
- Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, 2016.
- Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219, 2006.
- Rodrigo Costas and María Bordons. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of informetrics*, 1(3):193–203, 2007.
- Rachel Couban. Covidence and rayyan. *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada*, 37(3), 2016.
- Jeanne Daly, Karen Willis, Rhonda Small, Julie Green, Nicky Welch, Michelle Kealy, and Emma Hughes. A hierarchy of evidence for assessing qualitative health research. *Journal of clinical epidemiology*, 60(1):43–49, 2007.
- David Denyer and David Tranfield. Producing a systematic review. In In D. A. Buchanan & A. Bryman (Eds.), editor, *The Sage handbook of organizational research methods*, page 671–689. Sage Publications Ltd., 2009.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Valérie Durieux and Pierre Alain Gevenois. Bibliometric indicators: quality measurements of scientific publication. *Radiology*, 255(2):342–351, 2010.
- Sandra Fabbri, Cleiton Silva, Elis Hernandez, Fábio Octaviano, André Di Thommazo, and Anderson Belgamo. Improvements in the start tool to better support the systematic review process. In *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*, pages 1–5, 2016.
- Ana M Fernández-Sáez, Marcela Genero Bocco, and Francisco P Romero. Slr-tool: A tool for performing systematic literature reviews. In *ICSOFT (2)*, pages 157–166, 2010.

- Eugene Garfield. Reviewing review literature. part 2. the place of reviews in the scientific literature. *Essays of an Information Scientist*, 10(19):117, may 1987.
- Gerald Gazdar, Ewan Klein, Geoffrey K Pullum, and Ivan A Sag. *Generalized phrase structure grammar*. Harvard University Press, 1985.
- Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, pages 473–483, 1992.
- Gene V Glass and Mary Lee Smith. Meta-analysis of research on class size and achievement. *Educational evaluation and policy analysis*, 1(1):2–16, 1979.
- PP Glasziou, L Irwig, CJ Bain, and GA Colditz. How to use the evidence: assessment and application of scientific evidence. In *School of Medicine Publications*. Canberra: National Health & Medical Research Council, 2000.
- Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.  
<http://www.deeplearningbook.org>.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, 2014.
- Trisha Greenhalgh. How to read a paper: Papers that summarise other papers (systematic reviews and meta-analyses). *Bmj*, 315(7109):672–675, 1997.
- Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.
- Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650):924–926, 2008. ISSN 0959-8138. DOI 10.1136/bmj.39489.470347.AD. URL <https://www.bmj.com/content/336/7650/924>.
- Anna-Bettina Haidich. Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29, 2010.
- Mary Dee Harris. Introduction to natural language processing. Technical report, Loyola Univ., 1984.
- Julian PT Higgins, Sally Green, et al. *Cochrane handbook for systematic reviews of interventions*. Wiley Online Library, 2008.

- William W. Hood and Concepción S. Wilson. The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2):291, Oct 2001. ISSN 1588-2861.  
**DOI** [10.1023/A:1017919924342](https://doi.org/10.1023/A:1017919924342).
- Ozan Irsoy and Claire Cardie. Deep recursive neural networks for compositionality in language. In *Advances in neural information processing systems*, pages 2096–2104, 2014.
- Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.
- Khalid S Khan, Gerben Ter Riet, Julie Glanville, Amanda J Sowden, Jos Kleijnen, et al. *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews*. Number 4 (2n in CRD Report. NHS Centre for Reviews and Dissemination, 2001.
- Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.
- Barbara Kitchenham, O Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. Systematic literature reviews in software engineering—a systematic literature review. *Information and software technology*, 51(1):7–15, 2009.
- Christian Kohl, Emma J McIntosh, Stefan Unger, Neal R Haddaway, Steffen Kecke, Joachim Schiemann, and Ralf Wilhelm. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on cadima and review of existing tools. *Environmental Evidence*, 7(1):8, 2018.
- András Kornai. *Mathematical linguistics*. Springer Science & Business Media, 2007.
- Sascha Kraus, Matthias Breier, and Sonia Dasí-Rodríguez. The art of crafting a systematic literature review in entrepreneurship research. *International Entrepreneurship and Management Journal*, pages 1–20, 2020.
- Steve Lawrence, C Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. *Computer*, 32(6):67–71, 1999.
- Loet Leydesdorff. Indicators of structural change in the dynamics of science: Entropy statistics of the sci journal citation reports. *Scientometrics*, 53(1):131–159, 2002.
- Yingming Li, Ming Yang, and Zhongfei (Mark) Zhang. Scientific articles recommendation. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, CIKM '13*, pages 1147–1156, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8. **DOI** [10.1145/2505515.2505705](https://doi.org/10.1145/2505515.2505705). URL <http://doi.acm.org/10.1145/2505515.2505705>.

