



UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Tecnologia

MARBILIA POSSAGNOLO SERGIO

PROCESSO DE REVISÃO LITERÁRIA USANDO MODELO DE TOPICOS EM
FERRAMENTA INTERATIVA

LITERARY REVIEW PROCESS USING TOPICS MODEL IN INTERACTIVE TOOL

Limeira

2023

MARBILIA POSSAGNOLO SERGIO

PROCESSO DE REVISÃO LITERÁRIA USANDO MODELO DE TOPICOS EM
FERRAMENTA INTERATIVA

LITERARY REVIEW PROCESS USING TOPICS MODEL IN INTERACTIVE TOO

Tese apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para obtenção do título de Doutora em Tecnologia, na área de Sistemas de Informação e Comunicação.

Thesis presented to the Faculty of Technology of the State University of Campinas as part of the requirements for obtaining the title of Doctor of Technology, in the area of Information and Communication Systems.

Supervisor/Orientador: Prof. Paulo Sérgio Martins Pedro

Co-supervisor/Coorientador: Dr. Marcelo Schneck de Paula Pessôa

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DA TESE
DEFENDIDA PELA ALUNA MARBILIA POSSAGNOLO SERGIO. E
ORIENTADA PELO PROF. DR PAULO SÉRGIO MARTINS PEDRO

Limeira
2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Tecnologia
Felipe de Souza Bueno - CRB 8/8577

Se66p Sergio, Marbilia Possagnolo, 1962-
Processo de revisão literária usando modelo de tópicos em ferramenta interativa / Marbilia Possagnolo Sergio. – Limeira, SP : [s.n.], 2023.

Orientador: Paulo Sérgio Martins Pedro.

Coorientador: Marcelo Schneck de Paula Pessoa.

Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Tecnologia.

1. Processamento de linguagem natural (Computação). 2. Revisão sistemática. 3. Inteligência artificial. 4. Software de aplicação - Desenvolvimento. I. Martins Pedro, Paulo Sérgio, 1967-. II. Pessoa, Marcelo Schneck de Paula, 1949-. III. Universidade Estadual de Campinas. Faculdade de Tecnologia. IV. Título.

Informações Complementares

Título em outro idioma: Literary review process using topics model in interactive tools

Palavras-chave em inglês:

Natural language processing (Computer science)

Systematic review

Artificial intelligence

Application software - Development

Área de concentração: Sistemas de Informação e Comunicação

Titulação: Doutora em Tecnologia

Banca examinadora:

Marcelo Schneck de Paula Pessoa

Edson Luiz Ursini

Marli de Freitas Gomes Hernandez

Artemis Maria Francelin Sanches Moroni

Mauro de Mesquita Spínola

Data de defesa: 31-05-2023

Programa de Pós-Graduação: Tecnologia

Identificação e informações acadêmicas do(a) aluno(a)

- ORCID do autor: <https://orcid.org/0000-0001-7575-9976>

- Currículo Lattes do autor: <https://lattes.cnpq.br/3950781480912294>

UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Tecnologia

**Processo de revisão literária usando modelo de tópicos em ferramenta
interativa**

Autor: Marbília Possagnolo Sergio

Orientador: Prof. Dr. Prof. Paulo Sérgio Martins Pedro

Coorientador Dr. Marcelo Schneck de Paula Pessôa

A Banca Examinadora composta pelos membros abaixo aprovou esta Tese:

Prof. Dr. Marcelo Schneck, de Paula Pessôa

Instituição: Politécnica – USP São Paulo

Prof. Dr. Edson Luiz Ursini

Instituição: Faculdade de Tecnologia - UNICAMP

Profa. Dra. Marli de Freitas Gomes Hernandez

Instituição: Faculdade de Tecnologia - UNICAMP

Profa. Dra. Artêmis Maria Francelin Sanches Moroni

Instituição: CTI Renato Archer – MCTI

Profa. Dra. Mauro de Mesquita Spínola

Instituição: Politécnica – USP São Paulo

A Ata de Defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

DEDICATORIA

Dedico este trabalho a todo a minha família. Ao meu esposo Amauri que, nestes mais de 37 anos de casados, sempre foi meu companheiro e principal motivador.

Aos meus filhos Raquel, Raphael e Gabriella pelo amor recebido em suas diferentes manifestações e pela grata oportunidade de acompanhar, suas conquistas pessoais, acadêmicas e profissionais. E por serem admiráveis exemplos no estabelecimento de um convívio harmônico familiar e com amigos.

Aos meus netos gêmeos Juliana e Matheus e ao meu genro Fernando que entraram em nossa família renovando energias e proporcionando novas alegrias.

A minha mãe Dulce, hoje prestes a completar 90 anos, por sempre estar ao meu lado reforçando sua crença em mim, afirmando “Você pode chegar lá”.

E ao meu pai Abel, que nos deixou ainda jovem, e de quem a presença sempre fez falta em todos os momentos de conquistas da minha família.

AGRADECIMENTOS

Foram muitos que contribuíram com esta minha caminhada, mas de forma especial, agradeço:

A instituição em que trabalho, CTI Renato Archer, órgão vinculado ao Ministério De Ciência, tecnologia e Inovação – MCTI que, através do atual diretor Dr. Jorge Vicente L. da Silva e da Coordenadora Geral de Projetos e Serviços Profa. Dra. Juliana K. M. B. Daguano por fornecerem o apoio necessário para a conclusão deste trabalho e a todos os meus bolsistas PIBIC que de alguma forma contribuíram para este trabalho e em especial a Talita e o Yuri que foram grandes parceiros trazendo alternativas para o desenvolvimento do Sw3T.

Ao Prof. Paulo que me aceitou como orientanda, por ter sido decisivo na escolha do tema final da pesquisa e ter permitido um coorientador. Tais decisões foram fundamentais para a conclusão deste trabalho, e apesar de estar afastado para tratamento de saúde, manteve-se presente através de suas mensagens de incentivo.

A Profa. Marli, ex-colega do corpo de docente da Faculdade de Tecnologia, onde tive a grata experiência de trabalhar por 3 anos, uma grande amiga que esteve o tempo todo ao meu lado, por acreditar em mim e não me deixar esmorecer. Sua presença teve um importante papel ao me manter conectada a instituição e em especial a coordenação de pós-graduação.

Por fim, agradeço imensamente ao Prof. Marcelo que considero um amigo pessoal. Trabalhamos juntos na ABNT, e em inúmeros trabalhos representando o CTI Renato Archer e a USP. Como meu orientador, o Prof. Marcelo trouxe a guiança segura e competente, ideias e sugestões valiosas. Compartilhou seu conhecimento e sua longa experiência em orientação. Me presenteou com seu apoio absoluto, cedendo horas de seus dias, inclusive noites de seu precioso tempo. Foi parceiro em extensas discussões e não mediu esforços na revisão deste trabalho, viabilizando a sua conclusão em tempo hábil. Para este grande homem, pela sua cordialidade, carinho e cumplicidade demonstrada nesse período de trabalho conjunto, deixo meu muito obrigado e eterna gratidão.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

A Internet revolucionou a forma de se fazer revisão literária, disponibilizando diversas bases de dados científicas e uma grande quantidade de publicações a serem analisadas. Embora existam diversas ferramentas de busca e alguns softwares de apoio bastante utilizados, a tarefa ainda é difícil, exaustiva e demorada, especialmente na análise de resumos, pela diversidade de tópicos encontrados nas publicações e a falta de um agrupamento dos temas semelhantes. O objetivo deste trabalho é propor um processo que contribua com a celeridade na execução de uma RSL. Integrado a uma ferramenta de software desenvolvida com técnicas de Processamento de Linguagem Natural (PLN), o novo processo trata e classifica textos, agrupando-os com temas semelhantes, inspirados no que as áreas de mineração de texto e inteligência artificial realizam. O método de pesquisa-ação foi empregado nesta pesquisa e a aplicação da solução em três diferentes revisões literárias, permitiu desenvolver o novo processo e a ferramenta denominada Sw3T, com modelagem de tópicos Latent Dirichlet Allocation (LDA). Participantes de outras instituições se propuseram a avaliar a ferramenta e o processo proposto, permitindo futuros aperfeiçoamentos. Os resultados, nos casos concluídos, demonstraram a eficácia do método, obtendo melhor estruturação, disciplina e rapidez, preservando transparência, rastreabilidade e confiabilidade das revisões realizadas. O novo processo proposto de revisão sistemática da literatura demonstrou ser eficaz e capaz de trazer benefícios para pesquisadores, além de manter acessível o histórico das informações das operações. Uma maior gama de aplicações pode trazer um aperfeiçoamento do processo e o balanceamento dos parâmetros do algoritmo. Como trabalhos futuros, pode-se estudar a aplicação de outras técnicas de inteligência artificial e de mineração de texto, técnicas de parametrização e inclusive a contribuição de outras abordagens como a realidade aumentada para auxiliar na análise e seleção das publicações em representações tridimensionais.

Palavras-chave: Revisão literária; Modelagem de tópicos; LDA; Processamento da Linguagem Natural; Processo de RSL; ferramenta de RSL.

ABSTRACT

The Internet revolutionized the way literary reviews are conducted by providing access to various scientific databases and a vast number of publications to be analyzed. Although there are numerous search tools and widely used support software, the task remains difficult, exhaustive, and time-consuming, particularly when dealing with abstracts, due to the diversity of topics covered in the analyzed publications and the lack of a sorting mechanism for similar themes. The objective of this study is to propose a process that contributes to expediting the execution of an SLR. Integrated with a software tool developed with Natural Language Processing (NLP) techniques, the new process handles and classifies texts, grouping them with similar themes, inspired by what the areas of text mining and artificial intelligence do.. The action-research method was employed in this research and the application of the solution in three different literary reviews, allowed the development of the new process and the tool called Sw3T, with Latent Dirichlet Allocation (LDA) topic modeling. Participants from other institutions volunteered to evaluate the proposed tool and process, allowing for future improvements. The results from completed cases demonstrated the effectiveness of the process, achieving better structuring, discipline, and speed, while preserving the transparency, traceability, and reliability of the systematic literature reviews. The proposed SLR process demonstrated that it is effective and brings benefits to researchers while keeping the history of operations information accessible. A broader range of applications could enhance the process and balance the algorithm's parameters. IN future work, the application of other artificial intelligence and text mining techniques, parameterization techniques, and even the contribution of other approaches, such as augmented reality, to assist in the analysis and selection of publications in three-dimensional representations, could be explored.

Keywords: Literary review; Topic modeling; LDA; Natural Language Processing; RSL process; RSL tool

LISTA DE ILUSTRAÇÕES

FIGURA 1 - <i>QUERY (STRING)</i> GENÉRICA UTILIZADA NA BUSCA DE PUBLICAÇÕES.	26
FIGURA 2 - PRISMA - TOTAIS DAS BUSCAS DA RSL DE 2021 E MOTIVOS DE EXCLUSÃO	29
FIGURA 3 - CRESCIMENTO DE PUBLICAÇÕES SOBRE O TEMA EM ATÉ 2022 (DA AUTORA)	30
FIGURA 4 - PASSOS DO CICLO DA PESQUISA (ADAPT. DE COUGHLAN&COUGHLAN, 2002)	46
FIGURA 5 - CARACTERÍSTICAS DAS ETAPAS DE CONDUÇÃO.....	48
FIGURA 6 - ESTRUTURA DA PESQUISA - ETAPAS E FASES	49
FIGURA 7 - HIERARQUIA DO APRENDIZADO DE MÁQUINA (MONARD&BARANAUSKAS, 2003)...	57
FIGURA 8 - FLUXOGRAMA DA EXECUÇÃO DE UM PLN (YUEKANG ET AL, 2020).....	58
FIGURA 9 - PROCESSO DE REVISÃO DA LITERATURA COM O SW3T	73
FIGURA 10 – VISÃO PARCIAL DA TELA DO SW3T DE VISUALIZAÇÃO DE TÓPICOS	77
FIGURA 11 - GRÁFICO LDAVIS (<i>NATURE</i> , 2022) DO SW3T - PROXIMIDADE DOS CLUSTERS.	78
FIGURA 12 – TELA DO SW3T COM RESULTADO DETALHADO DA ANÁLISE	79
FIGURA 13 – TELA DO SW3T COM TODAS AS ANÁLISES DO PROJETO.....	79
FIGURA 14 - TELA DO SW3T COM HISTÓRICO DE SELEÇÃO	80
FIGURA 15 - MODELO DE DADOS DO SW3T.....	81
FIGURA 16 - TELA PRINCIPAL DO APLICATIVO SW3T	82
FIGURA 17 - EXEMPLO DO ARQUIVO TXT COM A LISTA DE STOPWORDS DO USUÁRIO	86
FIGURA 18 - PLANILHA COM TÓPICOS, PALAVRAS-CHAVE E SEU PESO NO CORPUS.....	87
FIGURA 19 - 15 TÓPICOS DA RSL DE SEGURANÇA DA INFORMAÇÃO (SERGIO ET AL, 2019)	87
FIGURA 20 - TELA DO SW3T COM A MELHOR ANÁLISE PARA O TEMA WHATSAPP	89
FIGURA 21 - TELA DO SW3T COM HISTÓRICO DE SELEÇÃO DO CICLO 2.....	90
FIGURA 22 - PRISMA DO TERCEIRO CICLO DA PESQUISA-AÇÃO.....	92
FIGURA 23 - TELA DO SW3T COM 8 TÓPICOS DE 251 RESUMOS PUBLICADOS ATÉ 2016	93
FIGURA 24 - TELA DO SW3T DE 8 TÓPICOS PARA 329 PUBLICAÇÕES DE 2017 À 2019	93
FIGURA 25 – TELA DO SW3T COM 6 TÓPICOS PARA 102 PUBLICAÇÕES ENTRE 2020 E 2023.....	94
FIGURA 26 - DESDOBRAMENTO DA QUESTÃO DA PESQUISA	100
FIGURA 27 - TELA INICIAL NO SW3T.....	110
FIGURA 28 - TELA GERAL DO SW3T - TODOS OS PROJETOS DE UM USUÁRIO	110
FIGURA 29 - TELA DO SW3T PARA INCLUIR NOVO PROJETO	111
FIGURA 30 - TELA COM AS FUNÇÕES DO SW3T NUM PROJETO	111
FIGURA 31 - TELA DO SW3T PARA INCLUIR ARQUIVO XML DE METADADOS.....	112
FIGURA 32 - TELA DO SW3T PARA TRATAR METADADOS DUPLICADOS	113
FIGURA 33 - TELA DO SW3T PARA VER TODOS OS METADADOS DE UM PROJETO	113
FIGURA 34 - TELA DO SW3T DE PESQUISA NOS RESUMOS.....	114
FIGURA 35 - TELA DO SW3T PARA ATUALIZAR A TABELA DE BDS FONTE	115
FIGURA 36 - TELA DO SW3T PARA ATUALIZAR TABELA DE/PARA.....	115
FIGURA 37 - TELA DO SW3T PARA ATUALIZAR TABELA <i>STOPWORDS</i> DO USUÁRIO	116
FIGURA 38 - TELA DO SW3T PARA EXPORTAR TABELAS PARA OUTRO PROJETO	116
FIGURA 39 - TELA DO SW3T PARA PREPARAR RESUMOS E FAZER ANÁLISE.....	117

FIGURA 40 - TELA DO SW3T PARA VISUALIZAR UMA ANÁLISE	118
FIGURA 41 - TELA DO GRÁFICO DO SW3T QUE MOSTRA PROXIMIDADE ENTRE CLUSTERS	118
FIGURA 42 - TELA DO SW3T COM RESUMOS DE UM CLUSTER	119
FIGURA 43 - TELA DO SW3T DAS ETAPAS DO TRATAMENTO DOS RESUMOS.....	119
FIGURA 44 - TELA DO SW3T COM TODAS AS ANÁLISES DE UM PROJETO	120
FIGURA 45 - TELA DO SW3T COM HISTÓRICO DE SELEÇÃO DOS METADADOS DO PROJETO.....	121

LISTA DE TABELAS

TABELA 1 - FASES E ATIVIDADES DO MÉTODO KITCHENHAM (KITCHENHAM, 2004).....	23
TABELA 2 - STRING, BDS E RECURSOS DE DOWNLOAD DE PUBLICAÇÕES SELECIONADAS	28
TABELA 3 - RELAÇÃO DE PUBLICAÇÕES DE TRABALHOS CORRELATOS DE RLS 2018 E 2022.....	31
TABELA 4 - ESTRATÉGIA DE PESQUISA, ADAPTADO DE YIN (YIN, 2005)	43
TABELA 5 - DESCRIÇÃO DAS FASES DA PESQUISA E RESPECTIVOS ENTREGÁVEIS.....	50
TABELA 6 - RELAÇÃO DOS OBJETIVOS ESPECÍFICOS E SUAS SIGLAS	51
TABELA 7 - PROPOSIÇÕES, SUA DESCRIÇÃO E RELAÇÃO COM OS OBJETIVOS ESPECÍFICOS	52
TABELA 8 - PONTO DE ANÁLISE PA1 - SELEÇÃO DO PROCESSO	53
TABELA 9 - PONTO DE ANÁLISE PA2 - NECESSIDADE DE PROGRAMAÇÃO.....	53
TABELA 10 - PONTO DE ANÁLISE PA3 - MOTIVO PARA ADOÇÃO DO PROCESSO.....	54
TABELA 11 - TRATAMENTO DE CADA CORPORA (ELABORADO PELA AUTORA).....	66
TABELA 12 - QUESTÕES ESPECÍFICAS DA PA ATENDIDAS.....	88
TABELA 13 - RESULTADO DA ANÁLISE DAS QUESTÕES ESPECÍFICAS DO CICLO 2	91
TABELA 14 - RESULTADO DA ANÁLISE DAS QUESTÕES ESPECÍFICAS DO CICLO 3	95

Nomenclaturas e Siglas

AM - Aprendizado de Máquina (ML - *Machine Learning*)

BDs - Bases de Dados (forma de armazenar informação em ambiente digital)

BERT - Acrônimo para *Bidirectional Encoder Representations from Transformers* - modelo de pré-treinamento de PLN

BERTopic - Bidirectional Encoder Representations from Transformers like a Topic Modeling Technique

CSS - *Cascading Style Sheets*

Doc2Vec - *Document To Vector*

FAWOC - *Fast Word*

HDP - *Hierarchical Dirichlet Process* (PHD - Processo Hierárquico de Dirichlet)

HTML - *Hypertext Markup Language*

IA - Inteligência Artificial (AI – *Artificial Intelligent*)

InPRA – *Intelligent Personal Research Assistant*

LDA - *Latent Dirichlet Allocation*

LDA2vec - *Latent Dirichlet Allocation to Vector*

LSA - *Latent Semantic Analysis*

LSI - *Latent Semantic Indexing* (LSA do Google)

MySQL - Sistema de gerenciamento de banco de dados

PHP - acrônimo de *Hypertext Preprocessor*

PLN - Processamento da Linguagem Natural (NLP)

pLSA - *Probabilistic Latent Semantic Analysis*

pLSI - *Probabilistic Latent Semantic Indexing* (o mesmo que pLSA para o Google)

Python - Linguagem de programação de alto nível

RSL - Revisão Sistemática da Literatura (SLR - *Systematic Literature Review*)

Skip-gram or N-gram - Algoritmo de previsão de palavras de contexto de PLN

SVD - Singular Values Decomposition

Sw3T - Software Tópico Tem Termos (*Software Term To Topic*)

TF-IDF - *Term Frequency–Inverse Document Frequency*

TI - Tecnologia da Informação

Top2Vec - *Topics To Vector*

Word2vec - *Words To Vector*

SUMÁRIO

1. INTRODUÇÃO	16
1.1 OBJETIVOS.....	17
1.1.1 OBJETIVOS ESPECÍFICOS	18
1.2 ESTRUTURAÇÃO DO TRABALHO	18
2. CONCEPÇÃO DA REVISÃO LITERÁRIA DESTA PESQUISA	20
2.1 INTRODUÇÃO	20
2.2 ESTUDOS SOBRE REVISÕES DE LITERATURA E PRODUÇÕES CIENTÍFICAS	20
2.2.1 MÉTODO DE REVISÃO SISTEMÁTICA DA LITERATURA PROPOSTA POR KITCHENHAM	21
2.3 CONCEPÇÃO E PROTOCOLO DA RSL DESTA PESQUISA	24
2.3.1 O MODELO PICO.....	24
2.3.2 TERMOS DE BUSCA E BASES DE DADOS	25
2.4 A REVISÃO LITERÁRIA E OS ACHADOS OBSERVADOS.....	25
2.4.1 TRABALHOS CORRELATOS.....	30
2.4.2 FERRAMENTAS QUE APOIAM UMA REVISÃO LITERÁRIA	37
2.5 CONSIDERAÇÕES FINAIS	38
3. METODOLOGIA – ESTRUTURAÇÃO DA PESQUISA.....	39
3.1 INTRODUÇÃO	39
3.2 METODOLOGIA DE PESQUISA E MÉTODOS EXISTENTES.....	39
3.3 CARACTERIZAÇÃO DA PESQUISA	40
3.3.1 QUANTO AOS CRITÉRIOS DE CARACTERIZAÇÃO.....	40
3.3.2 QUANTO À NATUREZA DAS VARIÁVEIS	42
3.3.3 QUANTO À ESTRATÉGIA DA PESQUISA	43
3.3.4 FECHAMENTO DA CARACTERIZAÇÃO.....	44
3.4 CONCEITOS DE PESQUISA-AÇÃO.....	45
3.5 ESTRUTURA DA PESQUISA.....	48
3.5.1 ETAPA DE PLANEJAMENTO DA PESQUISA	50
3.6 CONSIDERAÇÕES FINAIS	54
4. PROCESSAMENTO DA LINGUAGEM NATURAL (PLN)	55
4.1 INTRODUÇÃO	55
4.2 INTELIGENCIA ARTIFICIAL, MINERAÇÃO DE TEXTO E RAMIFICAÇÕES.....	55
4.3 CONCEITO E TÉCNICAS DO PLN	57
4.4 TIPOS DE MODELAGENS DE TÓPICOS.....	62
4.4.1 LATENT SEMANTIC ANALYSIS (LSA) / LATENT SEMANTIC INDEXING (LSI p/ GOOGLE) .	63

4.4.2 ANÁLISE SEMÂNTICA LATENTE PROBABILÍSTICA (PLSA)	63
4.4.3 LATENT DIRICHLET ALLOCATION (LDA)	64
4.4.4 LDA EM DEEP LEARNING: LDA2VEC.....	64
4.4.5 INDEXAÇÃO SEMÂNTICA LATENTE (LSI).....	65
4.4.6 PROCESSO HIERÁRQUICO DE DIRICHLET (HDP)	65
4.4.7 BERTOPIC E TOP2VEC	65
4.4.8 EXEMPLO DE LEMATIZAÇÃO E N-GRAM	66
4.5 OUTRAS TECNOLOGIAS UTILIZADAS.....	67
4.6 CONSIDERAÇÕES FINAIS	70
5. PROCESSO PROPOSTO DE REVISÃO LITERÁRIA COM SW3T	71
5.1 INTRODUÇÃO	71
5.2 REQUISITOS DO PROCESSO DE RSL COM SW3T.....	71
5.3 O PROCESSO DE RSL COM O SW3T	72
5.4 A FERRAMENTA SW3T	80
5.4.1 A ARQUITETURA DO BANCO DE DADOS MYSQL	80
5.4.2 FUNÇÕES DA FERRAMENTA SW3T.....	82
5.5 CONSIDERAÇÕES FINAIS	84
6. APLICAÇÃO DO PROCESSO DE RSL COM SW3T	85
6.1 INTRODUÇÃO	85
6.2 PLANEJAMENTO DA PESQUISA-AÇÃO	85
6.3 CICLO 1 – RSL NA PESQUISA SOBRE SEGURANÇA DA INFORMAÇÃO	85
6.4 CICLO 2 – RSL NA PESQUISA SOBRE WHATSAPP COMO PROVA JUDICIAL	89
6.5 CICLO 3 – RSL NA PESQUISA SOBRE RSL COM MODELAGEM DE TÓPICOS	92
6.6 EFICÁCIA DA FERRAMENTA NUMA RSL	96
6.6.1 REVISÃO LITERÁRIA DE CIDADES TROPICAIS	96
6.6.2 REVISÃO LITERÁRIA DE UMA PESQUISADORA DE DOUTORADO	97
6.6.3 CONSIDERAÇÕES SOBRE A QUESTÃO ESPECÍFICA DO PA1	97
6.7 CONSIDERAÇÕES FINAIS	98
7. AVALIAÇÃO DO PROCESSO RSL COM SW3T	99
7.1 INTRODUÇÃO	99
7.2 FASE 5 – APRESENTAÇÃO DE RESULTADOS	99
7.3 QUANTO AO ATENDIMENTO À PROPOSTA DE PESQUISA.....	99
7.4 REFLEXÕES SOBRE A EXPERIÊNCIA VIVENCIADA	103
7.5 CONCLUSÕES E PESQUISAS FUTURAS	103

REFERÊNCIAS.....	105
APÊNDICE I – MANUAL DO SW3T	110
APÊNDICE 2 – POSTER PIBIC E INDICAÇÃO AO PRÊMIO IC 2018	122
APÊNDICE 3 – ARTIGO PUBLICADO EM 2019 NO IEEE.....	124
APÊNDICE 4 – TRECHO DO TCC COM USO DO SW3T EM 2022	125

1. INTRODUÇÃO

Nas pesquisas científicas, a atividade de revisão literária sempre foi um desafio mesmo antes do advento da internet. Os levantamentos se restringiam ao acervo das bibliotecas a que os pesquisadores tinham acesso físico, acarretando até mesmo a necessidade de seu deslocamento para outras instituições de pesquisa. A Internet revolucionou a forma de fazer revisão literária, estabelecendo uma situação inversa onde grande quantidade de publicações está disponível (Facchinetti et al, 2022), criando um novo desafio a ser superado - Como lidar com o excesso de informações?

Atualmente, dentre os vários desafios de uma revisão literária destacam-se a necessidade de se estabelecer uma estratégia para lidar com a abundância de informações disponíveis em ambiente digital, que inclui um grande número de bases de dados científicas e a popularização da literatura cinzenta com informações de origem não científica (Dudziak, 2021)(Nature, 2022) (Newcastle, 2023),mas que contribuem com informações relevantes, trazendo, em primeira mão, inovações tecnológicas.

Existem barreiras para pesquisadores com recursos limitados que precisam de acesso as publicações, uma vez que muitas Bases de Dados (BDs) o acesso as publicações na íntegra não é gratuito. Pesquisadores que possuem ligação com as universidades públicas ou mesmo privadas podem ter diferentes níveis de acesso dependendo de quais assinaturas a instituição disponibiliza. Vale ressaltar que, na maioria das BDs de acesso particularizado, o acesso irrestrito está limitado aos metadados (Facchinetti et al, 2022), como as informações de título, palavras-chave, resumo e referências dos textos.

Manter-se atualizado em um determinado campo também é um desafio. A realização de uma revisão literária abrangente leva tempo e precisa ser célere. Constantemente, novas pesquisas estão sendo publicadas, o que pode levar à rápida desatualização da revisão, comprometendo a validade dos resultados da pesquisa.

Há a necessidade de definir estratégias de como selecionar informações com uma visão objetiva e imparcial. É necessário evitar que o pesquisador se baseie

em seu próprio viés e subjetividade o que pode influenciar na interpretação da literatura e levar a um processo de seleção não replicável.

No geral, a realização de uma revisão literária requer um alto nível de rigor, pensamento crítico e atenção aos detalhes. Os pesquisadores precisam estar atentos aos desafios que podem enfrentar e tomar medidas para enfrentá-los e garantir que a revisão literária seja relevante e confiável (Kitchenham&Brereton, 2013). Este trabalho de pesquisa adota o método de RSL de Kitchenham (Kitchenham, 2004) por fornecer uma análise abrangente e objetiva de pesquisas existentes e ser referência na engenharia de software, área desta pesquisa.

Estão disponíveis diversos aplicativos conceituados que auxiliam a atividade de revisão. Aplicativos como StArt, JabRef e Parsifal são referências no suporte a RSLs e apresentam valiosas funcionalidades (Vosgerau&Romanowski, 2014). Porém, quando se trata de dar suporte à seleção pela leitura dos resumos, ainda são precários os recursos disponíveis. O conjunto de publicações pré-selecionadas geralmente diversifica na abordagem de pesquisa e apresenta diferentes ângulos sobre o mesmo tema. Assim, sem uma classificação prévia dos resumos por tema ou escopo de pesquisa, o processo de seleção é lento e cansativo.

No entanto, existem soluções tecnológicas vinculadas ao aprendizado de máquina e à Inteligência Artificial (IA), envolvendo o uso de algoritmos matemáticos e modelos estatísticos que identificam padrões e relacionamentos entre dados para que a máquina entenda a linguagem humana e faça previsões ou decisões. Há frentes de pesquisa que já são capazes de interagir com a voz humana, textos, expressões faciais, dentre outras possibilidades (Khanzode et al., 2020).

1.1 OBJETIVOS

Para responder à pergunta “**O que é?**”, que define o objetivo geral deste trabalho acadêmico, temos:

Desenvolver um processo de revisão literária que utilize técnicas de Processamento da Linguagem Natural (PLN) com base no método revisão sistemática da literatura (RSL) para engenharia de software proposto por Kitchenham e Brereton (Kitchenham&Brereton, 2013).

1.1.1 Objetivos específicos

Nos objetivos específicos respondemos “**Como**” será desenvolvida a pesquisa. Os objetivos específicos desta pesquisa são:

- Identificar soluções de RSL para preencher lacunas que existem no sentido de atender aos requisitos de Qualidade: Rapidez, Transparência, Confiabilidade e Rastreabilidade.
- Estabelecer um processo que possa dar celeridade a uma revisão sistemática da literatura pelo uso de tecnologia de mineração de texto.
- Identificar técnicas de PLN que possam ser aplicadas a uma revisão literária.

1.2 ESTRUTURAÇÃO DO TRABALHO

Este trabalho está organizado em sete capítulos, sendo o primeiro a contextualização do problema, objetivo geral e objetivos específicos.

O segundo capítulo trata da fundamentação teórica do trabalho. Apresenta um breve relato das fases e etapas da RSL de *Kitchenham* (Kitchenham&Brereton, 2013) e identifica em que espectros o novo processo pode contribuir. Descreve a revisão sistemática da literatura (RSL) desta pesquisa, identifica trabalhos correlatos que mapeiam o estado da arte no tema e apontam lacunas para o desenvolvimento da agenda de pesquisa.

O terceiro capítulo tem como foco descrever o método pesquisa-ação adotado neste trabalho e justificar a aderência ao projeto. Apresenta a estrutura da pesquisa organizada em três etapas e relata o resultado da Etapa 1 – planejamento da pesquisa.

O quarto capítulo descreve os principais conceitos técnicos que embasaram este trabalho e desenvolve o entendimento sobre modelagem de tópicos, justificando a escolha do modelo probabilístico *Latent Dirichlet Allocation* (LDA).

O quinto capítulo descreve o processo de RSL proposto e a ferramenta denominada Sw3T desenvolvidos neste trabalho.

O sexto capítulo discute o realizado na Etapa 2 – ciclo da pesquisa-ação da estrutura da pesquisa. Descreve como foram desenvolvidos o novo processo de RSL e a ferramenta Sw3T. Detalha os casos reais executados, o que foi realizado em cada ciclo e a avaliação dos pontos de análise, discutindo a eficácia do processo e da ferramenta.

Finalmente o sétimo capítulo discute os resultados, como previsto na Etapa 3 da estrutura da pesquisa. Avalia o trabalho realizado. Retoma a pergunta da pesquisa e o atendimento aos seus objetivos. Conclui com a reflexão sobre pontos fortes e lições aprendidas, sugerindo futuros trabalhos.

2. CONCEPÇÃO DA REVISÃO LITERÁRIA DESTA PESQUISA

2.1 INTRODUÇÃO

Neste capítulo é feito um breve relato sobre os diferentes tipos de produções científicas que realizam revisão de literatura, dando destaque ao método de revisão sistemática da literatura para engenharia de software, proposto por *Kitchenham* (Kitchenham, 2004) usado como base deste trabalho. Relata a revisão literária realizada nesta tese, desde seu planejamento até a identificação de trabalhos relacionados estabelecendo o estado da arte no tema e lacunas de pesquisa exploradas.

2.2 ESTUDOS SOBRE REVISÕES DE LITERATURA E PRODUÇÕES CIENTÍFICAS

Vosgerau e Romanowski (Vosgerau&Romanowski, 2014) apresentam uma diversidade de termos encontrados para denominar os estudos que realizam revisões. Uma vez que as expressões representam o mesmo tipo de estudo, para melhor entender a relação existente propõem o agrupamento dos tipos de revisão, em dois grupos: as revisões que mapeiam e as revisões que avaliam e sintetizam.

Em outro ponto, Vosgerau e Romanowski (Vosgerau&Romanowski, 2014) esclarecem que as revisões que mapeiam levantam as referências encontradas sobre um determinado tema, para um primeiro contato com o objeto de estudo investigado, com a finalidade de delimitar, clarificar e caracterizar o objeto de estudo. Inclui os seguintes termos: Levantamento Bibliográfico; Revisão de Literatura ou Revisão Bibliográfica; estudo do tipo estado do conhecimento; Revisão Narrativa; estudos bibliométricos e análise de indicadores bibliométricos.

Contudo, as revisões que mapeiam produzem um grande volume de evidências no seu apontamento para novas pesquisas. Thomas (Thomas, 2007) afirma que é necessário verificar como e em que condições essas evidências se repetem e mesmo sobre a aplicabilidade dessas tanto para a formulação de políticas de fomento à pesquisa, como no processo de formação dos profissionais da educação (Thomas, 2007).

Para Vosgerau e Romanowski (Vosgerau&Romanowski, 2014), revisões que avaliam e sintetizam visam formular a questão de investigação, no estabelecimento de estratégias de diagnóstico crítico e na exigência na transparência para o estabelecimento de critérios para inclusão e exclusão dos estudos, necessariamente primários, ou seja, coletados pelo próprio pesquisador, tais como surveys, entrevistas, observações, relatos etc. (Davies, 2007) . Estes estudos têm sua origem na área de ciências da saúde, pois buscam identificar as condições em que determinadas evidências ocorrem e a possibilidade de identificação de padrões de ocorrência. Inclui as seguintes denominações: revisões sistemáticas, revisão integrativa, síntese de evidências qualitativas, metassíntese qualitativa, meta-análise e metassumarização.

Uma vez que este trabalho visa uma contribuição abrangente de uma RSL, adotou-se o Método de Kitchenham (Kitchenham, 2004) revisão sistemática da literatura para engenharia de Software, referência na área, descrito a seguir.

2.2.1 Método de revisão sistemática da literatura proposta por Kitchenham

O método de revisão sistemática da literatura (RSL) para engenharia de software foi desenvolvido por Kitchenham (Kitchenham, 2004) com o objetivo de estabelecer uma avaliação justa de um tópico de pesquisa, usando uma metodologia confiável, rigorosa e auditável. Sua diretriz deriva de três diretrizes usadas por pesquisadores médicos, adaptadas para refletir os problemas específicos da engenharia de software.

A finalidade do método Kitchenham (Kitchenham, 2004), é fornecer uma análise abrangente e objetiva de pesquisas existentes numa área específica desejada. É levar o pesquisador a identificar lacunas no conhecimento e tomar decisões sobre como direcionar sua pesquisa com base em evidências obtidas na revisão literária. Envolve localizar, analisar, sintetizar e interpretar pesquisas já existentes sobre o tema de interesse do pesquisador.

Cardoso, Alarcão e Celorico (Cardoso&Alarcão&Celorico, 2010) ressaltam que a qualidade das publicações a serem selecionadas deve atender a critérios como relevância, validade e confiabilidade do conteúdo da publicação.

Selecionar publicações empíricas, estudos de caso ou pesquisa-ação indispensáveis para definir bem o problema e obter uma ideia precisa sobre o estado da arte de um determinado tema o que deve levar o pesquisador a identificar lacunas de conhecimento e oportunidades de contribuição da pesquisa e, garantindo que o esforço efetivo da pesquisa em curso contribuirá para o desenvolvimento de novos conhecimentos.

Segundo Kitchenham (Kitchenham et al, 2010), entre as razões para se realizar a RSL, estão a de resumir as evidências relativas a um tratamento de informação, identificar lacunas observadas e sugerir pesquisas adicionais em uma determinada área. Uma RSL deve fornecer uma base para posicionar adequadamente atividades de uma nova pesquisa.

Uma RSL surge para obter todas as informações existentes sobre algum fenômeno de maneira completa e imparcial. Em outro ponto, Kitchenham (Kitchenham et al, 2010) coloca que a maioria das pesquisas começa com uma revisão literária de algum tipo. No entanto, a menos que ela seja completa e justa, terá pouco valor científico.

Para reduzir um possível viés do pesquisador, Kitchenham (Kitchenham, 2004) sugere que outro pesquisador participe no processo de análise e decisão sobre as publicações. Em RSL de teses, a autora preconiza que o orientador pode desempenhar esse papel.

Kitchenham e Brereton (Kitchenham&Brereton, 2013) afirmam que uma das principais características de uma RSL é o protocolo de revisão, o método de revisão, a documentação da estratégia de pesquisa, os critérios de inclusão e exclusão, o registro das informações obtidas e o resultado da avaliação de estudos primários incluindo critérios de qualidade. O relatório da RSL deve ser claro e transparente, incluindo um resumo das publicações selecionadas na revisão.

O método Kitchenham (Kitchenham, 2004) envolve oito atividades distribuídas em três fases: [1] planejamento da revisão, [2] realização da revisão, [3] relato da revisão conforme

Tabela 1 - Fases e atividades do método Kitchenham (Kitchenham, 2004)

FASE	ATIVIDADES
[1] planejamento da revisão	1. Identificação da necessidade de uma revisão 2. Desenvolvimento de um protocolo de revisão.
[2] realização da revisão	3. Identificação de pesquisa 4. Seleção de estudos primários 5. Avaliação da qualidade do estudo 6. Extração e monitoramento de dados 7. Síntese de dados.
[3] relato da revisão	8. Envolve a redação do resultado do trabalho.

Enquanto a “Fase Planejamento da Revisão” se destaca por depender exclusivamente do conhecimento do pesquisador, a “Fase Realização da Revisão” contém etapas com atividades mecânicas e repetitivas no que se refere a seleção das publicações nas bases de dados (Sergio et al, 2019) .

Schoppfel (Schopf et al., 2021) ressalta a importância de estar bem definido o conjunto de termos de busca e as fontes de dados a serem consultadas a partir da questão de pesquisa.

Uma importante estratégia para reduzir o volume de publicações encontrado na busca, está na escolha do conjunto de termos de busca. Porém, nem sempre os termos são facilmente identificados pelo pesquisador. Geralmente a busca é iniciada utilizando um único termo e à medida que as publicações são estudadas, o pesquisador vai introduzindo novos termos, de inclusão ou de exclusão, filtrando melhor a busca pelas publicações. Esse processo se repete até que o pesquisador entenda que o subconjunto de artigos selecionados represente, potencialmente, os trabalhos científicos mais relevantes para sua pesquisa.

Com base na estratégia de seleção previamente definida na etapa de planejamento, estas publicações são submetidas a leitura detalhada de conteúdo a fim de validar sua relevância. As etapas “**selecionar estudos primários**” e “**documentar a revisão**” demandam trabalho e tempo do pesquisador, além de exigirem muita precisão na seleção o que faz com que o processo proposto neste trabalho seja uma ajuda valiosa.

2.3 CONCEPÇÃO E PROTOCOLO DA RSL DESTA PESQUISA

Nesta subseção é descrita a revisão sistemática da literatura desta pesquisa evidenciando como se chegou ao estado da arte e à identificação de lacunas de conhecimento. A revisão literária foi realizada, inicialmente, em 2018 e replicada em meados de 2021 utilizando o aplicativo Mendeley. Uma nova pesquisa foi realizada em dezembro de 2022 como parte do Ciclo 3 da pesquisa-ação, visando apenas verificar o atendimento aos pontos de análise deste último ciclo de pesquisa.

Na subseção são relatados o modelo PICO (Durkin&Durkin, 1998) DURKIN, J., & DURKIN, J. 1998. Expert systems: design and development. Prentice Hall PTR. (Echevarria, 2014) utilizado para definir a questão de pesquisa, as bases de dados escolhidas e os termos de busca utilizados.

2.3.1 O modelo PICO

O modelo tem por objetivo auxiliar na reflexão sobre o problema de pesquisa e na definição da questão de pesquisa. O modelo PICO (Santos&Pimenta&Nobre, 2007). contém 4 perguntas: **P**opulação, **I**ntervenção, **C**omparação e **R**esultado esperado (**Outcome**). Uma vez esclarecidos esses pontos, a questão de pesquisa pode ser definida com consistência. Para esta pesquisa foram obtidas as seguintes respostas:

(P) – População: Pesquisadores e alunos de pós-graduação.

(I) – Intervenção: Busca de uma solução para melhorar o processo de revisão literária utilizando técnicas de PLN (Ellmann, 2018) na seleção de publicações na análise por resumo dentro de um determinado tema.

(C) – Comparação: Funcionalidades e abrangência de ferramentas de apoio à revisão literária.

(O) – Resultado esperado: Obter um processo de revisão literária que aumente a produtividade e permita a rastreabilidade das atividades realizadas com o apoio de uma ferramenta.

Neste contexto, obtém-se a seguinte questão de pesquisa:

Como as técnicas de Processamento da Linguagem Natural (PLN) podem contribuir com a revisão literária de uma pesquisa científica?

2.3.2 Termos de busca e bases de dados

Tendo como ponto de partida a “Questão de pesquisa”, os termos utilizados para seleção de publicações nas bases de dados foram: **Literature Review**, **Bibliografic Review**, **Bibliographical Review**, **Topic Model**, **LDA**, **Latent Dirichlet Allocation**. A seguir são descritas as justificativas dos termos utilizados.

Revisão Literária (literature review, bibliografic review, bibliographical review): Termos recorrentes em publicações científicas. Consiste na identificação de publicações relevantes que devem levar ao estado da arte e possíveis lacunas a serem preenchidas sobre o tema de pesquisa

Modelo de tópico (*Topic Model*): Utilizado como parte das técnicas do PLN. São algoritmos que analisam palavras em textos processados e identificam os tópicos latentes (Blei&Lafferty, 2007).

Há diversas técnicas que compõem o PLN. Estas técnicas, detalhadas no capítulo 3, são combinadas interagem de diferentes maneiras, dependendo da tarefa específica do PLN e do tipo de texto que está sendo analisado.

Latent Dirichlet Allocation (LDA): Entre os tipos de Modelagem de tópicos, relacionados no capítulo 3, o LDA é o algoritmo probabilístico mais utilizado uma vez que seus resultados são os mais precisos (Blei&Lafferty, 2007). Para este levantamento, utilizou-se o termo de busca LDA, substituindo o termo PLN, com o objetivo de reduzir o escopo das publicações a serem selecionadas.

Na próxima subseção é relatada a revisão literária e os trabalhos correlatos identificados na RSL apontando as contribuições dos mesmos para esta pesquisa.

2.4 A REVISÃO LITERÁRIA E OS ACHADOS OBSERVADOS

A *string* genérica combinando termos de busca é apresentada na Figura 1. Os termos **bibliografic** ou **bibliographical** são pouco utilizados, mas existem publicações que os utilizam. Também pode ocorrer de apenas o termo **topic model**,

em vez de LDA, ser utilizado em títulos, resumos e palavras-chave. Da mesma forma pode aparecer apenas a sigla LDA ou apenas seu nome por extenso.

```
("bibliogra* review*" and "Latent Dirichlet Allocation") OR
("bibliogra* review*" and LDA) OR
("bibliogra* review*" and "topic model") OR
("literature review*" and "Latent Dirichlet Allocation") OR
("literature review*" and LDA) OR
("literature review*" and "topic model")
```

Figura 1 - Query (string) genérica utilizada na busca de publicações.

Como critério de inclusão e exclusão de publicações tem-se:

- Exclusão de publicações que não possuem resumo em inglês ou que não possuem resumos.
- Exclusão de publicações com termos relacionados a redes sociais como: social media, Instagram, Facebook, internet, Wikipedia, blockchain e outros, assim como quando são tratados textos de conteúdo específicos como de coleções, documentos, diagnósticos, pareceres e outros.
- Exclusão de publicações que analisam imagens, figuras, impressões digitais e outros elementos que não textos de uma publicação.
- Exclusão de publicações cuja abreviação LDA tem significado diferente de "Latent Dirichlet Allocation".

O critério inicial para seleção dos resumos é que contenham o objetivo, tipo de pesquisa, uma síntese do realizado e dos resultados obtidos. Também devem mencionar o uso de ferramentas ou do LDA para analisar publicações científicas.

Para esta revisão, foram definidas bases de dados genéricas, específicas e cinzentas (Dudziak, 2021) a serem consultadas. Nos dois primeiros, procurou-se pelas que utilizam técnicas de revisão por pares ou similares e com valor acadêmico reconhecido através de indicadores de valores altos dentro das áreas de conhecimento de tecnologia e computação.

Dessa forma entre as bases de dados genéricas tem-se a **Scopus** que é a maior base de dados de literatura revisada por pares com resumos e citações de qualidade e a **Web of Science (WoS)**, que contém os principais bancos de dados

de citações do mundo, abrangendo mais de 12.000 periódicos de grande impacto, incluindo periódicos de acesso aberto e mais de 150.000 anais de conferências com cobertura nas áreas de ciências sociais, ciências, artes e humanidades.

Entre as bases de dados específicas foram utilizadas a BD **IEEEExplore** que contém periódicos, anais de conferências, normas técnicas e materiais relacionados à ciência da computação e engenharia elétrica e eletrônica e a BD **ACM Digital Library** com periódicos científicos, revistas e anais de conferências e congressos nas áreas de Ciência da Computação, Engenharia de Computação e Tecnologia da Informação.

Uma vez que na área de tecnologia, geralmente, inovações são publicadas primeiro em ambiente não acadêmico, com o objetivo de buscar possíveis novidades tecnológicas que ainda não estejam nas bases de dados científicas, definiu-se incluir a base de dados **Google Scholar** que contém um número expressivo de publicações. O repositório da empresa Google tornou-se um dos mais importantes para realização de buscas na internet e, ao longo do tempo, passou a ser incluído em buscas de cunho acadêmico.

Dessa forma chegou-se a 5 fontes de base de dados a serem consultadas: [i] SCOPUS; [ii] ACM Digital Library (ACM); [iii] IEEE; [iv] Web of Science (WoS); [v] Google Scholar (Google).

Cada Base de Dados apresenta particularidades nos recursos de busca, necessitando a customização da string genérica. Elas também disponibilizam diferentes recursos para realizar o download dos metadados das publicações selecionadas o que exigiu validar se o conteúdo dos metadados baixados estava completo.

A Tabela 2 apresenta a string de pesquisa de cada base de dados e os tipos de arquivos para *download* disponibilizados. A busca limitou-se às publicações de periódicos e anais de congressos da base principal e não houve restrição quanto ao ano de publicação uma vez que o modelo de tópicos é um tema recente na literatura e até aquele momento haviam relativamente poucas publicações. Foram pesquisados os campos **título**, **resumo** e **palavra-chave**.

Tabela 2 - string, BDs e recursos de download de publicações selecionadas

Fontes BDs	String de pesquisa (queries)	Tipos de download
IEEE Explorer (IEEE)	("Full text & metadata": topic model and "full text & metadata": bibliogra* review** or "literature* review**") or ("all metadata": LDA and "all metadata": bibliogra* review** or "literature* review**") or ("all metadata": Latent Dirichlet Allocation and "all metadata": bibliogra* review** or "literature* review**")	Plain Text, BibTeX, RIS, CSV, RefWorks
The ACM Digital Library (ACM)	[[[all: "bibliographic review"] or [all: "literature review"]]] and [all: LDA]] or [[[all: "bibliographic review"] or [all: "literature review"]]] and [all: topic model]] or [[[all: "bibliographic review"] or [all: "literature review"]]] and [all: Latent Dirichlet Allocation]]	BibTeX, EndNote, ACM ref
Scopus	title-abs-key (("bibliogra* review**" and "Latent Dirichlet Allocation") or ("bibliogra* review**" and LDA) or ("bibliogra* review**" and "topic model") or ("literature* review**" and "Latent Dirichlet Allocation") or ("literature* review**" and LDA) or ("literature* review**" and "topic model"))	Mendeley, RIS, BibTeX, HTML, ciVal csv, ExLibris, EndNote
Web of Science (WoS)	"Literature review*" and LDA (topic) or "literature review*" and "Latent Dirichlet Allocation" (topic) or "literature review*" and "topic model" (topic) or "bibliogra* review**" and LDA (topic) or "bibliogra* review**" and "Latent Dirichlet Allocation" (topic) or "bibliogra* e review**" and "topic model" (topic)	EndNote, RefWorks, RIS, BibTeX, XLSX, CSV e HTML
Google scholar (Google)	["bibliographic review" AND LDA or "Latent Dirichlet Allocation" or "topic model"] OR ["bibliografic review" AND LDA or "Latent Dirichlet Allocation" or "topic model"] OR ["literatura review" AND LDA or "Latent Dirichlet Allocation" or "topic model"]	BibTexGeneric Name, CSV, EndNote e RefMan

As informações das publicações selecionadas foram baixadas e colocadas no Mendeley para seleção. A Figura 2 apresenta o diagrama PRISMA (Sarkis-Onofre et al, 2021) com as etapas de busca e seleção das publicações da revisão de 2021, sem uso do Sw3T.

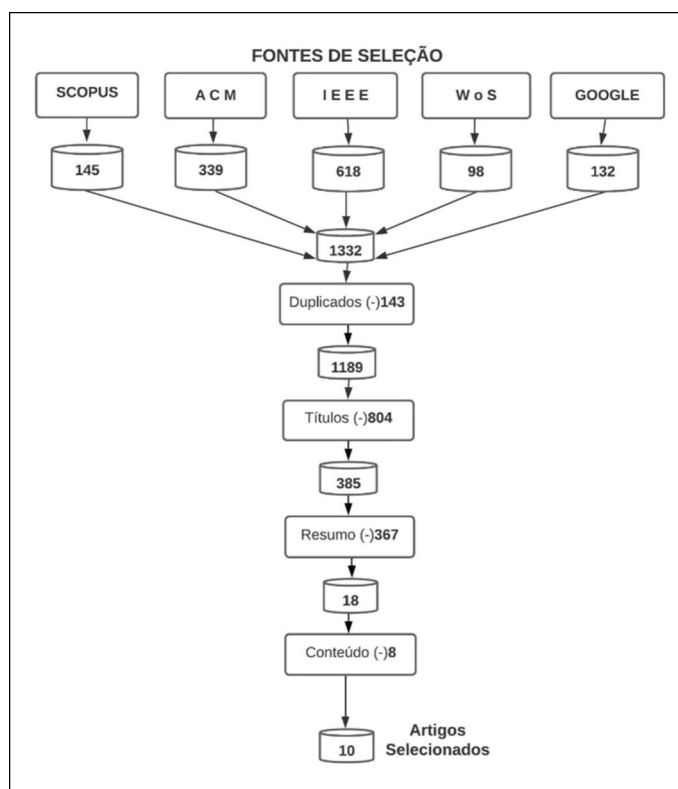


Figura 2 - Prisma - totais das buscas da RSL de 2021 e motivos de exclusão

A Figura 2, através do diagrama PRISMA observa-se que foram selecionados 1332 metadados iniciais. As operações para eliminar duplicados, por título e por resumo foram realizadas na ferramenta Mendeley. No entanto, a falta de tecnologia para agrupar resumos semelhantes, tornou o esforço para análise moroso e desgastante.