- Chin-Yew Lin and FJ Och. Looking for a few good metrics: Rouge and its evaluation. In *Ntcir Workshop*, 2004.
- Alfred J Lotka. The frequency distribution of scientific productivity. *Journal of the Washington academy of sciences*, 16(12):317–323, 1926.
- Richard Mallett, Jessica Hagen-Zanker, Rachel Slater, and Maren Duvendack. The benefits and challenges of using systematic reviews in international development research. *Journal of development effectiveness*, 4(3):445–455, 2012.
- Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.
- Christopher Marshall and Pearl Brereton. Systematic review toolbox: a catalogue of tools to support systematic reviews. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, pages 1–6, 2015.
- Iain J Marshall and Byron C Wallace. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8(1):163, 2019.
- André Martinet and Elisabeth Palmer. *Elements of general linguistics*. Faber & Faber, 1966.
- Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O’Blenis. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17(4):446–453, 2010.
- Brockway McMillan et al. The basic theorems of information theory. *The Annals of Mathematical Statistics*, 24(2):196–219, 1953.
- John Mingers and Loet Leydesdorff. A review of theory and practice in scientometrics. *European Journal of Operational Research*, 246(1):1 – 19, 2015. ISSN 0377-2217. DOI <https://doi.org/10.1016/j.ejor.2015.04.002>.
- Farman A Moayed, Nancy Daraiseh, Richard Shell, and Sam Salem. Workplace bullying: a systematic review of risk factors and outcomes. *Theoretical Issues in Ergonomics Science*, 7(3):311–327, 2006.
- David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLOS Medicine*, 6(7):1–6, 07 2009. DOI [10.1371/journal.pmed.1000097](https://doi.org/10.1371/journal.pmed.1000097).
- Jefferson Seide Molléri and Fabiane Barreto Vavassori Benitti. Sesra: a web-based automated tool to support the systematic literature review process. In *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering*, pages 1–6, 2015.

- PP Morgan. Review articles: 2. the literature jungle. *CMAJ: Canadian Medical Association Journal*, 134(2):98, 1986.
- Rogério Mugnaini, Asa Fujino, and Nair Yumiko Kobashi. Bibliometria e cientometria no brasil: infraestrutura para avaliação da pesquisa científica na era do big data. *São Paulo: ECA/USP*, 2017.
- Cynthia D Mulrow. The medical review article: state of the science. *Annals of internal medicine*, 106(3):485–488, 1987.
- Zachary Munn, Edoardo Aromataris, Catalin Tufanaru, Cindy Stern, Kylie Porritt, James Farrow, Craig Lockwood, Matthew Stephenson, Sandeep Moola, Lucylynn Lizarondo, Alexandra McArthur, Micah Peters, Alan Pearson, and Zoe Jordan. The development of software to support multiple systematic review types: the Joanna Briggs Institute System for the Unified Management, Assessment and Review of Information (JBI SUMARI). *INTERNATIONAL JOURNAL OF EVIDENCE-BASED HEALTHCARE*, 17(1):36–43, MAR 2019.  
**DOI** [10.1097/XEB.0000000000000152](https://doi.org/10.1097/XEB.0000000000000152).
- Vasilli Vasilevich Nalimov and Zinaida Maksimovna Mulchenko. Measurement of science. study of the development of science as an information process. Technical report, FOREIGN TECHNOLOGY DIV WRIGHT-PATTERSON AFB OHIO, 1971.
- Nicholas Israel Nii-Trebi. Emerging and neglected infectious diseases: insights, advances, and challenges. *BioMed research international*, 2017, 2017.
- Enrique Orduna-Malea, Alberto Martín-Martín, Emilio Delgado López-Cózar, et al. Google scholar as a source for scholarly evaluation: a bibliographic review of database errors. *Revista española de Documentación Científica*, 40(4), 2017.
- Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, 5(1):210, 2016.
- Annette M O'Connor, Guy Tsafnat, Stephen B Gilbert, Kristina A Thayer, and Mary S Wolfe. Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of international collaboration for the automation of systematic reviews (icasr). *Systematic reviews*, 7(1):3, 2018.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- By Tina Poklepović Peričić and Sarah Tanveer. Why systematic reviews matter. internet, jul 2019. available at:  
<https://www.elsevier.com/connect/authors-update/why-systematic-reviews-matter>.