Entre as 10 publicações selecionadas, apenas 9 são apresentadas como trabalhos correlatos uma vez que uma delas se refere ao artigo publicado pela autora desta tese sobre este tema com título “*A Semantic approach to support the analysis of abstracts in a bibliographical review*” (Sergio et al, 2019) .

Em dezembro de 2022, na execução do terceiro ciclo de pesquisa-ação, uma nova RSL desta pesquisa foi realizada utilizando na íntegra o novo processo com a ferramenta Sw3T, então satisfatoriamente desenvolvida. Embora o resultado não tenha sido incorporado integralmente nesta pesquisa, algumas informações estão incluídas neste relato, como o gráfico da Figura 3, que mostra o crescimento significativo de publicações por ano nas bases de dados pesquisadas.

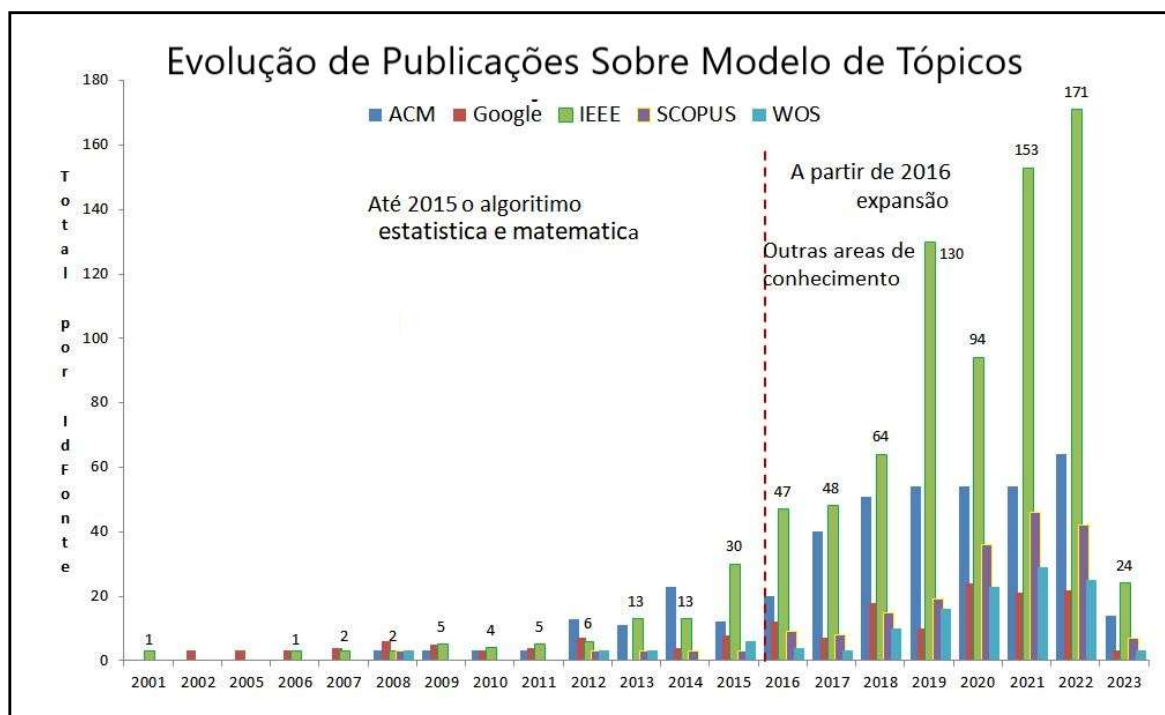


Figura 3 - Crescimento de publicações sobre o tema em até 2022 (da autora)

É esperado que as bases ACM e IEEE tenham um aumento expressivo por serem da área específica, mas todas as bases apresentam uma evolução significativa desde 2016.

Observa-se que surge uma maior quantidade de relatos de uso de modelos de tópicos em revisão literária, porém, na análise de conteúdo observou-se pouquíssimas propostas que visam o desenvolvimento de uma ferramenta com o uso dessa tecnologia. As que abordam este tipo de solução, revelam lacunas ao apontarem oportunidades de melhoria em suas soluções para prover apoio a todas as atividades de uma RSL.

2.4.1 Trabalhos correlatos

A análise mais detalhada das nove publicações selecionadas em meados de 2021 resultou em importantes contribuições para identificar lacunas no conhecimento e como direcionar a pesquisa baseada em evidências.

Na tabela 3 estão relacionadas por ano de publicação, o título e a base de dados onde as publicações foram encontradas.

Tabela 3 - Relação de publicações de trabalhos correlatos de RLS 2018 e 2022

n.	Ano/ Fonte	Autores	Título	Origem
1	2016/ Scopus	Kometani e Nagaoka	Construction of a literature review support system using LDA	Journal Lecture Notes in Artificial Intelligence
2	2017/ ACM	Ros et al.	A Machine Learning Approach for Semi-Automated Search and selection in literature studies	ACM Conference Proceeding Series
3	2018/ Scopus	Tiba et al.	Firms with benefits: A systematic review of responsible entrepreneurship and corporate social responsibility literature.	Corporate Social Responsibility and Environmental Management
4	2019/ Scopus	Asmussen e Møller	Smart literature review a practical topic modelling approach to exploratory literature review	journal Lecture Notes in Artificial Intelligence
5	2019/ Google	Chakkarwar e Tamane	Quick insight of research literature using topic modeling	Conference on Smart Innovation, Sys and Tech
6	2020/ Scopus	Lin et al.	Text mining in a literature review of urothelial cancer using topic model	BMC Cancer
7	2020/ Scopus	Tauchert et al.	Towards an Integrative Approach for Automated Literature Reviews Using Machine Learning	Journal in Artificial Intelligence and Lecture Notes
8	2021/ IEEE	Karunananda et al.	Intelligent personal research assistant	5th SLAAI-ICAI 2021 - Conference on AI and 17th Annual Sessions
9	2021/ Google	Mateen et al.	An Analysis on Text Mining Techniques for Smart Literature Review	Journal of Advanced Trends in Computer Science and Engineering

A seguir, as publicações relacionadas são resumidas evidenciando suas principais contribuições para esta pesquisa.

Kometani e Nagaoka (Kometani&Nagaoka, 2016) criam uma aplicação denominada SMS - Sistema de Gerenciamento de Seminários com uso do LDA para acompanhar alunos em sua primeira pesquisa. Relatam o uso do SMS por 10 alunos e um total de 22 artigos. Foram inseridos no SMS um total de 123 resumos, sendo 22 resumos dos próprios artigos e os demais resumos elaborados pelos alunos e professores. Utilizam uma lista de *Stopwords* atualizada pelos professores. O LDA é executado apenas com os substantivos e adjetivos dos resumos revelando tópicos. Esses tópicos foram utilizados para orientar os alunos na escolha de seus temas de pesquisa. Uma contribuição importante para esta pesquisa foi a forma de utilização dos *Stopwords* e o uso do LDA em revisão literária para identificar questões de pesquisa, que serviu de inspiração para a construção do Sw3T.

Ros et al. (Ros et al., 2017) simulam o uso de técnicas de aprendizado de máquina em um processo semiautomático de busca e seleção de publicações em BDs indexados. Propõem a automação desde a geração das strings de busca para cada BD, a realização da busca e obtenção dos artigos completos no formato PDF. Na análise, são utilizados o LDA e técnicas de tratamento de texto como *N-gram*, tokenização, lematização, *Stopwords* e IF-IDF para eliminar termos raros (Wu et al., 2008) . Embora o artigo tenha sido um dos pioneiros no uso de LDA para RSL, identificaram-se várias questões a serem equacionadas, desde a dificuldade de conversão dos artigos em PDF para texto, tendo como premissa o acesso aos textos completos e por não apresentar alternativas para eliminar publicações duplicadas. O relato da revisão não informa o processo realizado o que prejudica a transparência e a rastreabilidade do processo. Para busca, propõem, mas não aplicam, técnica “bola de neve” com base nas citações dos artigos. Esse artigo inspirou a definição de diversas funcionalidades para o Sw3T.

Tiba et al (Tiba et al, 2018) fazem uma RSL sobre empreendedorismo responsável e utilizam o algoritmo LDA para identificar tópicos sem declarar o uso de um parâmetro. Em sua conclusão, afirmam que mitigaram o tema usando os resultados na restrição do foco, mas como o LDA, um método baseado em máquina, é restrito em sua capacidade de explicar nuances de significado dentro de um texto, entenderam não poderiam esperar igualar o resultado à análise de conteúdo de um especialista em termos de riqueza de resultados. Assim, optaram por complementar o resultado do LDA com uma análise quantitativa utilizando ferramentas de bibliometria. Esse artigo colabora com esta pesquisa no sentido de apontar a necessidade de permitir diversas análises de LDA e prever consultas aos resumos das publicações em análise visando rever decisões de exclusão/inclusão, assim como de prever interface com outras ferramentas usadas em uma RSL como as de bibliometria.

As demais publicações foram identificadas na revisão de 2021, durante a execução do ciclo 2 da pesquisa-ação. Observou-se o crescimento significativo de publicações sobre uso do LDA em RSL.

Asmussen e Møller (Asmussen&Møller, 2019) propõem o uso do LDA para analisar PDF de artigos inteiros, convertidos em texto e usando uma limpeza interativa que, segundo os autores, se apresenta mais como arte do que ciência. O LDA é usado mais para utilidade do que para precisão. Tem como critério eliminar termos que ocorrem em um único artigo. Usa a métrica de perplexidade para orientar na quantidade de tópicos a ser adotada. Segundo os autores, a quantidade de tópicos necessária pode variar em função do tipo de revisão. A revisão exploratória necessita de um menor número de tópicos e a revisão completa de um maior número. Eles rotulam e identificam os assuntos de cada tópico. A pesquisa é validada pela semântica que compara os resultados com o raciocínio de especialistas. Os artigos distribuídos em tópicos são aferidos se estão nos tópicos certos ou deslocados. O resultado é válido se a maioria estiver no tópico correto. Uma planilha é utilizada para visualizar os títulos e as palavras-chave de cada tópico. Este trabalho possui como pontos fortes o uso do indicador de perplexidade e a rotulagem de tópicos e apresenta como fragilidade n o uso do texto completo da publicação desde o início da pesquisa.

Chakkarwar e Tamane (Chakkarwar&Tamane, 2019). apresentam um caso concreto do uso do LDA na análise de diversos tipos de documentos, mas não de publicações científicas. Discutem as tecnologias utilizadas para processar os textos e gerar os tópicos, sem abordar os parâmetros do LDA utilizado. O método termina com a identificação de rótulos dos tópicos gerados. Os pontos fortes deste trabalho estão na descrição conceitual abrangente valorizando a escolha do LDA e no uso de um processo completo de análise de texto gerando tópicos.

Lin et al (Lin et al, 2020) fazem uma RSL sobre Câncer Uterino, utilizam o LDA2VEC um método de mineração de texto não supervisionado, determinam o número crítico de tópicos. No entanto, afirmam que não há uma melhor maneira de escolher o número ideal de tópicos. Usam a medida de perplexidade para estimar o número ótimo de tópicos, mas colocam que o número ideal de tópicos geralmente é dado pelos pesquisadores. Este trabalho contribui na aplicação do LDA2VEC e no entendimento do uso da métrica perplexidade.

Tauchert et al (Tauchert et al, 2020) automatizam parcialmente o processo de revisão da literatura. Desde a coleta de artigos até a avaliação. Usam rastreador

personalizado não autorizado pela mantenedora da base de dados, o algoritmo word2vector, modelagem de tópicos LDA, extração automática rápida de palavras-chave e clustering hierárquico aglomerativo, processamento e agrupamento automáticos de literatura relevante e subsequente apresentação gráfica dos resultados usando ilustrações como dendrogramas. Fornecem informações sobre quais tópicos cada cluster aborda e quais palavras-chave os definem. Analisam por citação, texto ou híbrido e baseiam-se em texto de resumo, palavras-chave ou texto completo que entendem ter resultados superiores para categorização de documentos. O objetivo de classificação, recomendação, extração de conteúdo, agrupamento e técnica utilizada são bag-of-words, vetorização, classificador bayesiano, modelos de tópicos, extração de palavras-chave. Esta publicação contribui no estabelecimento de clusters, uso de técnicas de NPL e uso de word2vector com LDA.

Karunananda et al (Karunananda et al, 2021) apresentam a solução InPRA – Assistente de Pesquisa Pessoal Inteligente para apoiar todo processo de pesquisa científica. Tem como foco subsidiar alunos no desenvolvimento de pesquisa. O InPRA recomenda leituras, ajuda na definição da questão de pesquisa, formula a metodologia e controla toda a execução até a conclusão da pesquisa. Restringe-se a buscar nas BDs públicas Arxiv da *Cornell University* que disponibilizam artigos de várias áreas, mas não fazem revisão por pares. Permite a conexão com o Mendeley, EndNote, Zotero, Ms-word e verificador gramatical (*Grammarly*). Descrevem as tecnologias utilizadas e deixam claro que o software conduz toda a pesquisa obrigando o aluno a executar as atividades na sequência pré-estabelecida e restringindo o acesso a módulos posteriores. Nota-se que mesmo a solução sendo orientada a pesquisador iniciante, pode não ser desejado o rigor que ela impõe. Outra restrição é a BD única que ela acessa.

Mateen et al. (Mateen et al, 2021) trazem uma revisão de trabalhos relacionados à análise da literatura que utilizam mineração de texto e apontam as técnicas de análise de frequência e a modelagem de tópicos como sendo as mais utilizadas. Concluem usando como critérios de avaliação o indicador de coerência de tópicos, indicador de perplexidade e medida qualitativa, entre outros. Encontram semelhança semântica entre as palavras de um tópico usando o indicador de Coerência. Confirmam que o LDA é a técnica mais utilizada com o modelo *bag of*

words que não considera a ordem ou o contexto das palavras e indicam o uso da *word2vector* (Silva, 2021) . Para análise de frequência de palavras, indicam a técnica *POS tagging*. Alguns estudos fornecem evolução de tópicos para um período específico de tempo e tendências em relação a países e autores. Citam o modelo de *tópico STM* para análise da literatura e indicam seu estudo para melhorar o desempenho e comparação com as demais. Apontam o *t-SNE* como um bom método para visualizar documentos pela distribuição de tópicos. A contribuição desse trabalho é a indicação das melhores técnicas de mineração. Falta informação sobre o modelo STM e o t-SNE para visualizar tópicos.

No final de 2022, uma nova RSL sobre este tema foi realizada durante a execução do Ciclo 3 da pesquisa-ação utilizando o novo processo de RSL com Sw3T. Nessa oportunidade foram encontradas mais duas publicações de 2022 que trazem novidades sobre o uso do LDA em RSL. Assim, mesmo não tendo sido incorporadas nesta pesquisa, elas são citadas a seguir e inspiram trabalhos futuros.

Facchinetti et al. (Facchinetti et al, 2022) encontrada na Scopus e publicada no Journal Knowledge-Based Systems, apresentam um fluxo de trabalho de RSL semiautomática. Consideram o uso de um conjunto de ferramentas já existentes, com funções específicas incorporadas no fluxo RSL semiautomática denominado SRL-kit. Sua contribuição para o SRL-kit é o desenvolvimento da ferramenta chamada *FAWOC – FastWord*. *FAWOC*, programa que apoia a classificação de tokens em: rótulos (palavras-chave) ou ruídos (não relevante, *Stopwords* ou adiada). As *Stopwords* são colocadas nesta última classificação. O corpus com a classificação dos Tokens atualizada é submetido ao LDA. O resultado do processo consiste em um conjunto de estatísticas da relação entre artigos, tópicos e sua tendência de publicação em periódicos e anais de conferências. A precisão é avaliada pelo exame dos resultados por especialistas do domínio. A contribuição desse trabalho é a classificação semiautomática de tokens, conversão de siglas e a análise de subgrupos de resumos.

Tanhaei et al. (Tanhaei et al, 2022) retratam o estudo de neurocientistas que para encontrar a relação entre as regiões anatômicas do cérebro e as doenças que as afetam utilizaram os tópicos gerados pelo LDA. Analisaram um grande

conjunto de artigos para encontrar os termos comuns em pelo menos dois tópicos. Alcançaram um melhor resultado do que se a análise fosse realizada por publicações. Utilizaram realidade virtual para exibir a correlação entre as publicações. O trabalho mostra que é possível utilizar o LDA com técnicas de realidade aumentada representando em 3D a correlação entre os tópicos.

A partir da leitura das publicações selecionadas nas duas RSL, foram identificadas melhores práticas e lacunas na aplicação de modelagem de tópicos em revisões literárias. A seguir, esses achados são relacionados em forma de questões específicas contemplados nos pontos de análise, conforme definido na estrutura da pesquisa a ser descrita no capítulo 3, subseção 3.5. São eles:

- [1] Criar arquivo de *Stopwords* que contemple as necessidades de exclusão identificadas pelo usuário.
- [2] Criar solução para trocar/corrigir termos similares ou para correção de grafia.
- [3] Definir o algoritmo de modelagem a ser utilizado.
- [4] Estabelecer quais classificações gramaticais de palavras devem ser preservadas para análise.
- [5] Criar soluções que não restrinjam pesquisadores com dificuldade de acesso a BDs indexadas e/ou que não tenham facilidade tecnológica e nem de diferentes áreas do conhecimento.
- [6] Criar uma solução para validar a consistência dos metadados selecionados e baixados das BDs quanto à completude necessária para a seleção.
- [7] Permitir o tratamento de artigos de várias fontes ao mesmo tempo mantendo seu histórico de origem.
- [8] Permitir a eliminação de artigos duplicados.
- [9] Preservar as características de uma RSL de: transparência, confiabilidade, rapidez e permitir produtividade do processo.
- [10] Contemplar funções de registro de informações da RSL e resgate do histórico das decisões do pesquisador durante a seleção para subsidiar o relatório.
- [11] Possibilitar a autonomia do pesquisador no que se refere à escolha das BDs e na condução de sua pesquisa.
- [12] Prover uma solução que possa ser utilizada por pesquisadores de qualquer área do conhecimento e independente de sua habilidade com tecnologia.

- [13] Adiar até a análise do conteúdo do artigo para reduzir a quantidade de PDFs a serem baixados.
- [14] Possibilitar a integração da ferramenta com outras ferramentas que atendam de forma específica o RSL, como gráficos.
- [15] Permitir a escolha dos parâmetros na execução dos modelos de tópico, dando autonomia na decisão do melhor resultado obtido.
- [16] Definir a métrica de validação do resultado dos tópicos.
- [17] Viabilizar a análise de grande quantidade de resumos, especialmente em relação ao ano de publicação.
- [18] Permitir o compartilhamento da RSL entre pesquisadores envolvidos.

Com a descrição das publicações relacionadas e a relação de achados identificados na RSL, conclui-se esta atividade da Pesquisa Científica. A continuidade do trabalho, isto é o estabelecimento do quadro de métodos da pesquisa e a definição da sua estrutura é desenvolvida no capítulo 3.

No entanto, na próxima subseção são relacionadas ferramentas que não apresentam evidências do uso de modelagem de tópicos, mas que apoiam revisões literárias.

2.4.2 Ferramentas que apoiam uma revisão literária

A RSL desta pesquisa, através da escolha dos termos de busca, visou encontrar publicações que abordassem o uso de ferramentas em RSL que utilizassem modelagem de tópicos. Porém, observando as publicações selecionadas, apenas algumas poucas publicações foram encontradas abordando o uso do LDA de forma insatisfatória ou incompleta. Tal fato motivou a ampliação da busca por outras ferramentas de apoio à revisão literária a fim de identificar boas práticas a serem adotadas e estabelecer possíveis interfaces que propiciem o fornecimento de informações mais completas no relato de uma RSL.

Utilizando o agrupamento proposto por Vosgerau e Romanowski (Vosgerau&Romanowski, 2014) conceituadas no início deste capítulo, os tipos de revisão podem ser classificadas em “revisões que mapeiam” e “revisões que avaliam e sintetizam”. A seguir, relaciona-se as ferramentas:

As “revisões que mapeiam” catalogam o material como planilhas eletrônicas, EndNote e o Mendeley. Apoiam a produção de análise de conteúdo como o software Atlas.ti. Viabilizam estudos bibliométricos e análise quantitativa, como o software Pajek. A apresentação de seus resultados é usualmente na forma textual, com uso de métodos estatísticos e com apresentação gráfica que ilustra e comprova os quantitativos obtidos, como os softwares Sphinx, NVivo, SPSS, Pajek.

As “revisões que avaliam e sintetizam” catalogam ferramentas como NVivo e Atlas.ti que permitem a recuperação rápida de extratos codificados no texto, a criação de categorias e subcategorias de análise, a confecção de redes de relação entre os códigos, categorias e extratos do texto e ainda a exportação para outros softwares de tratamento quantitativos, quando necessário. Para a contabilização das variáveis quantitativas levantadas e analisadas pelos estudos e aplicados a estes testes estatísticos, apontam o software SPSS.

Outras ferramentas encontradas, não citadas por Vosgerau e Romanowski (Vosgerau&Romanowski, 2014) são: *More*, *Citavi*, *citeUlike*, *Evernote*, *Minhas Citações*, *Zotero*, *Publions*, *Citespace*, *VosViewer* e outros. Essas ferramentas realizam análise por autores, por local de publicação, por país, permitem manter lista de publicações e provêm referências automatizadas. Algumas têm função de busca de artigos similares na WEB. Os softwares *StArt*, *JabRef* e *Parsifal* são os mais referenciados no Google. A *StArt* e *JabRef* apresentam características funcionais que atendem uma RSL, porém, não foi possível identificar se essas ferramentas utilizam modelagem de tópicos.

Portanto, desenvolver uma solução de RSL com o Sw3T se apresenta como uma importante contribuição para a ciência e durante seu desenvolvimento incorporou algumas funcionalidades desses aplicativos à medida que se faziam necessárias.

2.5 CONSIDERAÇÕES FINAIS

Neste capítulo foi identificado o estado da arte do processo de revisão literária, identificadas as lacunas de pesquisa e apontados os requisitos para o desenvolvimento de um novo processo. O próximo capítulo descreve aspectos de processamento da linguagem natural (PLN).

3. METODOLOGIA – ESTRUTURAÇÃO DA PESQUISA

3.1 INTRODUÇÃO

Este capítulo tem o objetivo de descrever o quadro metodológico adotado nesta pesquisa e está organizado da seguinte forma: conceitos sobre metodologia, descrição de métodos de pesquisa disponíveis, estruturação das etapas da pesquisa e a descrição da etapa de planejamento.

3.2 METODOLOGIA DE PESQUISA E MÉTODOS EXISTENTES

O desenvolvimento de um trabalho científico deve contar com um quadro metodológico que garanta seu resultado. Para o entendimento e justificação do quadro metodológico adotado, apresentam-se os métodos disponíveis para composição do quadro de uma pesquisa científica.

GIL (Gil, 2006) define **pesquisa** como sendo “*o procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos*”. Por sua vez Reinehr (Reinehr, 2008), define o termo **pesquisa científica** como:

“É a aplicação de um determinado conjunto de métodos, adequado às circunstâncias e objeto da investigação, que levará ao caminho que representa a opção mais segura e econômica para a pesquisa, auxiliando o pesquisador na definição das melhores opções a serem adotadas ... para o aproveitamento de experiências e conhecimentos acumulados acerca do próprio uso do método, contribuindo para aumentar a eficiência da investigação científica”.

A pesquisa científica é um processo composto por diferentes etapas interligadas (Kitchenham&Brereton, 2013). Ela é concebida como um processo de múltiplas etapas relacionadas entre si, que acontecem ou não de maneira sequencial ou contínua.

Reinehr (Reinehr, 2008) define o termo **metodologia** como “filosofia da ciência, ou seja, aquela que estuda os métodos, e, ainda, analisa e critica a sua validade em situações diversas de aplicação. De forma resumida, pode-se compreender que o termo metodologia se refere ao estudo dos métodos e não constitui propriamente um sinônimo para o termo método”.

Por sua vez, Lakatos e Marconi (Lakatos&Marconi, 2006) definem o termo **método**, como sendo: “(...) conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar conhecimentos válidos e verdadeiros, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista”.

Como a pesquisa é um processo de investigação sistemática de um assunto, um quadro metodológico deve ser empregado.

A compreensão dos diferentes métodos disponíveis, possibilita o entendimento de como se definiu o quadro metodológico desta tese selecionando os métodos adequados a serem adotados no desenvolvimento da pesquisa.

3.3 CARACTERIZAÇÃO DA PESQUISA

Nesta seção, serão apresentadas diversas abordagens para a classificação de pesquisas de forma a levar à caracterização desta pesquisa específica.

3.3.1 Quanto aos critérios de caracterização

SANTOS (Santos, 1999) afirma que os três **critérios** que caracterizam a pesquisa científica são: [i] objetivos da pesquisa, [II] procedimentos de coleta de dados e [III] fontes da coleta de dados.

Quanto aos **objetivos**, uma pesquisa pode ser do tipo exploratória, descritiva ou explicativa (Santos, 1999) (Gil, 2006) .

Pesquisa do **tipo exploratória** visa proporcionar maior familiaridade com o problema, torná-lo mais explícito ou construir hipóteses (Gil, 2006) . Para Santos (Santos, 1999), consiste em um passo preliminar para outro tipo de pesquisa, ajudando a expor e delimitar o tema e hipóteses.

Pesquisa do **tipo descritiva**, descreve as características de determinada população ou fenômeno (Gil, 2006) , ou seja, “*é um levantamento das características conhecidas, componentes do fato/fenômeno/problema*” (Santos, 1999).

Pesquisa do **tipo explicativa** busca criar uma teoria aceitável a respeito de um fato ou fenômeno (Santos, 1999) possibilitando a compreensão da relação

causal entre variáveis e explicando o porquê da ocorrência. Em geral aprofunda o conhecimento sobre a realidade (Gil, 2006)

*Esta pesquisa, quanto ao seu objetivo, se caracteriza como sendo do tipo **exploratória** pois proporciona uma maior compreensão do processo de revisão literária com vistas a melhorar a execução desta atividade.*

Quanto aos **procedimentos de coleta de dados** para alguns tipos de pesquisa, Gil (Gil, 2006) apresenta dois grandes tipos de procedimentos, aos que utilizam fontes de papel e os de campo, cujos dados são fornecidos por pessoas.

Os procedimentos de coleta de dados que utilizam **fontes de papel**, se dividem em pesquisa bibliográfica e documental. A **pesquisa bibliográfica** é desenvolvida com base em materiais publicados, como livros e artigos científicos e trata exclusivamente do levantamento das fontes escritas (Gil, 2006) enquanto a **pesquisa documental** trata de papéis que ainda não foram organizados ou não receberam tratamento analítico e nem publicações (Santos, 1999).

Os procedimentos de coleta de dados em **pesquisas de campo** se subdividem em pesquisa experimental, ex-post facto, levantamento de campo, estudo de caso, pesquisa-ação e pesquisa participante. Para Santos (Santos, 1999) as duas últimas também se situam nesse tipo de coleta de dados.

A pesquisa de campo **experimental** consiste na reprodução, sob condições controladas, de um determinado fato ou fenômeno, com o objetivo de descobrir os fatores procurados. Gil (Gil, 2006) conclui que tratando-se de pessoas, grupos ou organizações, este entendimento não é tão simples de se delinear como quando se trata de entidades físicas como cobaias e elementos químicos.

A pesquisa de campo **ex-post facto** é quando o estudo é realizado após os acontecimentos. Não dispondo de controle sobre a variável independente do fenômeno, porque ele já ocorreu (Gil, 2006)

A pesquisa **levantamento de campo** (ou **survey**) caracteriza-se pela interrogação direta às pessoas cujo comportamento se deseja conhecer, para depois inferir conclusões com análises quantitativas. Segundo Forza (Forza, 2002), o uso de *surveys* teve um crescimento acentuado nas décadas que o antecedem, pois,

tornou as pesquisas mais úteis no campo prático uma vez que conjugam dados empíricos.

A pesquisa de campo **estudo de caso** caracteriza-se, segundo GIL (Gil, 2006) , pelo estudo profundo e exaustivo de um ou poucos objetos, com o objetivo de adquirir amplo e detalhado conhecimento sobre eles. Yin (Yin, 2005) define estudo de caso como sendo “Uma investigação empírica que investiga um fenômeno contemporâneo dentro de seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos”.

Na pesquisa de campo, segundo Coughlan e Coghlan (Coughlan&Coghlan, 2002) a **pesquisa-ação** é caracterizada por focar na condução do pesquisador que atua para resolver problemas organizacionais junto com aqueles que vivem o problema diretamente. Entende que a pesquisa-ação utiliza um processo emergente, cuja forma é se delineia à medida em que desenrola a ação e depende diretamente do resultado desta ação para adquirir sua forma.

A pesquisa de campo **pesquisa participante**, por sua vez, caracteriza-se pela interação entre pesquisadores e membros da situação investigada, assim como a pesquisa-ação (Gil, 2006) . Mas enquanto a pesquisa-ação pressupõe algum tipo de ação planejada, a pesquisa participante visa distinguir a ciência popular da ciência dominante (Gil, 2006) . Para Thiollent (Thiollent, 2018) toda pesquisa-ação é do tipo participante, pois o pesquisador participa da ação. No entanto, nem toda a pesquisa participante pode ser considerada uma pesquisa-ação.

Esta pesquisa, quanto aos procedimentos e fontes de coleta de dados, caracteriza-se como sendo do tipo pesquisa-ação pois é uma atividade que vai ser desenvolvida em campo e, além disso, o pesquisador vai vivenciar o problema e interferir na sua melhoria em um processo que se desenvolve em ciclos.

3.3.2 Quanto à natureza das variáveis

Para Miguel (Miguel, 2011) quanto à natureza das variáveis, a pesquisa pode ser qualitativa ou quantitativa. Segundo Bryman (Bryman, 2006), a pesquisa qualitativa busca uma profunda compreensão do contexto de uma situação e procura analisar a sequência de fatos no tempo, com um enfoque menos estruturado. Por outro lado, a pesquisa quantitativa utiliza modelagens matemáticas, estatísticas ou

simulação e normalmente é centrada em laboratório. Miguel (Miguel et al, 2010) ainda afirmam que a pesquisa-ação, a pesquisa participante e o estudo de caso são classificados como pesquisas qualitativas.

*Esta pesquisa, quanto à natureza, é do tipo **qualitativa** pois deverá desenvolver um processo de compreensão do fenômeno de uma revisão literária.*

3.3.3 Quanto à estratégia da pesquisa

Ainda com relação à **classificação da pesquisa**, Yin (Yin, 2005) contesta a visão hierarquizada de alguns metodologistas que afirmam que determinados métodos são apenas aplicáveis a determinados tipos de pesquisa. Apresenta três **critérios** para classificar a estratégia de pesquisa, sendo:

- (i) Tipo de questão da pesquisa;
- (ii) Controle sobre os eventos comportamentais efetivos; e
- (iii) Foco em fenômenos históricos ou fenômenos contemporâneos.

A Tabela 4 ilustra, de acordo com estes critérios, a estratégia de pesquisa a se adotada, segundo Yin (Yin, 2005) .

Tabela 4 - Estratégia de pesquisa, adaptado de Yin (Yin, 2005)

Estratégia da pesquisa	Forma de questão de pesquisa	Há controle sobre eventos comportamentais?	Foca em acontecimentos contemporâneos?
experimento	como, por que	sim	sim
levantamento	quem, o que, onde, quantos, quais	não	sim
análise de arquivo	quem, o que, onde, quando, quais	não	sim/não
pesquisa historica	Como, por que	não	não
estudo de caso	Como, por que	não	sim

Para Yin (Yin, 2005) , questões de pesquisa do tipo “**o que**” tendem a ser mais exploratórias e que qualquer uma das estratégias de pesquisa poderia ser adotada. No entanto, questões do tipo “**o que**” relacionas a questões do tipo “**quais**”, se aplicam melhor em um **levantamento** ou em uma **análise de arquivos**.

Explica que o estudo de caso é utilizado quando: “faz-se uma questão do tipo “**como**” e “**por que**” sobre um conjunto contemporâneo de acontecimentos, sobre

o qual o pesquisador tem pouco ou nenhum controle” (Carrara, 2011) . A principal diferença entre o **estudo de caso** e um **experimento** é que o segundo é isolado de seu contexto, uma vez que ocorre no ambiente controlado de laboratório. Já o estudo de caso difere da **pesquisa histórica** uma vez que a mesma se baseia em fatos não-contemporâneos.

Conclui-se a análise da Tabela 4 afirmando que é natural inferir que uma pesquisa no formato de **experimento** se dá quando ela exige, por parte do pesquisador, controle sobre eventos comportamentais. Ao contrário da **pesquisa histórica** e **estudo de caso** que não permitem qualquer controle dessa natureza.

Para sua definição, esta pesquisa corrobora com a afirmação de Bechhofer (Bechhofer, 2004) “... o processo de pesquisa não é uma sequência bem definida de procedimentos que seguem um padrão organizado, mas uma iteração desorganizada entre o mundo conceitual e prático, de dedução e indução ocorrem ao mesmo tempo”. Reinehr (Reinehr, 2008) também afirma:

“Uma pesquisa não se encaixa de forma precisa em uma ou outra tipologia. Em especial, porque o processo de construção do conhecimento não se dá de forma linear ao longo do tempo, mas sim, de forma iterativa e incremental. Nesse processo, mesclam-se métodos e ferramentas de acordo com a etapa da pesquisa, as características do objeto investigado e a evolução da compreensão do pesquisador acerca do fenômeno estudado (Reinehr, 2008)”.

Esta pesquisa, quanto à sua estratégia é do tipo experimento pois o tipo de questão de pesquisa é identificar como técnicas de Processamento da Linguagem Natural (PLN) podem contribuir com a revisão literária, o pesquisador possui controle sobre os eventos comportamentais e foca em acontecimentos contemporâneos.

3.3.4 Fechamento da caracterização

Com base no explanado este trabalho apresenta o seguinte quadro metodológico:

- Critérios de caracterização: **pesquisa exploratória**
- Procedimentos e fontes de coleta de dados: **pesquisa-ação**
- Natureza das variáveis: **pesquisa qualitativa**
- Estratégia de pesquisa: **experimento**

3.4 CONCEITOS DE PESQUISA-AÇÃO

Para Wood-Harper (Wood-Harper, 1995) a pesquisa-ação é percebida como a mais efetiva para o desenvolvimento de técnicas ou construção de teoria. Nela os motivos são mais bem observados, uma vez que o pesquisador atua na mudança do ambiente estudado.

A pesquisa-ação busca a forma participativa onde pesquisadores e participantes são envolvidos (Thiollent, 2018) . Ela valoriza a liberdade dada aos pesquisadores na busca pelo aperfeiçoamento do processo e na forte interação entre os membros do projeto a fim de se obter uma reflexão coletiva (Carrara, 2011) .

Coghlan e Coghlan e Brannick (Coghlan&Brannick, 2005) afirmam que é o método apropriado quando a questão de pesquisa está relacionada à descrição de uma série de eventos no tempo ocorrido em determinada iniciativa, buscando responder perguntas ‘**como**’ e ‘**por que**’ em relação às ações associadas à melhoria da solução e ao entendimento do processo estudado. Carrara (Carrara, 2011) descreve as dez principais **características da pesquisa-ação**, que são:

- 1) O pesquisador participa da ação (não é somente um observador).
- 2) Contempla dois objetivos: contribuir para a ciência (objetivo de conhecimento) e solucionar um problema (objetivo prático).
- 3) Ocorre interatividade e cooperação entre os envolvidos.
- 4) Desenvolve meta com entendimento holístico.
- 5) Está relacionada à mudança.
- 6) Necessita de entendimento de valores e normas (estrutura ética).
- 7) Pode incluir todos os tipos de métodos de coleta de dados (técnicas quantitativas e qualitativas).
- 8) Necessita de vasto entendimento prévio do ambiente, das condições, da estrutura e da dinâmica das operações.
- 9) Conduzida em tempo real.
- 10) Precisa de critérios próprios de qualidade para sua avaliação.

As três etapas da **condução de uma pesquisa-ação** para Coughlan e Coughlan (Coughlan&Coughlan, 2002), são:

- 1) **Passo prévio:** Etapa onde se entende o contexto e o propósito da pesquisa (o que e o como mudar e qual a escala de tempo – exploração teórica).
- 2) **Ciclo da pesquisa: Etapa** onde realiza-se seis passos:
 - (i) Observa os dados das ações do ciclo anterior. No primeiro ciclo são observados os coletados na fase passo prévio.
 - (ii) Feedback da equipe sobre os dados coletados e apresentados.
 - (iii) Analisa dos dados (em conjunto com a equipe).
 - (iv) Planeja da ação (quem faz o quê e quando).
 - (v) Executa da ação.
 - (vi) Avalia ação e realiza coleta de dados.

A Figura 4 apresenta o ciclo de pesquisa destacando a sequência de execução dos movimentos.

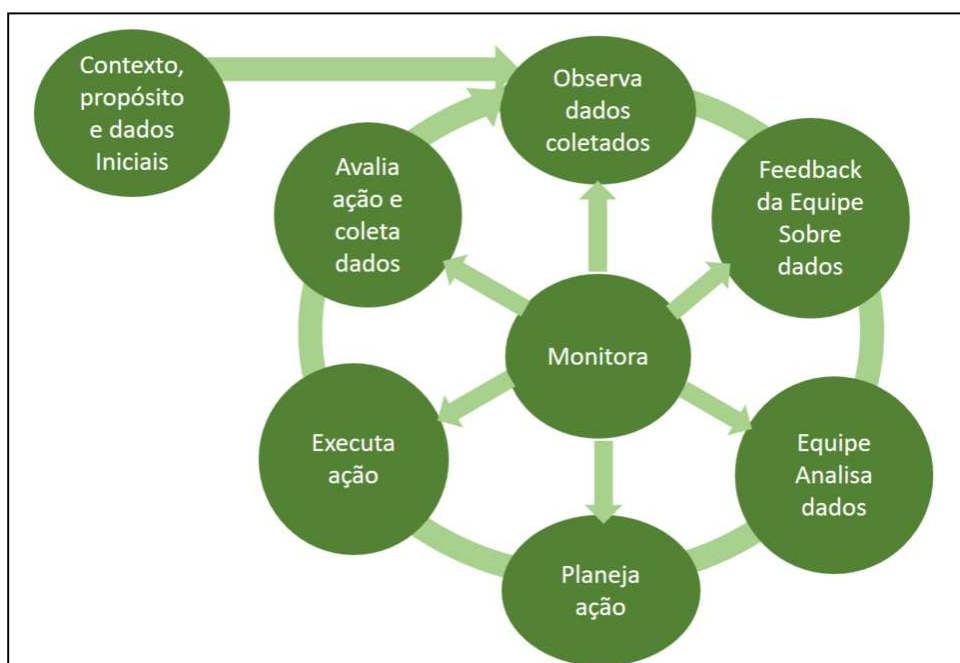


Figura 4 - Passos do Ciclo da pesquisa (adapt. de (Coughlan&Coughlan, 2002)

- 3) **Monitorar o processo** (Metapasso): Etapa de elaboração do relatório da pesquisa, descreve o observado no monitoramento dos passos anteriores e observa o processo de aprendizagem.

Para Carrara (Carrara, 2011) , apud Franco (Franco, 2005), a pesquisa-ação pode ser de três tipos:

- 1) **Colaborativa**: o pesquisador faz parte de um processo de mudança desencadeado pelos sujeitos do objeto de pesquisa.
- 2) **Crítica**: quando a necessidade de mudança é percebida valorizando a construção cognitiva da experiência.
- 3) **Estratégica**: quando a ação é previamente planejada sem a participação dos sujeitos.

Para o autor, a pesquisa-ação crítica é a mais apontada uma vez que é a que mais se aproxima das origens da pesquisa-ação.

Tem-se que este trabalho de pesquisa é de pesquisa-ação crítica pois visa atender a necessidade de mudança do processo de Revisão Literária em trabalhos científicos com a perspectiva de valorizar a experiência desta construção.

Para Carrara (Carrara, 2011) , apud Westbrook (Westbrook, 1995), há outros **benefícios no uso da pesquisa-ação** como a validação da pesquisa com maiores oportunidades para correção de desvios e por fim, o considerado mais importante, a equipe dar o ritmo da pesquisa considerando as novas evidências possam vir a serem estudadas motivando uma nova implementação. Assim, o alto grau de interação que o método provê entre pesquisadores e membros da equipe permite insights para ajustes na sua condução.

Na elaboração da estrutura da pesquisa deste trabalho, são consideradas as duas abordagens, as dez principais características da pesquisa-ação de Coghlan e Brannick (Coghlan&Brannick, 2005) e as três fases de condução de uma pesquisa-ação apresentadas por Coughlan e Coughlan (Coughlan&Coughlan, 2002).

Westbrook (Westbrook, 1995) alerta que a subjetividade é principal ponto fraco da pesquisa-ação, o que pode ser mitigado ao agregar mais sujeitos na equipe, sejam outros pesquisadores ou outros membros do grupo. Essa subjetividade fica equacionada neste trabalho, uma vez que a pesquisa teve a colaboração de bolsistas e da pesquisadora.

Ao considerar as dez características do método de pesquisa-ação de Coughlan e Brannick (Coughlan&Brannick, 2005), tem-se que a escolha do método é consistente uma vez que todas são observadas na execução do projeto, nas etapas de planejamento (Passo prévio) e execução da pesquisa (ciclo de pesquisa), conforme demonstrado na Figura 5.

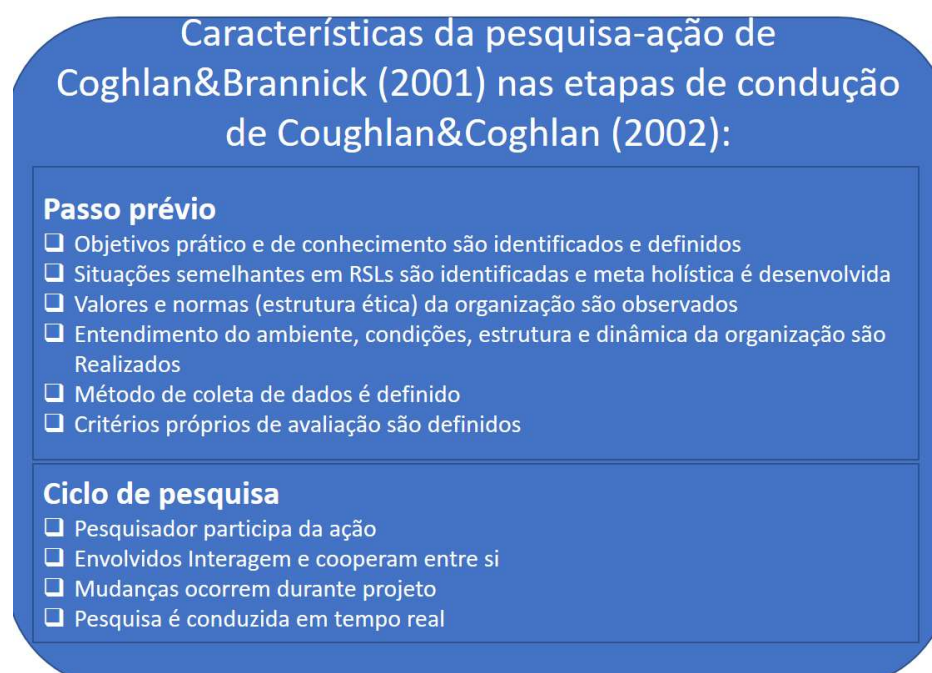


Figura 5 - Características das etapas de condução

A seguir é apresentada a estrutura da pesquisa, etapas de condução e entregáveis de cada fase.

3.5 ESTRUTURA DA PESQUISA

Esta seção descreve a estrutura da pesquisa tendo como base as etapas de condução e características definidas na subseção 3.3.4 Fechamento da caracterização da pesquisa-ação. Ela foi desenvolvida a partir do objetivo geral, dos objetivos específicos e da questão de pesquisa. A estrutura da pesquisa está

organizada em cinco fases distribuídas em três etapas a saber: Planejamento da pesquisa, ciclo da pesquisa e apresentação dos resultados, conforme ilustração apresentada na Figura 6 onde se destaca a sequência de execução.

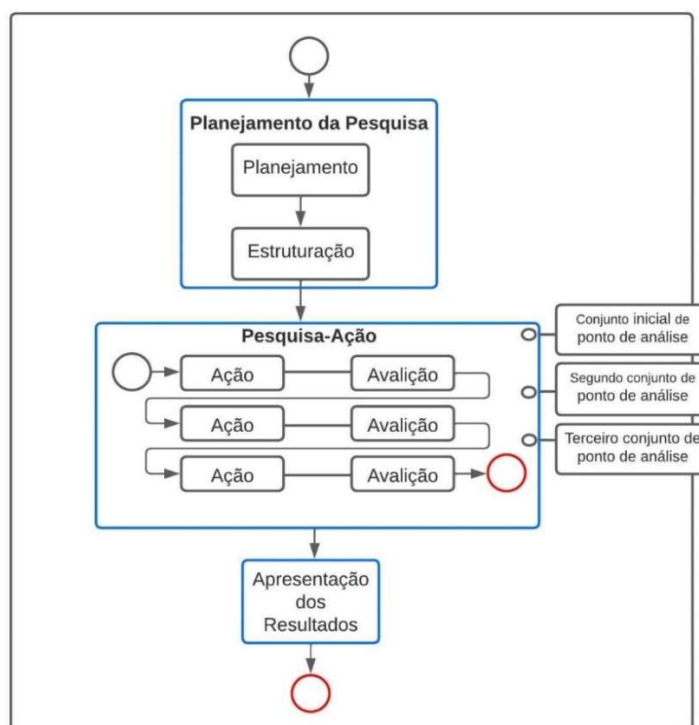


Figura 6 - Estrutura da pesquisa - etapas e fases

A Tabela 5 descreve as etapas, respectivas fases com seus entregáveis - pontos de análise para a mudança das fases.

Tabela 5 - Descrição das fases da pesquisa e respectivos entregáveis

ETAPA DA PESQUISA		DESCRIÇÃO	ENTREGÁVEIS
Planejamento da Pesquisa - exploratória	Fase 1 - Posicionamento teórico e metodológico	Realiza a RSL e dá o embasamento teórico do trabalho pela Identificação questão de pesquisa, objetivos geral, específicos e proposições.	RSL da pesquisa Questões e Proposições da pesquisa.
	Fase 2 - Estruturação da pesquisa	Desenvolve arquitetura da pesquisa e constrói o protocolo detalhando ciclos de pesquisa-ação e entregáveis assim como os pontos de análise iniciais.	Pontos de análise iniciais
Ciclo da Pesquisa-ação	Fase 3 - Ação	Desenvolvimento do novo processo de RSL, construção da ferramenta e identificação de novos pontos de análise	Ações relacionadas aos pontos de análise
	Fase 4 - Avaliação	Análise dos resultados da Fase 3 Identificação de novos elementos a serem tratados na solução.	Relatório do Ciclo da Pesquisa-ação
Fase 5 - Apresentação de resultados		Apresentação de resultados e a entrega do documento formal de pesquisa.	Relato da pesquisa

Nas próximas subseções, são apresentadas a Fase 1 e Fase 2 que compõem a primeira etapa da pesquisa descrevendo a organização, os resultados dessas fases e os entregáveis. A Etapa 2 que abrange as Fase 3 e Fase 4, assim como a Etapa 3, com a Fase 5, têm suas organizações e resultados apresentados, respectivamente, no Capítulo 6 – Aplicação do Processo de RSL com SW3T e no Capítulo 7 – AVALIAÇÃO DO PROCESSO RSL COM SW3T.

3.5.1 Etapa de planejamento da pesquisa

Esta etapa só ocorre uma única vez e está dividida em:

- Fase 1 — Posicionamento teórico e metodológico;
Os entregáveis são: Revisão Sistemática Literária da pesquisa; Questões da pesquisa; Objetivo geral, Objetivos específicos e Proposições da pesquisa.
- Fase 2 — Estruturação da pesquisa
Os entregáveis são: os pontos de análise iniciais estabelecidos com base nos resultados da RSL.

3.5.1.1 Fase 1 — Posicionamento Teórico

Esta fase se propõe a entregar o relato da Revisão Literária; a Questão de pesquisa, objetivo geral e objetivos específicos e é concluída com a definição das proposições de pesquisa, sendo:

Objetivo geral da pesquisa apresentado no Capítulo 1, é:

Desenvolver um processo de revisão literária que utilize técnicas de PLN com base no método revisão sistemática da literatura (RSL) para engenharia de software proposta por Kitchenham e Brereton (Kitchenham&Brereton, 2013).

Assim como o **Objetivos específicos** relacionados na Tabela 6:

Tabela 6 - Relação dos Objetivos Específicos e suas siglas

Objetivos Específicos	Descrição
OE1	Identificar soluções de RSL para preencher lacunas que existem no sentido de atender aos requisitos de Qualidade: Rapidez, Transparência, Confiabilidade e Rastreabilidade.
OE2	Estabelecer um processo que possa dar celeridade a uma revisão sistemática da literatura pelo uso de tecnologia para mineração de texto.
OE3	Identificar técnicas de PLN que possam ser aplicadas a uma revisão literária.

No Capítulo 2, a **questão de pesquisa** foi desenvolvida, sendo:

Como as técnicas de Processamento da Linguagem Natural (PLN) podem contribuir com a revisão literária de uma pesquisa científica?

O próximo passo é identificar as **proposições da pesquisa**. Yin (Yin, 2005) sugere a formulação de proposições, por meio das quais o investigador consiga focar nos aspectos a serem estudados, dentro da questão de pesquisa. As proposições permitem refletir sobre questões teóricas e identificar onde procurar evidências relevantes. Constituem conclusões deduzidas no levantamento teórico. Sem as proposições o investigador pode ficar tentado a recolher “tudo”, algo absolutamente impossível de fazer, afirma Yin (Yin, 2005) .

A revisão literária e os objetivos deixam claro que a base teórica atende aos domínios de modelos de revisão literária e métodos de PLN. Tendo em vista os objetivos específicos e refletindo sobre o conhecimento obtido na revisão literária

desta pesquisa. A Tabela 7 apresenta as proposições (P) de pesquisa estabelecidas, sua descrição e a sua correlação com os objetivos específicos (OE).

No decorrer deste trabalho, os Objetivos específicos (OE) e as proposições (P) são identificados, pela numeração recebida na tabela 6 e na Tabela 7. Carrara (Carrara, 2011) acrescenta que uma proposição (P) pode atender a mais de um objetivo específico (OE), como observado na Tabela 7.