- Olle Persson, Wolfgang Glänzel, and Rickard Danell. Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60(3):421–432, 2004.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
- James M Rippe and Theodore J Angelopoulos. Relationship between added sugars consumption and chronic disease risk factors: Current understanding. *Nutrients*, 8(11):697, 2016.
- Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.*, 28(1): 4:1–4:38, January 2010. ISSN 1046-8188. DOI 10.1145/1658377.1658381. URL <http://doi.acm.org/10.1145/1658377.1658381>.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- Harrison Scells and Guido Zuccon. Searchrefiner: A query visualisation and understanding tool for systematic reviews. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, page 1939–1942, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. DOI 10.1145/3269206.3269215. URL <https://doi.org/10.1145/3269206.3269215>.
- Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.
- Kirsty R Short, Katherine Kedzierska, and Carolien E van de Sandt. Back to the future: lessons learned from the 1918 influenza pandemic. *Frontiers in cellular and infection microbiology*, 8: 343, 2018.
- Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011.
- Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1017–1024, 2011.



- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- James Thomas, Jeff Brunton, and Sergio Graziosi. Eppi-reviewer 4.0: software for research synthesis. *London: Social Science Research Unit, UCL Institute of Education*, 2010.
- James Thomas, John McNaught, and Sophia Ananiadou. Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1):1–14, 2011.
- James Thomas, Anna Noel-Storr, Iain Marshall, Byron Wallace, Steven McDonald, Chris Mavergames, Paul Glasziou, Ian Shemilt, Anneliese Synnot, Tari Turner, et al. Living systematic reviews: 2. combining human and machine effort. *Journal of clinical epidemiology*, 91:31–37, 2017.
- Geng Tian and Liping Jing. Recommending scientific articles using bi-relational graph-based iterative rwr. In *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13*, pages 399–402, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2409-0. DOI 10.1145/2507157.2507212. URL <http://doi.acm.org/10.1145/2507157.2507212>.
- Mercedes Torres Torres and Clive E Adams. Revmanhal: towards automatic text generation in systematic reviews. *Systematic reviews*, 6(1):27, 2017.
- David Tranfield, David Denyer, and Palminder Smart. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, 14(3):207–222, 2003.
- Guy Tsafnat, Adam Dunn, Paul Glasziou, and Enrico Coiera. The automation of systematic reviews, 2013.
- Anthony FJ Van Raan. Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1):133–143, 2005.
- Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11*, pages 448–456, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. DOI 10.1145/2020408.2020480. URL <http://doi.acm.org/10.1145/2020408.2020480>.
- Terry Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC, 1971.
- William A Woods. Transition network grammars for natural language analysis. *Communications of the ACM*, 13(10):591–606, 1970.

---

Richard Wormald and Jennifer Evans. What makes systematic reviews systematic and why are they the highest level of evidence? *Ophthalmic Epidemiology*, 25(1):27–30, 2018.

**DOI** [10.1080/09286586.2017.1337913](https://doi.org/10.1080/09286586.2017.1337913). PMID: 28891724.

G. K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, 1932.