Tabela 7 - Proposições, sua descrição e relação com os objetivos específicos

(P)	Descrição Proposição	Conceitos de apoio as proposições a serem atendidos.	OE.
P1	Suprir lacunas apresentadas em ferramentas de apoio a RSL, quanto à melhor forma de atender requisitos de qualidade.	Registrar execução do processo, garantindo transparência e rastreabilidade (Ros et al., 2017) (Taichert et al, 2020) Chakkarwar&Tamane, 2019) (Karunananda et al, 2021) Utilizar substantivos e adjetivos para análise pelo modelo de tópico (Kometani&Nagaoka, 2016) Manter registro de troca de termos (Kometani&Nagaoka, 2016) Permitir excluir termos (Kometani&Nagaoka, 2016) (Asmussen&Møller, 2019) Identificar e permitir exclusão de metadados duplicados (Ros et al., 2017) Permitir o armazenamento de motivos da exclusão/inclusão assim como anotações	OE 1 OE 2
P2	É possível uma solução que respeite a liberdade de condução do pesquisador e que disponibilize informações adequadas para o relato da revisão e disponibilize dados para aplicativos	Preservar a condução da pesquisa pelo pesquisador (Karunananda et al, 2021) Exibir parâmetros da utilizados na análise. (Tiba et al, 2018) Chakkarwar&Tamane, 2019) Estabelecer métricas de avaliação da quantidade de tópicos e de coerência na distribuição dos resumos (Asmussen&Møller, 2019). (Lin et al, 2020) (Mateen et al, 2021) Fornece informações para o relato da RSL (Ros et al, 2018) Integrar com ferramentas de bibliometria para análise quantitativa (Tiba et al, 2018)	OE 1 OE 2
P3	É possível prover uma solução de fácil entendimento e execução com funcionalidades amigáveis que contemple diferentes habilidades com uso de tecnologias	Desenvolver ferramenta com funcionalidades amigáveis (Ros et al., 2017). Amplia o uso da ferramenta a pesquisadores sem habilidade técnica (Asmussen&Møller, 2019). Definir técnicas para tratamento de texto como a Técnica POS tagging (Ros et al., 2017) (Chakkarwar&Tamane, 2019) (Taichert et al, 2020). (Mateen et al, 2021) Usa modelagem de dados com base em LDA como LDA2VEC ou Word2Vec + LDA (Kometani&Nagaoka, 2016) (Ros et al., 2017) (Tiba et al, 2018) (Asmussen&Møller, 2019) (Chakkarwar&Tamane, 2019) (Lin et al, 2020) (Taichert et al, 2020) (Mateen et al, 2021) Criar clusters para agrupar resumos com tópicos semelhantes. (Taichert et al, 2020) Prever consultas aos resumos, títulos e termos dos metadados em análise permitindo reverter decisões de exclusão/inclusão. (Tiba et al, 2018)	OE 1 OE 2 OE 3
P4	É possível uma solução que trate dados de diversas BDs postergando o acesso as publicações completas permitindo compartilhamento online de toda equipe.	Tratar grande volume de informações (Asmussen&Møller, 2019). Reduzir a necessidade de baixar PDFs de publicações inteiras. (Ros et al., 2017) (Asmussen&Møller, 2019). Não automatizar busca em BDs. (Karunananda et al, 2021) (Taichert et al, 2020) (Karunananda et al Permitir que o pesquisador defina a quantidade de tópico a ser gerado (Lin et al, 2020) Permitir compartilhamento de dados entre pesquisadores do projeto	OE 2 OE 3

3.5.1.2 Fase 2 — Estruturação da Pesquisa

Nesta fase é definida a arquitetura da pesquisa, construído o protocolo de pesquisa, os pontos de análise iniciais resultantes do universo pesquisado, bem como os entregáveis de cada fase.

O modelo de pesquisa se encontra neste capítulo e está esquematizado na Figura 6 onde a estrutura está detalhada em etapas e fases.

Pontos de análise

Os **pontos de análise** derivam das proposições da pesquisa e servem como base para a sua condução. Foram definidos três pontos de análise, descritos na Tabela 8, na Tabela 9 e na Tabela 10, contendo as referências teóricas, as questões específicas e as proposições relacionadas, uma para cada ponto de análise.

Tabela 8 - Ponto de Análise PA1 - Seleção do Processo

PA 1 – Desenvolvimento do processo		
Referências teóricas	Questões específicas	(P)
(Asmussen&Møller, 2019) (Ros et al., 2017) (Karunananda et al, 2021)	Criar solução que atenda pesquisadores, independente de área do conhecimento, do perfil tecnológico e da disponibilidade de acesso a BDs indexadas.	P1 P2 P4
(Chakkarwar&Tamane, 2019) (Kometani&Nagaoka, 2016)	Manter registro do histórico de execução da RSL para subsidiar o relatório final e garantir a rastreabilidade, transparência e confiabilidade do realizado.	

Tabela 9 - Ponto de Análise PA2 - Necessidade de Programação

PA 2 – Necessidade de programação		
Referências teóricas	Questões específicas	(p)
(Asmussen&Møller, 2019)(Ros et al., 2017) (Karunananda et al, 2021)	Receber arquivos com resultado de busca de quaisquer BDs e em grande volume. Garantir que arquivos baixados das BDs contenham informações necessárias para análise.	P2 P3 P4
(Chakkarwar&Tamane, 2019) (Kometani&Nagaoka, 2016)	Permitir que as atividades de seleção e análise reúnam publicações das diversas BDs consultadas, preservando a informação de origem e eliminar duplicados.	
(Lin et al, 2020)	Definir melhores práticas para tratar de textos e termos. Armazenar motivo exclusão ou troca pelo usuário. Definir métrica de validação de resultado da análise de tópicos.	

Tabela 10 - Ponto de Análise PA3 - Motivo para adoção do processo

PA 3 – Motivo para adoção do processo		
Ref. Teóricas	Questões específicas:	(p)
(Asmussen&Møller, 2019)	Dar autonomia para estabelecer o melhor resultado de agrupamento em tópicos, atendendo objetivo específicos e permitindo análise parcial (ano de publicação).	P1
	Postergar e reduzir a quantidade de PDFs de artigos completos a serem baixados.	P2
	Viabilizar o uso da ferramenta na WEB.	P3
	Possibilitar a integração da ferramenta com outras ferramentas complementares de interesse de uma RSL.	

Nestas tabelas as questões específicas serão utilizadas para avaliar o quanto cada ponto de análise é atendido em cada ciclo tendo como granularidade as questões específicas.

Os pontos de análise descritos anteriormente são o resultado final de todos os ciclos da pesquisa. Em cada ciclo houve identificações parciais que serão detalhadas na execução dos ciclos.

A apresentação dos pontos de análise encerra a primeira etapa do projeto descrita neste capítulo. A Etapa 2 e a Etapa 3 são descritas e detalhadas nos Capítulo 6 e Capítulo 7.

3.6 CONSIDERAÇÕES FINAIS

Este capítulo descreveu os principais métodos que poderiam ser utilizados nesta pesquisa e justificou a escolha da pesquisa-ação. Conclui a etapa de planejamento da pesquisa e os resultados a serem observados.

O próximo capítulo, visando o entendimento do realizado, apresenta conceitos das tecnologias empregadas e, em especial, do PLN e técnicas.

4. PROCESSAMENTO DA LINGUAGEM NATURAL (PLN)

4.1 INTRODUÇÃO

Neste capítulo é descrito o conceito de Processamento da Linguagem Natural (PLN), e seu papel como parte das áreas de conhecimento denominadas Inteligência Artificial e mineração de texto. Descrevem-se técnicas de PLN e são comentadas as principais técnicas de Modelagem de Tópicos que integram as ferramentas de PLN justificando-se a escolha do LDA com Word2Vec.

4.2 INTELIGENCIA ARTIFICIAL, MINERAÇÃO DE TEXTO E RAMIFICAÇÕES

Os computadores fornecem o meio perfeito para a experimentação e aplicação da tecnologia de Inteligência Artificial (IA). Eles são destinados a estimular o comportamento humano e são construídos para aplicações tecnológicas como instrução assistida por computador.

Para Durkin e Durkin (Durkin&Durkin, 1998) , a Inteligência Artificial (IA) é o campo de estudo da ciência que pesquisa formas de fazer o raciocínio do computador simular o raciocínio humano. Enquanto Weber (Weber, 1999) conceitua a Inteligência Artificial (IA) como o ramo da Ciência da Computação, dedicado ao estudo das técnicas computacionais que representam alguns aspectos da cognição humana. A ciência em IA está dividida em subáreas sistema especialista, processamento da linguagem natural (PLN), reconhecimento de padrões e robótica. Algumas das **aplicações computacionais** utilizadas nas áreas de desenvolvimento da IA são:

Aplicações da Ciência Cognitiva: sistemas especialistas, sistemas de aprendizagem, lógica difusa, algoritmos genéticos, redes neurais, agentes inteligentes

Aplicação de Robótica: percepções visuais, tato, destreza, locomoção, navegação

Aplicações de Interface Natural: linguagens naturais, reconhecimento de fala, interfaces multissensoriais, realidade virtual

Ryan (2023) traz o conceito de **mineração de texto** como sendo o processo ou prática de examinar grandes coleções de recursos escritos para gerar novas informações. O objetivo da mineração de texto é descobrir informações relevantes no texto, transformando em informações que podem ser usadas para análise posterior. Para isso, a mineração de texto usa uma variedade de metodologias de análise; o PLN é uma delas. O autor define a mineração de texto como a “ extração de informações e percepções do texto usando técnicas de IA e PLN ”. Essas técnicas transformam dados não estruturados em informações estruturados para facilitar o trabalho dos cientistas e analistas de dados. Mineração de texto e PLN estão intimamente ligados. Resumindo, é possível realizar PLN sem análise de texto, mas é difícil fazer análise de texto sem PLN.

O **aprendizado de máquina** (AM) (ou *Machine Learning- ML*) é um método de análise de dados que automatiza a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com um mínimo de intervenção humana.

Monard e Baranauskas (Monard&Baranauskas, 2003) conceituam AM como uma ferramenta poderosa para a aquisição automática de conhecimento e que não existe um único algoritmo que apresente o melhor desempenho para todos os problemas. Bishop (Bishop, 2006) complementa que os algoritmos de aprendizado de máquina constroem um modelo matemático com base em dados de amostra, conhecidos como dados de treinamento, para fazer previsões ou decisões sem serem explicitamente programados para executar a tarefa. Quanto ao tipo de aprendizado, o AM está organizado em uma hierarquia, conforme Figura 7.

Monard e Baranauskas (Monard&Baranauskas, 2003) esclarecem que o aprendizado indutivo é efetuado a partir do fornecimento de exemplos resolvidos para o sistema poder elaborar respostas a outros casos por meio de uma indução. O aprendizado indutivo pode ser dividido em supervisionado e não-supervisionado. No aprendizado supervisionado é fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento utilizando uma classificação ou uma regressão.

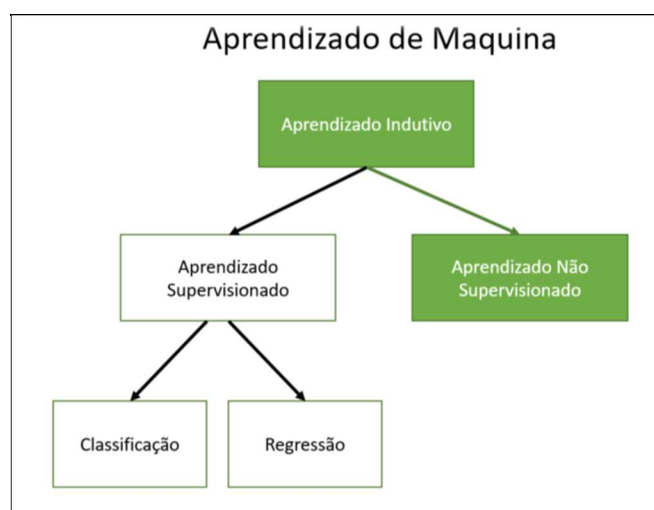


Figura 7 - Hierarquia do aprendizado de máquina (Monard&Baranauskas, 2003)

No aprendizado não-supervisionado, foco para compreender o LDA, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando clusters. Após a determinação dos clusters é necessária uma análise para determinar o que cada agrupamento significa no contexto do problema que está sendo analisado.

O aprendizado de máquina tem muitas aplicações no campo de tecnologia da informação (TI), incluindo processamento da linguagem natural (PLN), visão computacional, análise preditiva e sistemas de recomendação. É amplamente utilizado em vários setores, incluindo saúde, finanças e comércio eletrônico, para melhorar a tomada de decisão, automatizar processos e aprimorar as experiências do cliente (Khanzode et al., 2020).

4.3 CONCEITO E TÉCNICAS DO PLN

A PLN constitui um interesse central no campo da inteligência artificial e da ciência da computação. Os estudos de PLN compreendem teorias e métodos que permitem a comunicação efetiva entre humanos e computadores em linguagem natural. Como um campo de estudo científico, a PLN assimila ciência da computação, linguística e matemática com o objetivo principal de traduzir a linguagem humana (ou natural) em comandos que podem ser executados por computadores (YueKang et al, 2020) .

Uma das primeiras aplicações de PLN (MATHIAS, 2018) foi em atividades como sistemas multimídia, sistemas de informações geográficas e computação

colaborativa. A PLN tem muitas aplicações práticas, incluindo chatbots e assistentes virtuais, análise de sentimentos, tradução de idiomas, reconhecimento de fala e extração de informações. A PLN é usada em uma ampla variedade de setores, incluindo saúde, finanças, comércio eletrônico e atendimento ao cliente, para automatizar tarefas, melhorar as experiências do cliente e extrair informações dos dados.

O processo de PLN pode ser dividido em: pré-processamento de texto, representação de texto, treinamento de modelo e avaliação de modelo. A Figura 8 - Fluxograma da execução de um PLN (YueKang et al, 2020) mostra as etapas de sua condução.

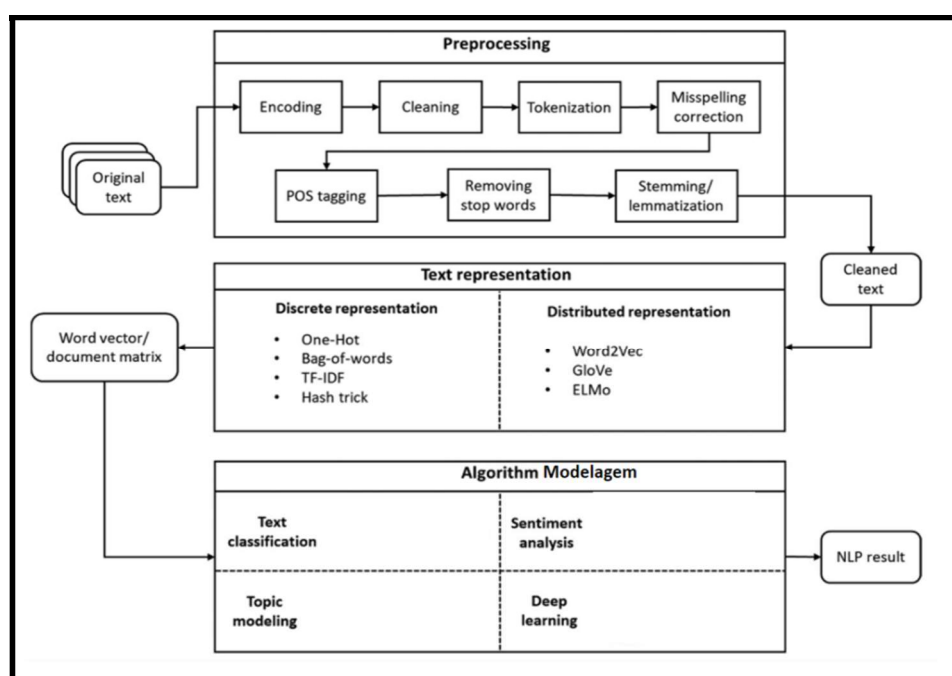


Figura 8 - Fluxograma da execução de um PLN (YueKang et al, 2020)

O **PRÉ-PROCESSAMENTO (PREPROCESSING)** visa obter um texto "limpo" para melhorar a eficiência e a precisão da análise a seguir, removendo símbolos sem sentido, verificando erros de ortografia, tokenização, marcação **POS-tagging**, a ser detalhado, removendo palavras de parada e lematização (YueKang et al, 2020) . Sendo:

Codificação (encoding): é o manuseio da codificação de caracteres deixando o texto no formato "correto", ou seja, em UTF-8, codificação padrão do Windows.

Limpeza (*Cleaning*): Dada a abundância de dados online e a irregularidade das informações a serem removidas, a limpeza é um processo muito complexo. *Tags* HTML e informações não textuais, como imagens e *emojis* devem ser limpas ou removidas do conjunto de dados. Com isso, os pesquisadores podem manter o texto exato que desejam usar para análise na pesquisa.

Tokenização (*tokenization*): A tokenização é o processo de quebrar o texto em unidades, muitas vezes na forma de palavras e frases. Existem muitos pacotes de tokenização em Python que fornecem procedimentos inteligentes de tokenização, e na nossa solução, usamos o NLTK, bastante utilizado e referenciado. Os tokens são os blocos de construção básicos do PLN. Por exemplo, na sentença:

Frase original: *Eu aprendo PLN. Vou conhecer mineração de dados.*

Tokenizar por frase cria 2 tokens:

['Eu aprendo PLN.', 'Vou conhecer mineração de dados']

Tokenizar por palavra cria 8 tokens:

['Eu', 'aprendo', 'PLN.', 'Vou', 'conhecer', 'mineração', 'de', 'dados'].

Erros ortográficos (*misspelling correction*): Consiste em verificar se palavras estão corretamente escritas e corrigi-las, se necessário. N-gram e pesquisa de dicionário são duas técnicas conhecidas de detecção de erros ortográficos. Usa algoritmos, como distância de edição, chaves de similaridade, técnicas baseadas em regras, técnicas baseadas em n-gram, técnicas probabilísticas e redes neurais.

POS tagging: Marcação de parte do discurso (*Part Of Speech tagging* – **POS):** envolve identificar a parte do discurso de cada palavra em uma frase, a fim de compreender a estrutura gramatical. O processo visa anexar a cada palavra uma "*tag*" que determina a classe gramatical (substantivo, adjetivo, verbo, etc.) de acordo com o contexto da frase. Pode envolver o reconhecimento de entidade nomeada (*Named-Entity Recognition* – NER), isto é, identificar e categorizar nomes de pessoas, organizações e locais, e a análise da estrutura gramatical de uma frase, identificando as relações de dependência entre as palavras. A análise de dependência é usada para entender o significado de uma frase e suas partes constituintes.

Remoção de palavras (*Removing Stopwords*): São palavras de ocorrência frequente em um idioma que não contribuem para o contexto semântico

significativo e são desnecessárias em aplicativos de PLN, como "a" e "the" em inglês. A biblioteca NLTK (*Natural Language Toolkit*) de Python contém listas predefinidas de tais palavras por idioma.

Lematização (*lemmatization* ou Radical) (*Stemming*): É o processo de reduzir as palavras em formas flexionais (palavras com significado, formas gerais). O tratamento denominado **LEMATIZAÇÃO** ocorre através de um **analisador morfológico** que realiza a identificação de cada palavra. O *Stemming* (radical) refere-se ao radical das palavras que não podem necessariamente expressar pensamento semântico completo. *Stemming* é frequentemente aplicado na recuperação de informações estendida. A PLN usada para análise de textos, independentemente de qual modelo de tópico é adotado, usa algoritmos Python como *WordNet*, *SpaCy* e *Snow-ball*.

Após o pré-processamento, ou seja, da preparação do texto, um documento, que passa a ser chamado de **corpora**, **assim como um** conjunto de corpora passa a ser chamado de **corpus**. Depois de pré-processar o texto, os pesquisadores devem escolher a melhor forma de **representar as palavras** para que possam ser calculadas com eficácia e precisão por computadores.

A REPRESENTAÇÃO DE TEXTO (*Text Representation*): É a conversão de palavras em números, em vetores ou matrizes. A conversão de texto pode ser classificada em representação discreta ou representação de distribuição.

Yue Kang (YueKang et al, 2020) inclui como representação discreta os algoritmos: **One-hot** que representa uma palavra como um vetor cuja dimensão é o comprimento de um corpus; **bag-of-words** (BOW) é usado para formar um vetor que representa um documento usando a contagem de frequência de cada termo no documento; **Hash Trick**, o algoritmo de redução de dimensão de texto mais comum e **TF-IDF** usado para representação de texto onde TF é a frequência de ocorrência de uma palavra no documento e IDF representa a proporção do texto que contém determinada palavra no corpus.

Wu et al. (Wu et al., 2008) descrevem TF-IDF como uma medida estatística que indica a importância de uma palavra de um documento em relação a uma coleção de documentos. O TF-IDF faz a redução de dimensionalidade e

encontra tópicos latentes que representam o relacionamento entre as palavras e os documentos. Para redução da dimensionalidade usa-se a técnica de álgebra linear para fatoração de matrizes denominada Decomposição em Valores Singulares (SVD) truncado (Faleiros, 2016). Se a palavra aparece com frequência em um corpora, mas aparece raramente no restante no corpus, ela terá uma pontuação mais alta.

A ideia-chave da representação distribuída é que o significado de uma palavra é dado de acordo com as palavras que frequentemente aparecem nas proximidades. Como representação destacam-se os algoritmos: **Word2Vec** (Rong, 2014), que faz uso da informação de contexto da palavra, tornando a informação semântica mais rica onde o vetor de palavras treinado é de baixa dimensão e denso; o **Global Vector (GloVe)** (Egger, 2022), **que** combina as estatísticas globais da Latent Semantic Analysis (LSA) e a janela de contexto local, incorporando informações estatísticas globais a priori; e o **ELMo** (Lyu, 2022), **que** representa palavras profundamente contextualizadas e modela as características complexas das palavras (como sintaxe e semântica).

Os **ALGORITMOS DE MODELAGEM** podem ser de classificação de texto, análise de sentimento, modelo de tópicos ou *Deep Learning*.

No campo da **classificação de texto**, estão a maioria dos métodos de aprendizado. que se enquadram em duas categorias: aprendizado supervisionado e aprendizado não supervisionado. A categoria de aprendizado supervisionado tem o desafio da rotulagem dos dados e os algoritmos supervisionados mais importantes são: *Logistic Regression (LR)*, *Support Vector Machine (SVM)*, *Naïve Bayes (NB)*, *Decision Tree (DT)* e *Random Forest*. Na categoria de aprendizado não supervisionado estão os algoritmos de agrupamento calculados com base na distância entre os vetores, como *Kmeans* e *K-Nearest-Neighbour (KNN)*. Os resultados calculados a partir desses algoritmos diferem dos gerados pela análise lógica dos especialistas (YueKang et al., 2020).

O campo da **análise de sentimento** visa determinar o sentimento geral ou o tom de um texto, como positivo, negativo ou neutro, de um texto ou frase. Cada

palavra recebe uma pontuação de sentimento no dicionário, e a soma de todas as palavras em uma frase ou documento é a pontuação de sentimento final para a frase ou documento em questão denominado léxico. Os léxicos mais populares são o dicionário psicolinguístico *Linguistic Inquiry and Word Count (LIWC)*, *SentiWordNet*, *MPQA Subjectivity Cues Lexicon*, *General Inquirer* e *Bing Liu Opinion Lexicons*.

O campo do **aprendizado profundo** usa algoritmos aplicados na classificação de texto, como CNN, RNN, rede neural BP, LSTM. O campo da **modelagem de tópicos** imita o processo de geração de dados e envolve a identificação dos principais tópicos ou temas em um texto, com base na frequência e distribuição de palavras. A modelagem de tópicos é usada para extrair insights de grandes coleções de documentos. Os tópicos são definidos como distribuições de palavras que comumente ocorrem simultaneamente. Um documento é descrito como uma mistura probabilística de tópicos. A ferramenta mais usada para modelagem de tópicos é a alocação latente de Dirichlet (LDA) (YueKang et al., 2020).

4.4 TIPOS DE MODELAGENS DE TÓPICOS

Os modelos de tópicos são construídos em torno da ideia de que a semântica do documento é governada por algumas variáveis ocultas ou "latentes" não observadas. Assim, o objetivo da modelagem de tópicos é descobrir essas variáveis latentes (criando tópicos) que moldam o significado dos corpora ou do corpus.

Modelos de tópicos são técnicas não supervisionadas que extraem prováveis tópicos de um corpus, criando associações probabilísticas de palavra-tópico e tópico-documento (Waal&Barnard, 2008). Entre as técnicas mais populares de modelagem de tópicos têm-se: LSA (LSI), pLSA (pLSI), LDA e LDA2vec ou LDA em Deep Learning, onde esta última é baseada em aprendizado profundo. Para Waal (Waal&Barnard, 2008), todos os modelos de tópicos são baseados nas suposições básicas de que:

- Cada documento consiste em uma mistura de tópicos e
- Cada tópico consiste em uma coleção de palavras.

Avaliar os diferentes modelos de tópicos é um desafio, porque: (a) são frequentemente empregados em dados não rotulados, de modo que não existe uma

verdade básica e (b) os agrupamentos de documentos "suaves" (probabilísticos) são criados por modelos de tópicos modernos, dificultando comparações, mesmo se os rótulos de verdade estejam disponíveis (Waal&Barnard, 2008). A seguir são descritos alguns desses modelos.

4.4.1 Latent Semantic Analysis (LSA) / Latent Semantic Indexing (LSI p/ Google)

É uma das técnicas fundamentais de modelagem de tópicos. Ela é rápida, mas exige grande volume de documentos. A ideia central é decompor uma ‘matriz documento-termos’ e uma ‘matriz documento-tópicos’ e finalmente obter uma ‘matriz tópico-termos’. Na primeira versão do LSA, faz-se a contagem bruta do número de vezes que uma palavra aparece no documento (De Almeida, 2014). No entanto, há problemas porque a frequência em que a palavra aparece, não define seu significado no documento. Os modelos LSA substituíram esta contagem bruta por uma pontuação *Term Frequency–Inverse Document Frequency* (TF-IDF).

4.4.2 Análise Semântica Latente Probabilística (pLSA)

O pLSA (*Probabilistic Latent Semantic Analysis*) no lugar de usar a técnica de álgebra linear SVD, utiliza métodos probabilísticos, cuja a ideia é adicionar um tratamento probabilístico aos tópicos e às palavras. Visa encontrar um modelo probabilístico com tópicos latentes que possam gerar dados que se observa na ‘matriz documento-termos’. Ele é um algoritmo iterativo que permite encontrar os parâmetros de máxima verossimilhança de um modelo probabilístico.

O pLSA é um modelo mais flexível do que LSA, mas ainda tem problemas como: não saber atribuir probabilidades a novos documentos e o número de parâmetros para pLSA cresce linearmente com o número de documentos, por isso é propenso a **overfitting**¹.

¹ *overfitting* - quando o modelo tem um **desempenho excelente**, porém quando utilizamos os dados de **teste** o resultado é **ruim**.

4.4.3 Latent Dirichlet Allocation (LDA)

Entre os modelos de tópicos, o LDA é o mais comumente utilizado, descrito como uma versão bayesiana do pLSA, proporcionando uma melhor generalização.

O LDA utiliza o algoritmo de '*Priori-Dirichlet*' ou distribuição à *priori* para as distribuições de tópicos por documento e de palavras por tópico. Assim, a distribuição de Dirichlet pode ser entendida como uma “distribuição sobre distribuições”, e é a distribuição a priori para distribuições multinomiais. E se propõe a responder à pergunta: “ Quais são as distribuições de probabilidade reais que provavelmente serão vistas num dado tipo de distribuição? ” (Faleiros, 2016).

Por exemplo, se um corpus tem documentos de **3 áreas de assunto diferentes**, ao modelar, o tipo de distribuição dará muito peso a um tópico específico e pouco peso aos demais. Provavelmente, a distribuição de **probabilidade poderia ser: 90 % tópico A, 5% tópico B, 5% tópico C**, indiferente de quem são os tópicos A, B e C. O resultado **improvável** será $A = B = C = 33\%$. A distribuição de Dirichlet fornece uma maneira de **amostrar distribuições de probabilidade** de um tipo específico. Assim, no LDA, *o conjunto de dados também serve como dados de treinamento para a distribuição Dirichlet de tópicos de documentos com algumas características:*

- O número de tópicos deve ser conhecido de antemão.
- Para reduzir a quantidade de termos do dicionário, desconsiderando a representação semântica das palavras em um corpus, adotou-se o algoritmo word2vector. Requer uma extensa fase de pré-processamento para obter uma representação significativa dos dados de entrada textual.
- Estudos relatam que o LDA pode produzir tópicos muito gerais ou irrelevantes (Alnusyan et al., 2020). Os resultados também podem ser inconsistentes em diferentes execuções (Yu&Egger, 2021).

4.4.4 LDA em Deep Learning: LDA2vec

O LDA2vec, uma extensão de word2vector e do LDA, aprende vetores de palavras, documentos e tópicos conjuntamente. Representa o texto dos documentos como uma mistura de tópicos. Das palavras obtém-se representações vetoriais

usando word2vector. O LDA2vec se baseia no modelo de rede ***skip-gram (n-gram)*** do ***word2vector*** para gerar vetores de palavras. A rede neural é treinada alimentando com os pares de palavras dos documentos usados no treinamento. Para criar o modelo, constrói-se um vocabulário de palavras a partir dos documentos de treinamento, encontrando, por exemplo, um vocabulário de 1.000 palavras únicas. A rede de saída é um único vetor contendo, para cada palavra do vocabulário, a probabilidade da palavra próxima, a ser selecionada aleatoriamente, ser aquela palavra do vocabulário de entrada. Se duas palavras diferentes tiverem “contextos” semelhantes (ou seja, palavras iguais aparecem ao redor), o modelo precisará gerar resultados semelhantes para essas duas palavras (Albanese, 2022) .

4.4.5 Indexação Semântica Latente (LSI)

Constrói uma matriz que contém contagens de palavras por documento de um grande pedaço de texto. Usa uma técnica matemática chamada decomposição de valor singular (SVD) para reduzir o número de linhas e preservar a estrutura de similaridade entre as colunas. A matriz é representada nas linhas por palavras únicas e nas colunas por cada documento. Configurar o modelo LSI é da mesma forma que a configuração do modelo LDA (Albanese, 2022) .

4.4.6 Processo Hierárquico de Dirichlet (HDP)

É um Modelo de associação mista para análise não supervisionada de dados agrupados. Ao contrário do LDA (sua contraparte finita), o HDP infere o número de tópicos dos dados (STACK, 2023).

4.4.7 BERTopic e Top2Vec

Grootendorst¹¹ (2022) e Angelov¹² (2020) propuseram novas abordagens para modelagem de tópicos, BERTopic e Top2Vec, respectivamente. Esses modelos abordam as limitações das estratégias convencionais com os modelos anteriores (Albanese, 2022) . BERTopic e Top2Vec fabricam incorporações semânticas a partir de documentos de entrada. A BERTopic aproveitou o BERT *Sentence Transformers* (SBERT) para fabricar representações vetoriais de palavras e frases contextuais de alta qualidade. Enquanto o Top2Vec usou o Doc2Vec para criar vetores de palavras,

4.4.8 EXEMPLO DE LEMATIZAÇÃO e N-GRAM

Existem vários algoritmos de lematização. Independentemente do modelo de tópico a ser adotado para analisar documentos é necessário, antes, tratar as palavras dos documentos através da análise semântica. A análise semântica é considerada um dos maiores desafios do PLN pois, a semântica está relacionada ao significado, não só de cada palavra, mas também da expressão. Ela se ocupa com a morfologia e a estrutura sintática e, em alguns casos, com informações pragmáticas (considerações de ordem prática, realista). Este tratamento é denominado **LEMATIZAÇÃO**, a identificação de cada palavra por meio de um **analisador morfológico**. Em seguida uma **análise sintática** classifica as palavras e o sentido das estruturas das palavras que foram reagrupadas. Para esta pesquisa, os algoritmos de Lematização utilizados estão relacionados na Tabela 11 e executados na sequência apresentada como a descrição e a exemplificação da saída tendo como entrada o conteúdo:

E = ('*Você encontrou o livro que eu te falei, Carla?*')

Tabela 11 - Tratamento de cada corpora (elaborado pela autora)

Função	Descrição	Saída
Primeira	Tokenizar palavras do documento	['Você', 'encontrou', 'o', 'livro', 'que', 'eu', 'te', 'falei', 'Carla?']
Segunda	Separar as pontuações das palavras	['Você', 'encontrou', 'o', 'livro', 'que', 'eu', 'te', 'falei', ',', 'Carla', '?']
terceira	análise ortográfica – separa símbolos	['Você', 'encontrou', 'o', 'livro', 'que', 'eu', 'te', 'falei', '.', 'Carla', '?']
quarta	elimina pontuações	['Você', 'encontrou', 'o', 'livro', 'que', 'eu', 'te', 'falei', 'Carla']
quinta	calcula similaridade	tokens[0].similarity(tokens[5]) => 0.29921812
	(2 tokens por vez)	tokens[0].similarity(tokens[3]) => -0.067587696
s exta	análise gramatical de cada token	[('Você', 'PRON'), ('encontrou', 'VERB'), ('o', 'DET'), ('livro', 'NOUN'), ('que', 'PRON'), ('eu', 'PRON'), ('te', 'PRON'), ('falei', 'VERB'), (',', 'PUNCT'), ('Carla', 'P -ROPN'), ('?', 'PUNCT')]
s étima	Lematiza- verbo p/infinitivo	['encontrar', 'falar']
oitava	lemma - com mesmo radical	['encontrei', 'encontraram', 'encontrarão', 'encontrariam']

N-gram – exemplo da função

Separa o conteúdo de um texto em N-letra, N-palavra ou N-frase.

Tendo-se o texto: *"remote sensing platforms acquire huge amounts of data every day as a result, large archives of data have been created"*

“Plataformas de sensoriamento remoto adquirem enormes quantidades de dados todos os dias, como resultado, ...”

N-gram de 2 letras, tem-se como os primeiros bigrams:

[‘P l’, ‘l a’, ‘t a’, ‘a f’, ‘f o’, ‘o r’, ‘r m’, ‘m a’, ‘a ’, ‘ d’, ‘d e’, ‘e ’, ‘ s’, ‘s e’, ‘e n’, ‘n s’, ‘s o’, ‘o r’, ‘r i’, ‘i a’, ‘a m’, ‘m e’, ‘e n’, ‘n t’, ‘t o’,]

N-gram de 1 palavra é:

[‘Plataformas’, ‘de’, ‘sensoriamento’, ‘remoto’, ‘adquirem’, ‘enormes’, ‘quantidades’, ‘de’, ‘dados’, ‘todos’, ‘os’, ‘dias’, ‘como’, ‘resultado’].

N-gram de 2 palavras é:

[‘Plataformas de’, ‘de sensoriamento’, ‘sensoriamento remoto’, ‘remoto adquirem’, ‘adquirem enormes’, ‘enormes quantidades’, ‘quantidades de’, ‘de dados’, ‘dados todos’, ‘todos os’, ‘os dias’, ‘dias, como’, ‘como resultado’].

N-gram de 3 palavras é:

[‘Plataformas de sensoriamento’, ‘de sensoriamento remoto’, ‘sensoriamento remoto adquirem’, ‘remoto adquirem enormes’, ‘adquirem enormes quantidades’, ‘enormes quantidades de’, ‘quantidades de dados’, ‘de dados todos’, ‘dados todos os’, ‘todos os dias’, ‘os, dias como’, ‘dias como resultado’]

4.5 OUTRAS TECNOLOGIAS UTILIZADAS

pyLDAvis

O LDAvis (Bem Mabey, 2015) foi projetado para ajudar os usuários a interpretar os tópicos em um modelo de tópico adequado a um corpus de dados de texto. O pacote extrai informações de um modelo de tópico LDA ajustado para informar uma visualização interativa baseada na web. O pyLDAvis é uma biblioteca Python de código aberto que ajuda na análise e criação de visualizações altamente interativas dos clusters criados pelo LDA. O LDA e pyLDAvis são usados para criar visualizações de *Topic Modeling Clusters*.

Clustering e modelagem de tópicos:

Na modelagem de tópicos são extraídos tópicos dos documentos e transformados em um espaço de dados menor, denominado espaço de tópicos. O número de tópicos extraídos é muito menor que a coleção de documentos e seu vocabulário. No clustering, a saída final contém um conjunto de clusters, cada um incluindo um conjunto de documentos (Faleiros, 2016) (Bettina, 2011).

Avaliação de modelo (Indicadores):

A seguir, são descritos os indicadores utilizados na pesquisa, os mais comumente aplicados para análise de tópicos para avaliação do modelo gerado pelo algoritmo. São eles:

Perplexidade: Medida de avaliação para modelos de tópicos comuns como o LDA. Quanto maior a perplexidade, pior a qualidade do modelo (Chuang, 2012). A perplexidade é calculada usando o conjunto de teste, que não foi visto pelo modelo durante o treinamento. Ela mede a capacidade do modelo de prever as palavras no conjunto de teste com base nos tópicos inferidos do conjunto de treinamento. Em outras palavras, quanto menor a perplexidade, melhor o modelo é para prever as palavras que aparecem no conjunto de teste.

A perplexidade pode ser usada para ajustar os hiperparâmetros do modelo LDA, como o número de tópicos. Isso implica que, à medida que o número de tópicos aumenta, a perplexidade diminui até alcançar um ponto em que os benefícios da adição de mais tópicos são superados pelos custos adicionais de processamento e interpretação. Portanto, a perplexidade pode ser usada como um guia para selecionar o número ideal de tópicos para um determinado conjunto de dados.

Afirmação: Este indicador foi criado para o Sw3T. Ele representa o valor inverso da perplexidade, o que facilita a análise visual do seu valor em conjunto com o indicador de acurácia. Dessa forma, quanto maior a afirmação, melhor é a escolha da quantidade de tópicos.

Acurácia: É uma medida da capacidade do modelo LDA de classificar corretamente os documentos em seus respectivos clusters. Essa medida é geralmente calculada através da taxa de acerto, que é a proporção de documentos corretamente classificados em relação ao total de documentos (Chase, 1995).

Uma medida de capacidade alta indica que o modelo é capaz de distinguir claramente entre os diferentes clusters e classificar corretamente a maioria dos documentos. Isso sugere que o modelo é uma boa escolha para a tarefa de classificação.

Por outro lado, um valor baixo de acurácia indica que o modelo tem dificuldade em distinguir entre os diferentes clusters e classificar corretamente os documentos. Isso sugere que o modelo não é uma escolha ideal para a tarefa de classificação e pode precisar ser refinado ou substituído por um modelo diferente com maior assertividade.

É importante notar que a medida de acurácia não deve ser usada isoladamente para avaliar a eficácia de um modelo de LDA, mas sim em conjunto com outras medidas de avaliação, como a Afirmação. Somente com uma análise abrangente é possível determinar a adequação do modelo para a tarefa específica em questão.

Dominância:

A dominância em um tópico do LDA refere-se à probabilidade de o documento pertencer a esse tópico em relação aos outros tópicos. Por exemplo, se a dominância de um determinado tópico for de 0,8, significa que a probabilidade de o documento pertencer a esse tópico é de 80% (De Almeida, 2014).

A interpretação da dominância é importante para entender a relevância de um tópico em um documento ou em um conjunto de documentos. Se um determinado tópico tem uma dominância alta em um conjunto de documentos, pode-se deduzir que esse tópico é importante e influente nesse conjunto de documentos. Isso pode ajudar a identificar tendências, padrões e interesses em um determinado conjunto de dados.

No entanto, é importante lembrar que a interpretação da dominância em um tópico não deve ser vista como uma verdade absoluta, mas sim como uma medida estatística que pode fornecer insights úteis para análises e decisões. É importante avaliar cuidadosamente os resultados do LDA e considerar outros fatores relevantes antes de tirar conclusões sólidas.

4.6 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentadas ferramentas que permitem fazer mineração de dados e, definir, após uma análise comparativa, que o LDA se mostrou mais adequado para ser utilizado neste caso específico.

No próximo capítulo, será descrito o processo desenvolvido para fazer a revisão sistemática da literatura com base em uma ferramenta de apoio. Em seguida será descrita a arquitetura desta ferramenta.

5. PROCESSO PROPOSTO DE REVISÃO LITERÁRIA COM SW3T

5.1 INTRODUÇÃO

Este capítulo descreve o processo proposto de revisão sistemática da literatura utilizando uma ferramenta de software, denominada Sw3T, desenvolvida para esta finalidade e que permite interação com o pesquisador e automação de algumas atividades. Vale ressaltar que um ponto importante do processo proposto é que se trabalha com termos de tópicos e não com palavras-chave, entendendo que a palavra-chave é dada pelo pesquisador e que os tópicos, compostos por termos latentes, são identificados pelo Sw3T.

A seguir, são descritas a arquitetura e funções do Sw3T.

5.2 REQUISITOS DO PROCESSO DE RSL COM SW3T

Baseado nos trabalhos correlatos e nas oportunidades identificadas para desenvolver um trabalho que corrobore com a melhoria do processo de revisão literária, foram identificados os seguintes requisitos para o processo:

- [1] Receber metadados de publicações, provenientes de diferentes BDs, que possam ter seus critérios de inclusão e exclusão refinados utilizando o aplicativo
- [2] Identificar metadados com informações incompletas e viabilizar sua correção
- [3] Permitir o recebimento de arquivos de diferentes BDs integrando-os na execução do processo
- [4] Criar tabela de *Stopwords* do pesquisador e de tabela de troca de termos **de/para** prevendo padronização de siglas, palavras similares e ajuste de erro de grafia
- [5] Criar solução para a eliminação de publicações duplicadas
- [6] Disponibilizar recursos para apoiar o processo de exclusão por termo, título, resumo ou conteúdo
- [7] Permitir a definição de parâmetros e a múltiplas execuções da análise na geração de tópicos
- [8] Manter o histórico das análises realizadas

- [9] Manter informações da seleção - motivo exclusão/inclusão dos metadados e comentários.
- [10] Estabelecer indicadores de desempenho da análise quanto ao número de tópicos, palavras-chave e ciclos de aprendizado
- [11] Possibilitar a análise por período dos artigos de um projeto
- [12] Disponibilizar visão gráfica para exibir a relação entre os tópicos
- [13] Prever a exportação dos registros da exclusão/inclusão e outras informações relevantes para compor o relatório da RSL.

Esses requisitos são provenientes das proposições e pontos de análise onde há a referência aos autores.

5.3 O PROCESSO DE RSL COM O SW3T

O processo de RSL com o Sw3T aqui descrito, contribui com as três fases previstas no modelo de Revisão Sistemática da Literatura – RSL de Kitchenham [2013].

Na Fase 1 – planejamento, este modelo possibilita concentrar todas as informações relevantes na ferramenta Sw3T, compartilhando com outros pesquisadores do projeto e possibilitando sua exportação para compor o relatório de uma RSL. Propõem-se que sejam armazenadas informações como a questão de pesquisa, os objetivos do projeto, as proposições e os pontos de análise, assim como informações do protocolo de pesquisa (strings de busca, BDs consultadas e critério de inclusão e exclusão).

A Fase 3 – relato da revisão é diretamente beneficiada com a disponibilização de todas as informações existentes no Sw3T, como as informações de planejamento armazenadas e todo o histórico da RSL realizada para compor o relatório. As informações podem ser exportadas para um arquivo no formato planilha, possibilitando a integração com outros aplicativos com funções específicas para atender a RSL como a criação de gráficos.

A Fase 2 – realização da seleção, é o gargalo de toda a RSL e a mais beneficiada neste processo proposto. O processo estabelece uma interação dinâmica com o pesquisador apoiando suas atividades.

O Sw3T não obriga a execução das atividades em uma sequência pré-definida. Todas as atividades podem ser repetidas à medida que o pesquisador adquire um maior entendimento sobre os metadados que estão sendo tratados.

A seguir, está descrito como o processo proposto de RSL com o Sw3T se desenvolve. A Figura 9 contempla as atividades com suas tarefas e a integração entre elas e o aplicativo.



Figura 9 - Processo de Revisão da Literatura com o Sw3T

Observa-se que a atividade “Buscar por publicações” não é coberta pela ferramenta Sw3T, porém, o resultado dessa atividade é um conjunto de arquivos com metadados selecionados das BDs. Esses arquivos são inseridos na ferramenta atendendo às demais atividades. As atividades do processo são:

ATIVIDADE DE BUSCA POR PUBLICAÇÕES: Esta atividade se caracteriza pelas tarefas de:

Consultar BDs na busca por publicações: O pesquisador, com acesso à internet e às BDs, utilizando os termos de busca, identifica publicações de interesse em potencial. Neste processo, os termos de busca rigorosos visando reduzir ao máximo os achados nas diversas BDs, passam a não serem críticos.

Gerar arquivo metadados em XML: O arquivo a ser baixado é gerado na busca nas BDs consultadas. O formato dos arquivos baixados difere para cada BD. Deve-se identificar qual padrão de exportação que traz, mas informação. Como o

Sw3T ainda está limitado a receber só arquivos XML, o arquivo baixado pode ser convertido em XML utilizando ferramentas disponíveis como Mendeley.

Esta atividade exaustiva também é minimizada pelo Sw3T uma vez que o refinamento, sobre o conjunto de metadados baixados, pode ter continuidade aplicando buscas com novos termos, sobre o repositório de metadados do projeto.

Um dos benefícios do processo RSL com Sw3T, é que ele trata apenas os metadados das publicações ampliando o universo de BDs que podem ser pesquisados, possibilitando buscas em quaisquer bases de dados que permitam gerar arquivos para exportação com informações apenas de metadados. Em soluções que processam artigos completos, a busca é restrita às BDs disponibilizadas pela organização onde o pesquisador está vinculado e tem acesso aos PDF das publicações.

Todas as demais atividades do processo RSL com Sw3T, descritas a seguir, são realizadas de forma interativa com a ferramenta.

ATIVIDADE DE IMPORTAÇÃO E EXPORTAÇÃO DE ARQUIVOS. Esta atividade trata do relacionamento do processo com o ambiente externo, tanto para receber ou deletar todas as informações do projeto como também para disponibilização de informações. Inclui as seguintes tarefas:

Incluir metadados: Esta tarefa deve ocorrer para cada arquivo baixado, viabilizando o registro da fonte de origem. O Sw3T avisa quando ocorrem metadados incompletos e permite uma nova inclusão após serem completados.

Deletar metadados: é uma tarefa que prevê a limpeza do banco de dados do projeto no Sw3T. É uma ação radical e foge do escopo da RSL propriamente dito. Em uma RSL, a seleção dos metadados deve ser registrada, assim, diferente do resultado da deleção, os metadados excluídos são mantidos no banco de dados do Sw3T com o status de excluído e com a descrição do motivo.

Criar planilha do projeto: é a recuperação das informações do projeto com informações do projeto e o histórico da seleção pela criação de um arquivo no formato planilha a ser exportado.

ATIVIDADE DE DEFINIÇÃO DE AJUSTES NO RESUMO: Esta atividade inclui três tarefas:

Excluir duplicados (individual ou coletiva): Todos os metadados duplicados, isto é, com informação de título, ano e resumo iguais, são exibidos e podem ser excluídos em uma única vez. A exclusão mantém o metadados no banco de dados do projeto mudando o status para excluído e motivo duplicado.

No caso de metadados parcialmente duplicados, isto é, quando uma das três informações não é igual, o pesquisador precisa decidir um a um qual metadados deve ser excluído. Uma vez que as informações vêm de várias fontes, é esperado que existam metadados iguais, mas muitas vezes eles podem divergir na data da publicação ou no conteúdo do resumo. No caso de divergência no resumo, algumas vezes está relacionado ao fato das BDs consultadas terem incluído no fim do resumo, suas informações como por exemplo: “@” e “IEEE”.

Definir regras de limpeza de texto: As regras de limpeza de termos observadas pelo pesquisador contam com duas tabelas, que podem ser atualizadas a qualquer momento e definem as regras de exclusão ou troca de termos.

ATIVIDADE DE ANÁLISE DOS RESUMOS: Aqui estão as tarefas para tratar os resumos e gerar tópicos, clusters e gráficos.

Tratar resumo: é a tarefa que atualiza o conteúdo dos resumos e os prepara para a próxima tarefa. As ações automatizadas são:

- **Transformar texto:** Modifica o texto do resumo original tornando todas as letras minúsculas e eliminando símbolos de pontuação. Esta ação segue as melhores práticas de limpeza de uma PLN, descrita no capítulo 4.
- **Efetivar as regras de limpeza:** é quando as regras estabelecidas pelo pesquisador que constam na tabela De/Para e na tabela *Stopwords* do Pesquisador são utilizadas para criar o resumo enxuto.
- **Retirar *Stopwords* padrão:** é a ação que se utiliza do arquivo, disponibilizado no GitHub, com o conjunto de palavras consideradas ruídos para uma análise, como as preposições, artigos, etc. Ela contém as melhores práticas de preparação de texto segundo especialista de PLN, conforme descrito no capítulo 4.
- **Lematizar Termos:** Trata-se da substituição de um termo pelo seu radical, termo de origem ou similar, de forma que as palavras com

mesmo sentido sejam igualadas. Esta ação segue as melhores práticas de lematização utilizadas numa PLN e são descritas no capítulo 4.

Gerar tópicos, clusters e gráficos: esta tarefa tem por função apresentar o resultado da execução do modelo de tópico probabilístico LDA. Ela contempla as seguintes ações:

- **Definir parâmetros de análise:** são os parâmetros a serem informados pelo pesquisador, como número de tópicos, número de palavras-chave por tópico, número de ciclos de aprendizado e intervalo de anos de publicação. As descrições desses parâmetros estão detalhadas no Apêndice I – Manual do Sw3T.
- **Gerar Tópicos:** esta tarefa executa a modelagem de tópico LDA sobre os resumos lematizados ainda mantidos, usando os parâmetros informados pelo pesquisador. A sua execução transforma os resumos em um conjunto de tokens, identifica tokens compostos utilizando o módulo 2-gram, cria um “dicionário” contendo, para cada token, informações de quantidade de resumos onde eles aparecem, reduz o número de tokens utilizando o modelo TF-IDF, estabelece o conjunto de palavras-chave para cada tópico latente, e identifica quais tópicos são observados em cada resumo.
- **Gerar clusters:** esta tarefa gera os clusters, com base nos tópicos identificados em cada resumo, identificando quais resumos possuem pelo menos dois tópicos latentes iguais e considerando a relevância das palavras-chave desses tópicos. Reuni num cluster, encabeçado por um dos tópicos, todos os resumos similares.
- **Calcular Indicadores:** esta tarefa trata os cálculos dos indicadores da análise com os parâmetros informados pelo pesquisador. O indicador afirmação sinaliza a assertividade da quantidade de tópicos parametrizada, o indicador da acurácia estabelece a precisão da distribuição dos resumos nos clusters e o indicador de dominância indica a afinidade de cada resumo com o cluster.
- **Gerar gráfico:** a execução desta tarefa é opcional e visa gerar o gráfico de visualização LDavis. Exibe o relacionamento entre os

tópicos da análise. A geração do gráfico tem um processamento pesado, mas gerá-lo ou não é uma decisão do pesquisador. No Apêndice I – Manual do Sw3T é feito um detalhamento desta função.

ATIVIDADE DE SELEÇÃO DE PUBLICAÇÕES: Esta atividade inclui as seguintes tarefas:

Exibir os tópicos de uma análise: É a visualização dos principais resultados de uma análise visando a seleção de publicações.

Incluir termos na tabela de Stopwords do pesquisador: Na Figura 10 são exibidos os tokens utilizados na análise que podem ser excluídos ou corrigidos fazendo uma atualização nas tabelas de Stopwords ou De/Para. São exibidos os tokens utilizados na análise que podem ser excluídos ou corrigidos fazendo uma atualização nas tabelas de Stopwords ou De/Para.

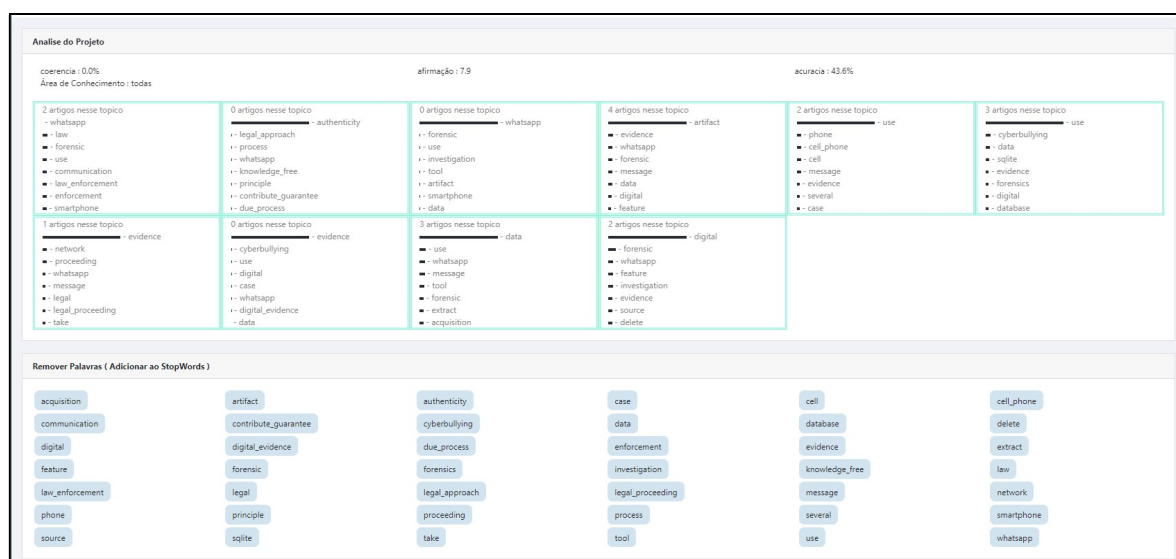


Figura 10 – Visão parcial da tela do Sw3T de visualização de Tópicos

Visualizar o gráfico de proximidade entre tópicos: Na Figura 11 tem-se a exibição do gráfico, quando requerido na execução de análise. O gráfico contribui com os indicadores de acurácia e de afirmação no sentido de estabelecer a qualidade do resultado da análise. Quanto menor a intersecção entre os tópicos, maior sua representatividade, afetando o indicador de acurácia no que se refere à precisão da distribuição dos resumos nos clusters.

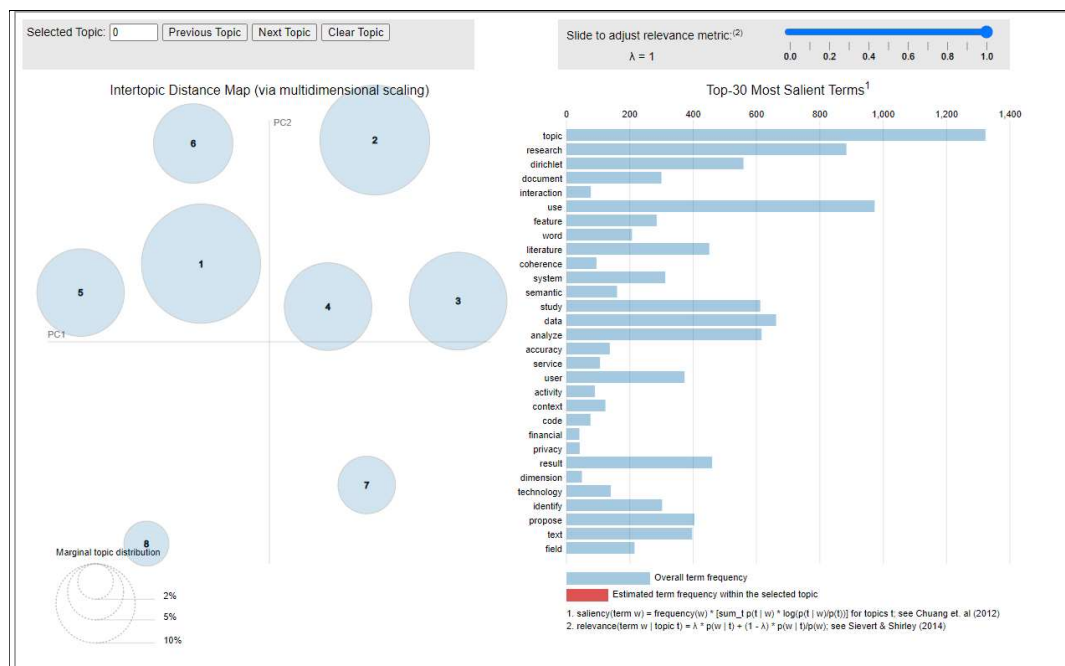


Figura 11 - Gráfico LDAvis (Nature, 2022) do Sw3T - proximidade dos clusters.

Os clusters concentram os resumos de artigos que possuem uma correlação alta de tópicos, facilitando a decisão de exclusão Individual ou coletiva. Todas as análises realizadas são mantidas no histórico do projeto prevendo comparação entre elas.

Exibir o cluster de resumos da análise: A Figura 12 informa o título e ano de publicação de todos os resumos. Apresenta o indicador de dominância de cada resumo indicando o seu alinhado com aquele cluster. Também é permitido excluir todos os artigos do cluster.

Título	Ano	% Dominancia	Status	#
browser forensic investigations of whatsapp web utilizing indexeddb persistent storage	2020	99.55%	Mantido Motivo	Ver Deletar
framework design for the retrieval of instant messaging in social media as electronic evidence	2020	99.54%	Mantido Motivo	Ver Deletar

Figura 12 – Tela do Sw3T com resultado detalhado da análise

No caso destas informações não serem o suficiente para tomada de decisão, pode-se visualizar o resumo com as várias versões.

ATIVIDADE GERENCIAMENTO DO PROJETO: É uma atividade que permite exibir informações gerais do andamento do projeto com as seguintes tarefas:

Visualizar as análises realizadas, esta visão de conjunto, observada na Figura 13 permite comparar os resultados das diversas análises, excluir e escolher a que será utilizada.

#	Qtd Artigos	Afirmação	Acuracia	Tópicos/Palavras	Ciclos de aprendizado	Anos	Status	Data	Área Conhecimento	Ver	Deletar
511	17	7.9	43.6%	10 / 8	8	0/2022	concluido	20/03/2023 22:28:36	todas	Ver	Deletar
500	17	7.9	43.6%	10 / 8	8	0/2022	concluido	15/03/2023 23:50:35	todas	Ver	Deletar
499	17	7.9	43.6%	10 / 8	8	0/2022	concluido	15/03/2023 23:50:29	todas	Ver	Deletar
478	17	7.9	43.6%	10 / 8	8	0/2022	concluido	13/03/2023 00:18:34	todas	Ver	Deletar
474	17	8.6	38.3%	10 / 10	8	0/2022	concluido	12/03/2023 12:02:25	todas	Ver	Deletar
470	17	8.5	14.2%	6 / 6	10	0/2022	concluido	10/03/2023 23:08:42	todas	Ver	Deletar

Figura 13 – Tela do Sw3T com todas as análises do projeto

Visualização da situação atual do projeto: A Figura 14 permite visualizar a evolução da seleção facilitando o relato da pesquisa e a construção do prisma.

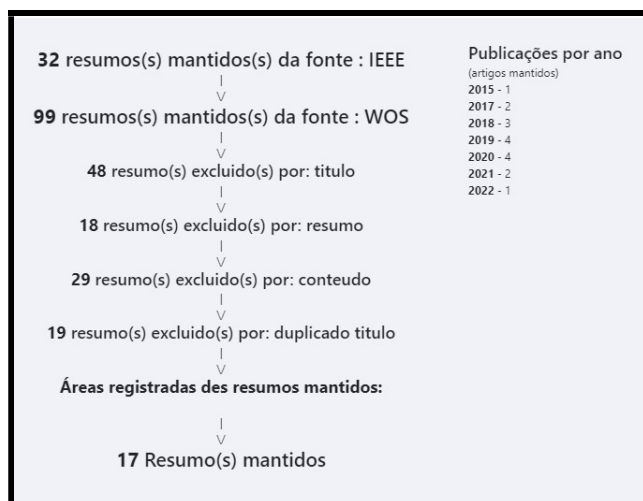


Figura 14 - Tela do Sw3T com histórico de seleção

Para ajudar na análise, alguns autores recomendam a estabelecer rótulos que representem cada tópico. O rótulo colabora com o entendimento do tópico auxiliando na seleção dos resumos.

Esta seção descreveu o processo de RSL com o Sw3T, na seção seguinte a ferramenta Sw3T é descrita. Observa-se, no entanto, que o detalhe do uso da ferramenta se encontra no Apêndice I - Manual do Sw3T deste trabalho.

5.4 A FERRAMENTA SW3T

Esta seção descreve a ferramenta Sw3T quanto à sua arquitetura e funções. A explicação de uso é detalhada no Apêndice A – Manual do Sw3T. A sigla Sw3T é o acrônimo de **Software Tópico Tem Termos** (e em inglês adaptado para **Software Term To Topic**).

Esta ferramenta foi implementada utilizando software livre com as tecnologias PHP, Python, HTML, CSS e banco de dados MySQL. Foram utilizadas as plataformas XAMPP e Anaconda para a etapa de codificação, teste e documentação do código. A documentação do código foi realizada utilizando o Jupyter notebook. Para apresentação gráfica dos tópicos, foi utilizado o aplicativo LDAvis construído com a linguagem R.

5.4.1 A arquitetura do banco de dados MySQL

O Sw3T suporta a realização de projetos diferentes e de diversos pesquisadores. Cada projeto corresponde a uma RSL realizada com um

determinado nome de projeto por um determinado pesquisador. Um pesquisador pode realizar mais de um projeto de RSL e existe a facilidade de aproveitamento de algumas tabelas de um projeto para outro.

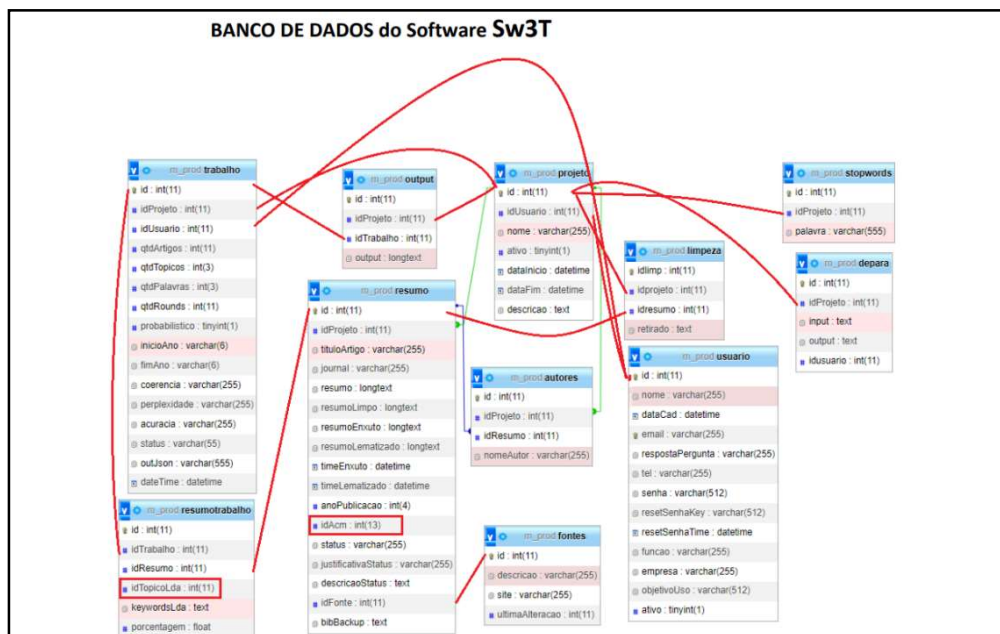


Figura 15 - Modelo de Dados do Sw3T

A Figura 15 descreve o modelo de dados relacional do Sw3T.

As Tabelas do MySQL e os relacionamentos entre elas ocorrem através das **chaves primárias (Keywords)** que estão representadas pelas linhas que as conectam. Esse modelo de dados contém 11 tabelas que são descritas a seguir.

1. Usuário – dados de cadastramento de cada usuário
2. Projeto – dados de cadastramento de cada projeto
3. De/para – tabela De/Para por usuário/projeto
4. Stopwords– tabela com termos do usuário a serem excluídos
5. Fontes – relação de BDs indexadas de onde originaram os metadados.
6. Resumo – informações do metadados mutações de conteúdo do resumo assim como as informações de status, motivo e detalhe do motivo.
7. Autores – relação de todos os autores de cada artigo
8. Outputs– relação de cada análise com o trabalho e projeto
9. Trabalhos – parâmetros e indicadores de uma análise
10. Resumo do trabalho – contém os tópicos resultantes da análise

11. Limpeza– são projetos excluídos possíveis de serem resgatados

5.4.2 Funções da ferramenta Sw3T

O Sw3T disponibiliza um conjunto de funções que atendem às atividades do processo RSL com Sw3T como descrito na seção anterior. As funções foram agrupadas da seguinte forma: iniciais, metadados do projeto, tabelas do projeto, análise dos resumos e acompanhamento dos resultados. A visualização da tela principal do projeto está na Figura 16, onde é possível visualizar grande parte de suas funções.

Funções Iniciais:

- ✓ Cadastrar usuário
- ✓ Cadastrar projeto, atualizar dados do projeto e excluir projeto

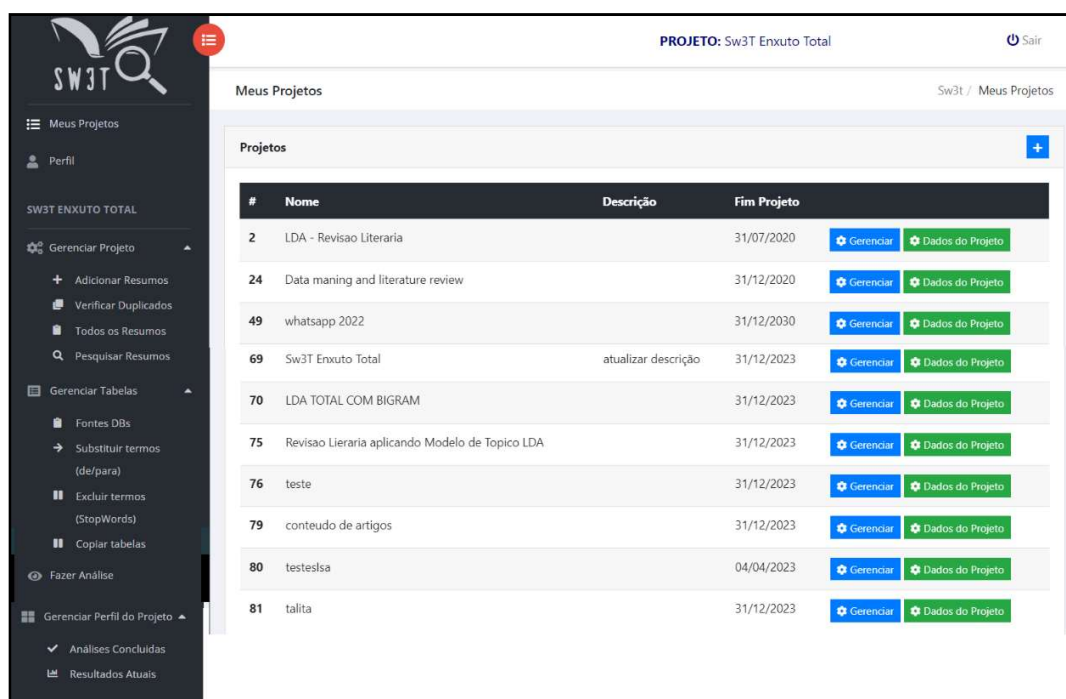


Figura 16 - Tela principal do Aplicativo Sw3T

Funções de metadados do projeto

São funções que fazem o relacionamento do projeto com o meio externo e permitem visualização de todos os metadados, inclusive a limpeza do projeto

- ✓ Adicionar arquivos com os metadados

- ✓ Verificar metadados duplicados (por título, resumo e ano) com visualização das informações e opção de visualização do resumo original, título, ano e autor.
- ✓ Visualizar todos os metadados de um projeto.
- ✓ Deletar os resumos da base de dados do projeto
- ✓ Exportar informações do projeto em planilha
- ✓ Pesquisar resumos e título utilizando diversos critérios
- ✓ Permite mudança no status de metadados de mantido para excluído e vice-versa.

Funções das tabelas do projeto

São as funções de atualização das tabelas com facilidade de migração dos dados para outro projeto.

- ✓ Atualizar tabela de substituição de termos (de/para) – definida pelo usuário estabelecendo a regra de troca de termos, podendo trocar siglas pelo seu significado ou o contrário e outras particularidades observadas pelo usuário.
- ✓ Atualizar tabela de Stopwords do usuário: Contém todos os termos (compostos ou não) que devem ser excluídos do resumo enxuto.
- ✓ Atualizar tabela de Fontes: mantém a relação das bases de dados de origem dos metadados
- ✓ Copiar tabelas: permite que o usuário migre o conteúdo das tabelas de/para e *Stopwords* para outro projeto do mesmo usuário.

Funções de análise do projeto

Disponibiliza as funções de tratamento dos resumos e criação dos tópicos.

- ✓ Transformar resumo (tira pontuações e deixa todas as letras em minúsculo)
- ✓ Tratar tabelas (executa as trocas de termos previstas na tabela de/para e exclui os termos existentes na tabela *Stopwords* do usuário)
- ✓ Retirar *Stopwords* padrão, lematizar os termos e cria os tokens
- ✓ Gerar tópicos – realiza a modelagem de tópicos de resumos lematizados com base nos parâmetros definidos pelo usuário, que são:

- Quantidade de tópicos
- Quantidade de palavras por tópico
- Quantidade de ciclos de aprendizado do modelo
- Intervalo de ano de publicação
- Área de conhecimento escolhida
- Gerar gráfico com as informações de relação entre tópicos e termos.

Funções de acompanhamento do projeto

Disponibiliza informações dos resultados do projeto, relacionando as análises realizadas:

- ✓ **Visualizar Análises concluídas:** apresenta as informações de todas as análises permitindo sua exclusão ou dando acesso às suas informações como: Tópicos, palavras por tópico, quantidade de resumos em cada tópico, indicadores de afirmação e acurácia e relação de tokens existentes no projeto possibilitando incluir um a um na tabela de *Stopwords* do projeto.
- ✓ **Visualizar os resumos de um determinado cluster:** exibe título, ano e seu índice de dominância do resumo dentro do cluster. Dá acesso à visualização do resumo permitindo sua exclusão. Contém a função de excluir todos os resumos daquele tópico.
- ✓ **Visualizar resultado atual do projeto:** Disponibiliza a visualização das etapas de exclusão e inclusão dos metadados a fonte de origem e em que momento da avaliação eles foram eliminados sendo: duplicado, título, resumo ou pelo conteúdo.

5.5 CONSIDERAÇÕES FINAIS

Neste capítulo foi descrito o processo de revisão sistemática da literatura com a ferramenta Sw3T. Também foi descrita a arquitetura do Sw3T e suas principais funcionalidades.

O próximo capítulo descreve os três ciclos da pesquisa-ação realizados na etapa 2 – ciclo de pesquisa-ação da estrutura da pesquisa e os resultados observados em cada ciclo.

6. APLICAÇÃO DO PROCESSO DE RSL COM SW3T

6.1 INTRODUÇÃO

Neste capítulo, será descrita a Etapa 3 – ciclo de pesquisa-ação, conforme o planejamento da pesquisa.

6.2 PLANEJAMENTO DA PESQUISA-AÇÃO

Esta etapa, que inclui a Fase 3 – ação e Fase 4 – avaliação, foi executada em 3 ciclos. Em cada ciclo, as ações foram definidas com base nas questões específicas dos pontos de análise e estes foram analisados quanto ao seu atendimento. As execuções de RSL foram feitas em diferentes temas. Em cada ciclo, é demonstrado que a solução se apresentou muito útil.

No primeiro ciclo, foi realizada uma revisão literária sobre segurança da informação. No segundo ciclo, a aplicação foi uma pesquisa sobre o uso de WhatsApp como prova judicial. O terceiro ciclo, foi a aplicação da revisão literária desta tese.

Conforme é recomendado no método pesquisa-ação, o pesquisador tem dois papéis: um profissional que executa as atividades dentro do contexto da ação em andamento e um pesquisador que analisa os resultados e propõe melhorias nas atividades que estão sendo realizadas. Por essa razão, às vezes é difícil realizar a separação de questões de cunho profissional, como a identificação de requisitos para o software, e questões acadêmicas de análise da eficácia do algoritmo, por exemplo.

6.3 CICLO 1 – RSL NA PESQUISA SOBRE SEGURANÇA DA INFORMAÇÃO

Este primeiro ciclo teve como principal objetivo testar o LDA analisando resumos de publicações para identificar se este algoritmo é adequado para a realização da revisão literária. Após o desenvolvimento do código, ainda bastante elementar, uma revisão da literatura com tema da pesquisa da Segurança da Informação foi executada.

As questões específicas do Ponto de Análise PA2 – Necessidade Programação² a serem atendidas foram:

- Receber arquivos com resultado de busca de quaisquer BDs e em grande volume.
- Definir as melhores práticas para o tratamento de textos e termos. (Parcial)
- Armazenar termos de exclusão identificados pelo usuário

A solução foi desenvolvida em Python e executada no prompt do DOS. Recebeu como entrada a planilha “Papers.xlsx” contendo os metadados de 604 artigos baixados da BD WoS e utilizou a coluna de resumos. A limpeza se deu utilizando um arquivo denominado “BlockWords.txt” com as *Stopwords* do usuário como na Figura 17 e o arquivo com as *Stopwords* padrão da biblioteca “NLTK” (Sergio et al, 2019) .

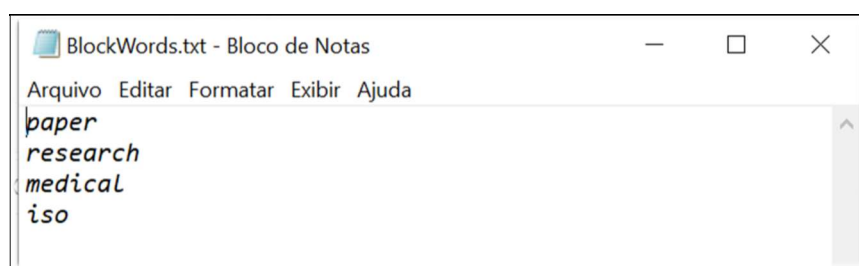


Figura 17 - Exemplo do arquivo TXT com a lista de Stopwords do usuário

Para a lematização dos textos, foi utilizado o “*WordNetLemmatizer*” da biblioteca “NLTK”. O resultado foi exibido no console do DOS com os tópicos, palavras chave e seu peso no corpus.

Na segunda versão, os resultados foram disponibilizados numa planilha CSV como na Figura 18. Para o desenvolvimento, foi utilizada a plataforma Kivy.

² PA2 – Ponto de Análise 2 descrito no item 3.5.1

Topic	Keywords (Weights)
1	Topic_0_%0_word,1_%1_word,2_%2_word,3_%3_word,4_%4_word,5_%5_word,6_%6_word,7_%7_word,8_%8_word,9_%9_word
2	(0, 0.027, security, 0.022, informationsecurity, 0.014, system, 0.014, management, 0.010, organization, 0.009, standard, 0.008, control, 0.007, informationtechnology, 0.007, information, 0.006, risk)
3	(1, 0.023, informationsecurity, 0.014, informationtechnology, 0.011, security, 0.010, project, 0.009, used, 0.008, information, 0.008, business, 0.007, level, 0.007, factor, 0.007, system)
4	(2, 0.015, informationtechnology, 0.014, document, 0.012, system, 0.011, informationsecurity, 0.011, industrial, 0.011, smbs, 0.009, leakage, 0.008, security, 0.008, availability, 0.006, technique)
5	(3, 0.025, informationsecurity, 0.020, informationtechnology, 0.015, standard, 0.013, security, 0.009, framework, 0.008, management, 0.007, control, 0.007, technology, 0.007, business, 0.006, telemedicine)
6	(4, 0.016, informationtechnology, 0.015, security, 0.015, model, 0.012, organization, 0.011, risk, 0.010, management, 0.010, project, 0.009, incident, 0.008, process, 0.008, effort)
7	(5, 0.021, security, 0.021, informationsecurity, 0.018, framework, 0.016, informationtechnology, 0.016, risk, 0.010, management, 0.010, internet, 0.009, method, 0.009, thing, 0.008, organization)
8	(6, 0.018, user, 0.014, informationsecurity, 0.013, information, 0.013, personal, 0.010, informationtechnology, 0.010, risk, 0.009, protection, 0.009, standardization, 0.008, process, 0.008, management)
9	(7, 0.029, security, 0.027, information, 0.018, management, 0.015, industry, 0.013, approach, 0.009, risk, 0.008, practice, 0.008, monitoring, 0.007, cloud, 0.007, business)
10	(8, 0.026, informationsecurity, 0.022, security, 0.017, egovernment, 0.017, system, 0.015, management, 0.014, informationtechnology, 0.011, analysis, 0.010, model, 0.010, risk, 0.009, organization)
11	(9, 0.022, security, 0.012, system, 0.009, configuration, 0.007, management, 0.007, informationtechnology, 0.006, user, 0.006, execution, 0.006, task, 0.006, universe, 0.005, computing)

Figura 18 - Planilha com tópicos, palavras-chave e seu peso no corpus

O resultado foi copiado para uma planilha e formatado para melhor visualização, como se observa na Figura 19. As nove palavras-chave de cada um dos 15 tópicos foram organizadas, permitindo um melhor entendimento.

TABLE II
IDENTIFIED THEMES AND TERMS

information	security	management	system	organization	process	assessment	risk	requirement
information	security	management	system	application	framework	logistics	27001	ecommerce
information	security	management	system	organization	model	flaw	enterprise	banking
information	security	management	company	organization	model	framework	27001	certification
information	security	management	model	organization	control	implementation	process	document
information	security	management	system	requirement	policy	control	27001	function
information	management	system	construct	organization	relationship	network	analysis	concept
information	security	analysis	process	organization	performance	supply	project	chain
information	security	sms	process	company	environment	implementation	business	however
information	security	system	27001	method	document	computing	cloud	requirement
information	security	enterprise	organization	method	level	current	asset	proposed
information	security	model	threat	framework	internal	centre	data	element
system	risk	method	organization	implementation	analysis	assessment	factor	level
system	quantity	organization	safety	protection	requirement	enterprise	27001	approach
data	risk	forensic	analysis	investment	value	must	digital	chain

Figura 19 - 15 tópicos da RSL de segurança da informação (Sergio et al, 2019)

Diversas rodadas foram executadas variando o número de tópicos e número de palavras-chave até que os tópicos revelaram temas para os pesquisadores. As palavras-chave que se repetiam na grande maioria dos tópicos, como *Information*, *security*, *management*, *ISO 27001* e *organization*, tiveram pouca contribuição para revelar temas. As demais palavras-chave, embora ainda tendo um grau de repetição, sinalizavam pequenas variações de temas como as palavras-chave *system*, *risk*, *requirement*, *policy*, *control*, *framework* e *approach* (Sergio et al, 2019). Na avaliação do ciclo, tem-se que as questões específicas da PA2 foram atendidas como sintetiza a Tabela 12 e o objetivo principal teve sucesso pois, o LDA foi considerado uma escolha adequada.

Tabela 12 - Questões específicas da PA atendidas

PA 2 – Necessidade de programação		
Questão específica	Observação	Situação Final
Receber arquivos com resultado de busca de quaisquer BDs e em grande volume.	Testado somente com arquivos do WoS com 604 metadados	OK
Definir as melhores práticas para tratamento de textos e termos.	LDA mostrou-se adequado e o tratamento do texto utilizou o arquivo padrão	OK
Armazenar termos de exclusão ou de troca identificados pelo usuário	Foi criado um arquivo de <i>Stopwords</i> do usuário	OK

O ciclo foi concluído com a identificação de novos pontos de análise e questões específicas a serem atendidas no ciclo 2, assim como novos pontos de análise que derivaram da execução da segunda revisão literária ocorrida em 2021, quando todas as questões específicas foram definidas. Reproduzida a estrutura da pesquisa, tem-se os seguintes pontos de análise:

PA 1 – Seleção do processo

- Criar solução que atenda pesquisadores, independentemente de área do conhecimento, do perfil tecnológico e da disponibilidade de acesso a BDs indexadas.
- Manter registro do histórico de execução da RSL para subsidiar o relatório final e garantir a rastreabilidade, transparência e confiabilidade do realizado.

PA 2 – Necessidade de programação

- Receber arquivos com resultado de busca de quaisquer BDs e em grande volume.
- Definir as melhores práticas para o tratamento de textos e termos.
- Armazenar termos de exclusão ou de troca identificados pelo usuário.

PA 3 – Motivo para adoção do processo

- Dar autonomia para estabelecer o melhor resultado de agrupamento em tópicos, atendendo objetivos específicos e permitindo análise parcial (ano de publicação).
- Definir métrica de validação de resultado da análise de tópicos
- Possibilitar a integração da ferramenta com outras ferramentas complementares de interesse de uma RSL.

Como resultado deste ciclo, houve a publicação de um pôster de Iniciação científica e a indicação da bolsista para concorrer ao prêmio de melhor iniciação científica de 2018 ambas reproduzidas no Apêndice 2 – Pôster PIBIC 2018, e indicação da bolsista para o prêmio. Foi apresentado um artigo sobre esta experiência no congresso da IEEE de 2019, cuja cópia da primeira folha está no Apêndice 3 – artigo publicado em congresso no IEEE. O arquivo completo pode-se obter acessando o repositório do IEEE.

6.4 CICLO 2 – RSL NA PESQUISA SOBRE WHATSAPP COMO PROVA JUDICIAL

O Sw3T foi aperfeiçoado incorporando as funcionalidades identificadas no ciclo 1 e, após a implementação, ficou pronto para aplicação em um novo ciclo. O ciclo 2 teve como principal objetivo desenvolver o processo de revisão literária com o Sw3T. Este teste foi realizado em dezembro de 2021 como uma revisão literária de outro tema.

A ferramenta foi desenvolvida na plataforma XAMPP utilizando Python, PHP, HTML, CSS e banco de dados MySQL. O ciclo foi realizado com um arquivo contendo metadados da revisão literária com o tema “WhatsApp e evidências jurídicas” para o TCC de um bolsista da equipe. O teste da solução foi bem-sucedido, e foram recebidas 134 publicações oriundas das bibliotecas WoS e IEEE.

Foram realizadas várias análises e a que apresentou melhor resultado orientou a seleção conforme visualizada na Figura 20.



Figura 20 - Tela do Sw3T com a melhor análise para o tema WhatsApp

Foram criados rótulos em forma de frases para cada tópico refletir o tema dos resumos daquele Cluster. São exemplificados a seguir, três rótulos decidindo-se que o “item a” e o “item c” continham os resumos de interesse para a pesquisa sobre WhatsApp.

- a) *Forense, message, digital, communication, data*: Análise forense recupera informações da mensagem digital.
- b) *Social, médium, method, vídeo, digital*: plataformas de mídia social usam vídeos para comunicação
- c) *Message, data, legal, apps*: mensagens de aplicativos com dados analisados em conformidade com os requisitos legais.

No final da seleção por título e resumo foram obtidos 17 artigos para serem recuperados no formato PDF para leitura. A Figura 21 mostra as informações para o PRISMA desta revisão literária.

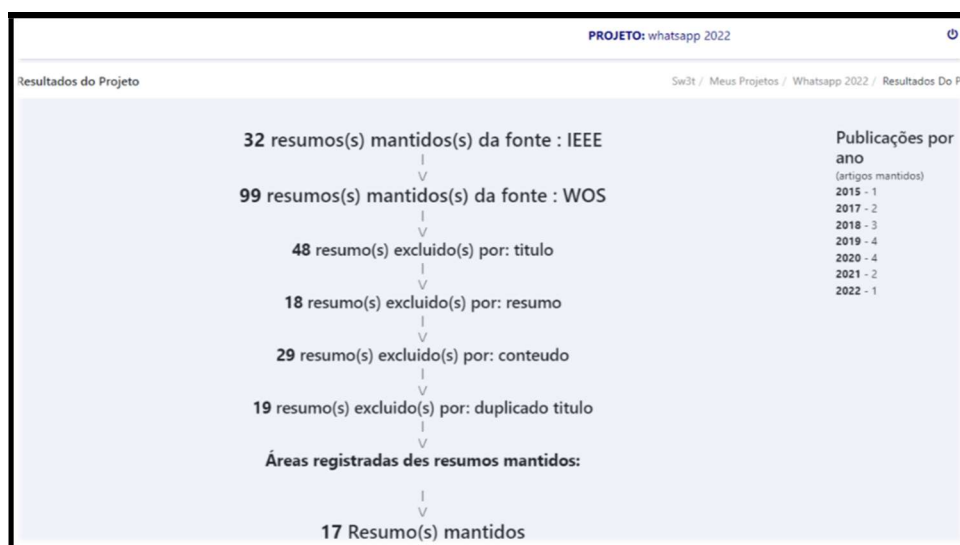


Figura 21 - Tela do Sw3T com histórico de seleção do Ciclo 2

O ciclo foi concluído, e na avaliação das questões específicas, constatou-se que algumas foram parcialmente atendidas. O resultado da validação é observado na Tabela 13.

Tabela 13 - Resultado da análise das questões específicas do ciclo 2

PA 1 – Seleção do processo		
Questões específicas	Observação	situação
Criar solução que atenda pesquisadores, independente de área do conhecimento, do perfil tecnológico e da disponibilidade de acesso a BDs indexadas.	Não foi possível analisar essa questão pois a revisão literária foi realizada por um programador do Sw3T	não
Manter registro do histórico de execução da RSL para subsidiar o relatório final e garantir a rastreabilidade, transparência e confiabilidade do realizado.	Foi incorporado na aplicação	OK
PA 2 – Necessidade de programação		
Questões específicas	Observação	situação
Receber arquivos com resultado de busca de quaisquer BDs e em grande volume.	Foi testado em duas bases de dados WoS e IEEE	OK
Definir as melhores práticas para tratamento de textos e termos.	Identificado o algoritmo N-GRAN para criação de termos compostos	parcial
Armazenar termos de exclusão ou de troca identificados pelo usuário	Foram criadas as tabelas Stopwords e De/Para no banco de dados	OK
PA 3 – Motivo para adoção do processo		
Questões específicas	Observação	situação
Dar autonomia para estabelecer o melhor resultado de agrupamento em tópicos, atendendo objetivo específicos e permitindo análise parcial (ano de publicação).	O processo permite que o usuário defina os parâmetros para realização da análise (número de tópicos e palavras-chave por exemplo)	OK
Definir métrica de validação de resultado da análise de tópicos	Definidas as métricas: acurácia, afirmação e dominância	OK
Possibilitar a integração da ferramenta com outras ferramentas complementares de interesse de uma RSL	Criada uma rotina de exportação dos metadados em formato de planilha	OK

As métricas da análise foram reavaliadas quanto à facilidade de interpretação na tela, e optou-se por inverter o valor do indicador *perplexidade*, que quanto menor melhor, criando-se a métrica *afirmação*, onde quanto maior melhor. Isto é, se a perplexidade é de 20, a afirmação é 80.

O ciclo foi concluído identificando as questões específicas dos pontos de análise a serem observados no ciclo 3, sendo:

- Manter registro do histórico de execução da RSL para subsidiar o relatório final e garantir a rastreabilidade, transparência e confiabilidade do realizado.
- Armazenar termos de exclusão ou de troca identificados pelo usuário
- Definir métrica de validação de resultado da análise de tópicos
- Possibilitar a integração da ferramenta com outras ferramentas complementares de interesse de uma RSL.
- Emitir informações sobre o projeto.

Como produto deste ciclo 2, tem-se o TCC submetido pelo bolsista à FT – UNICAMP onde a pesquisadora teve a oportunidade de fazer parte da banca. Parte

do TCC que se refere ao uso do Sw3T se encontra no Apêndice 4 – Trecho do TCC com uso Sw3T em 2022. O TCC completo se encontra na biblioteca da FT UNICAMP.

6.5 CICLO 3 – RSL NA PESQUISA SOBRE RSL COM MODELAGEM DE TÓPICOS

O ciclo 3 teve como principal objetivo implementar algumas melhorias na ferramenta para aperfeiçoar o novo processo de RSL com Sw3T. As melhorias foram realizadas atendendo às questões específicas observadas para o ciclo 2. Após efetivadas as melhorias previstas para este ciclo, um novo teste foi realizado. A RSL ocorreu em dezembro de 2022 com o tema “Revisão da Literatura e Modelagem de tópicos”. A busca foi realizada nas BDs WoS, IEEE, Scopus, ACM, Scopus e Google sem restrição de ano. A Figura 22 exibe os totais de metadados baixados, instância de seleção e publicações selecionadas.

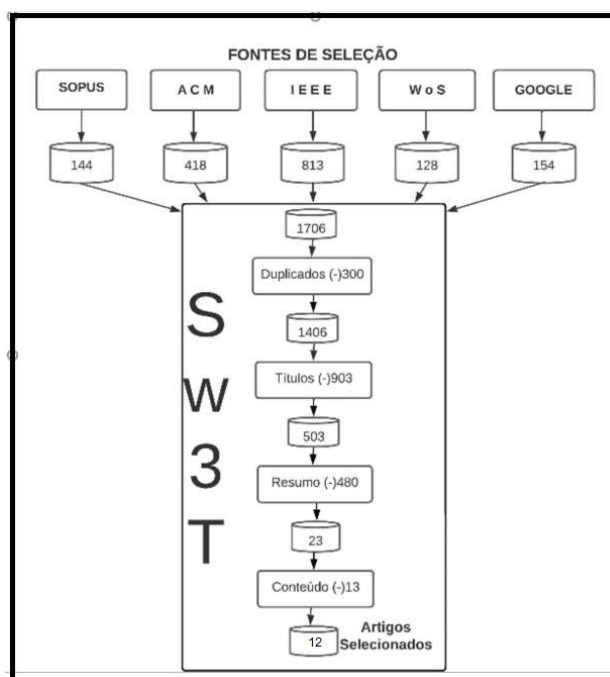


Figura 22 - Prisma do terceiro ciclo da pesquisa-ação

O relato desta revisão literária repete as mesmas condições da realizada na revisão realizada em 2021 e descrito no capítulo 2. A análise das publicações contemplou o resultado das três revisões literárias. O total de publicações relevantes é 12, pois inclui as 10 já identificadas na revisão do ciclo 2 e mais duas novas identificadas nesta revisão.

Foram criados tópicos para 3 períodos:

- [3] 2020 até 2023: Foram analisadas 102 publicações e obteve-se 6 tópicos. O gráfico contendo o resultado deste ciclo, está na Figura 25.



Figura 25 – Tela do Sw3T com 6 Tópicos para 102 publicações entre 2020 e 2023

Os seis termos que representam os seis tópicos são relacionados a seguir. Uma forma de deixar claro o tema dos tópicos é transformar os termos numa frase representativa. A seguir são apresentadas frases combinando os 6 termos que compõem cada tópico sugerido pelo ChatGPT (OpenAI, 2023), seguida da quantidade de publicações relacionadas ao tema. Sendo:

- (i) text, system, propose, result, classify, technology
The proposed technology uses a system to classify text and generate results (114).
- (ii) system, information, user, result, article, text
The system provides users with information and results based on the text of an article (125).
- (iii) result, bug, report, identify, system, propose
Propose a system to get results in bug reports (82).
- (iv) propose, result, feature, user, system, software
The proposed software is designed to improve the user experience (95).

- (v) provide, text, Information, process, field, result
The system can provide valuable information in various fields, processing text, and generating results (80).
- (vi) System, Propose, Result, Identify, Information, text
By identifying and proposing solutions based on text information, the system can produce accurate results (117).

Observa-se que o tópico (ii) que tem como rótulo prover informações resultantes da análise de artigos é o que concentra o maior número de publicações. O que é esperado pela *string* de busca utilizada é seguido do item (i), que aborda a classificação de texto e ao (vi), com tema de prover soluções para informações de textos. O tópico (iv) é o uso do LDA na era da programação e o (v), na análise de texto em geral.

O ciclo 3 foi concluído com sucesso uma vez que as questões específicas de cada ponto de análise também foram atendidas, conforme Tabela 14.

Tabela 14 - Resultado da análise das questões específicas do ciclo 3

PA 1 – Seleção do processo		
Questões específicas	Observação	situação
Criar solução que atenda pesquisadores, independente de área do conhecimento, do perfil tecnológico e da disponibilidade de acesso a BDs indexadas.	Não foi possível analisar essa questão pois a revisão literária foi realizada por um programador do Sw3T	Ok
PA 2 – Necessidade de programação		
Questões específicas	Observação	situação
Definir as melhores práticas para tratamento de textos e termos.	Identificado o algoritmo N-GRAN para criação de termos compostos	Ok

Vale observar na Figura 2 que o diagrama PRISMA revela que foram 1332 publicações iniciais. Diversas funções do Sw3T foram executadas pela pesquisadora para eliminar duplicados, excluir publicações por palavras-chave adicionais, por título e por resumo, registrando o motivo da exclusão e inclusão, motivo e comentários.

Nessa etapa, utilizam-se recursos do aplicativo para agrupar as publicações por tópico. Utilizando a classificação por tópico, a seleção ficou mais harmônica possibilitando a análise sequencial de assuntos semelhantes. Conforme as publicações foram sendo excluídas, novos tópicos foram gerados reorganizando os resumos ainda em análise, contribuindo a pesquisadora com o refinamento dos tópicos.

Após as análises por título e resumo com o apoio do Sw3T, observa-se na Figura 2 que 18 artigos foram analisados com profundidade consultando sua versão completa em PDF. Os artigos eliminados nesta última análise, também tiveram a justificativa registrada no Sw3T, garantindo a rastreabilidade do processo. Ao final da análise por conteúdo, obteve-se 10 publicações com conteúdo relevante. A rigor, ficaram 9 publicações, pois, o artigo publicado pela pesquisadora, também foi um dos selecionados.

6.6 EFICÁCIA DA FERRAMENTA NUMA RSL

Procurou-se, durante o andamento desta pesquisa, realizar revisão literária com pesquisadores que não tiveram nenhum contato com o desenvolvimento deste trabalho, mas infelizmente, por questões de prazo, não houve tempo de terminar a aplicação para transformar essas em ciclos da pesquisa-ação. No entanto, dois pesquisadores se dispuseram a utilizar a ferramenta em áreas diferentes.

6.6.1 Revisão literária de cidades tropicais

A primeira delas trata de uma pesquisa sobre **cidades tropicais** e o objetivo era criar uma taxonomia sobre esse termo. São realizadas pesquisas de áreas de conhecimento muito diferentes como saúde, política, educação e infraestrutura. O pesquisador se propôs a utilizar a ferramenta e fez a seguinte declaração:

“A ferramenta foi aplicada com o intuito de testar uma abordagem alternativa na construção de uma taxonomia sobre o termo “Cidades Tropicais” a partir de uma busca nas bases de referência citadas. A chave de busca era o termo amplo com pequenas variações. O resultado da busca gerou um conjunto relativamente pequeno de publicações, na ordem de 800 referências, mas com grande variabilidade no que se refere às áreas de conhecimento e metodologias de pesquisa e aplicação.”

O pesquisador identificou um universo de metadados de 800 artigos e foi selecionando e identificando os tópicos. O objetivo era, para criar a taxonomia, identificar os tópicos que seriam os elementos da classificação.

Sobre o uso do Sw3T, declarou:

“A ferramenta Sw3T permite uma grande flexibilidade de aplicação de estratégias de tratamento de documentos para a construção de análises de relacionamentos. A minha experiência pessoal com a ferramenta foi de desafio na construção de protocolos de interpretação dos resultados. É uma etapa de desenvolvimento das aplicações da ferramenta e da própria pesquisa de construção de taxonomia em condições em que o tema é amplo e com a presença de diferentes áreas de conhecimento e distintas abordagens metodológicas.

A ferramenta Sw3T mostrou-se interessante para desafiar os pesquisadores a explorarem as suas possibilidades na construção de relacionamentos com base no algoritmo LDA para diferentes aplicações.”

Esta aplicação motivou a inclusão, no Sw3T do campo “área de conhecimento” para permitir a realização de análises focadas por área de conhecimento.

6.6.2 Revisão literária de uma pesquisadora de doutorado

A pesquisadora é doutoranda no INPE e é professora em uma faculdade estadual, utilizou o Sw3T e deu o seguinte parecer:

“Eu estou gostando muito da sua ferramenta. A separação dos grupos após a análise ajuda a encontrar grupos que não tenham a ver com o tema, então nos concentramos apenas na leitura do título e conseguimos excluir a maioria dos artigos rapidamente sem perder o foco, alternando entre a leitura do título e resumo.

O algoritmo LDA ainda converge lentamente para o resultado. Cada grupo fornecido pela análise possuía de 50% a 66% de artigos que eu havia incluído na revisão sistemática inicial realizada de forma tradicional.

Ainda não cheguei ao final da separação dos artigos no Sw3T, até agora demorei por volta de 3h50m (conforme o histórico do Sw3T). Não medi o tempo que demorei na primeira revisão (sem o Sw3T), mas tenho sensação de que estou convergindo para a lista de artigos incluídos de forma mais rápida.”

6.6.3 Considerações sobre a questão específica do PA1

A questão do PA1 que trata do uso deste processo por pesquisadores independentes para avaliar o processo e a usabilidade do Sw3T não pode ser

verificada dentro dos ciclos da pesquisa ação. No entanto, para cobrir esta questão, essas duas aplicações sinalizam que é possível atender a esta questão também.

6.7 CONSIDERAÇÕES FINAIS

Neste capítulo foram descritas a aplicação do RSL com o Sw3T em 3 ciclos de uma pesquisa ação onde cada um deles apresentou uma aplicação diferente. Nos 3 ciclos o processo demonstrou ser bastante eficaz e foi evoluindo ao longo do tempo, com a identificação de melhorias e sua implementação nos ciclos seguintes.

No próximo capítulo, serão retomadas as perguntas de pesquisa e feito o fechamento da pesquisa com a verificação de cumprimento de seus objetivos.

7. AVALIAÇÃO DO PROCESSO RSL COM SW3T

7.1 INTRODUÇÃO

Este capítulo faz uma avaliação de todo o projeto sob vários ângulos, retomando a pesquisa na Fase 5 que apresenta os resultados da pesquisa-ação. São retomados os pontos da pesquisa para dar consistência e fechamento ao trabalho. São feitas reflexões do pesquisador sobre a experiência vivenciada na execução deste trabalho e conclui sinalizando uma continuidade e oportunidades para futuras pesquisas.

7.2 FASE 5 – APRESENTAÇÃO DE RESULTADOS

Esta fase da pesquisa trata da apresentação de resultados e da entrega do documento formal de pesquisa. Neste caso, o entregável é este documento contendo todo o relato da pesquisa no formato de tese.

Coughlan e Coughlan (2002) apontam que a pesquisa-ação tem dois objetivos: solucionar um problema real (objetivo prático) e contribuir para a ciência (objetivo de conhecimento).

Com relação ao objetivo prático foi desenvolvido um novo processo RSL com Sw3T que se apresenta como uma solução para o problema de complexidade e demora na realização de uma RSL e foi utilizado na realização de cinco RSLs diferentes.

Com relação ao objetivo de conhecimento, foi comprovada a oportunidade de realização de RSL utilizando o LDA e outras ferramentas de PLN com todos os cuidados acadêmicos em seu desenvolvimento, conforme descrito neste documento de pesquisa.

7.3 QUANTO AO ATENDIMENTO À PROPOSTA DE PESQUISA

Conforme descrito na estruturação da pesquisa (Capítulo 3), uma vez estabelecido o objetivo geral e a questão da pesquisa, foram feitas a derivação em objetivos específicos (OE) e estes em proposições (P) que, por sua vez, foram desdobradas em pontos de análise (PA) e questões específicas. Assim, o fechamento deste trabalho é feito no sentido inverso conforme esquematizado na Figura 26.

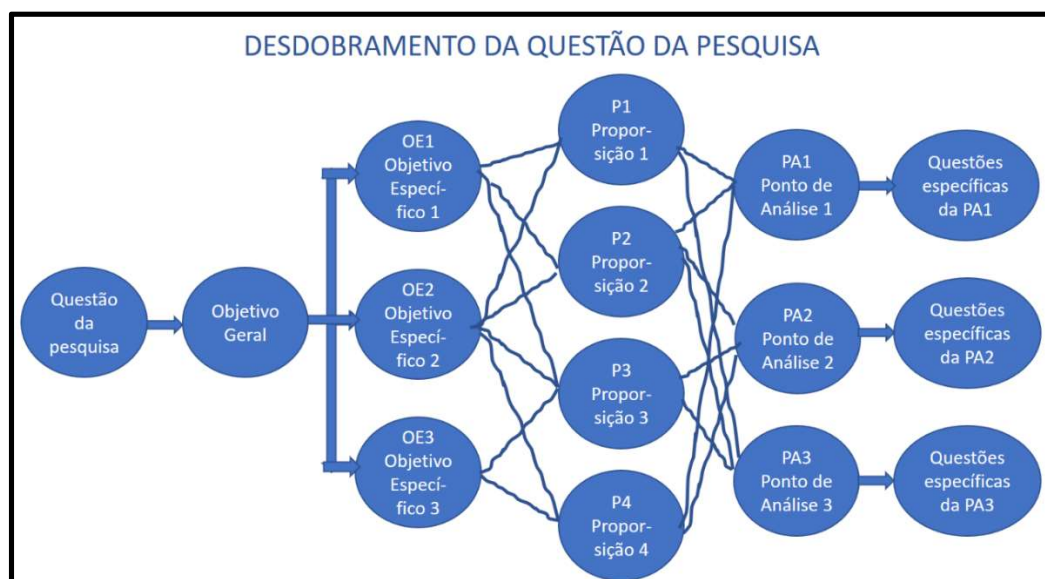


Figura 26 - Desdobramento da questão da pesquisa

O atendimento às questões específicas de cada ponto de análise é acompanhado no Capítulo 6, onde a cada ciclo as questões específicas são apontadas, executadas e avaliadas. Como consequência, os três pontos de análise a que elas atendem também estão atendidos.

Os pontos de análise conjugam uma relação com as proposições, sendo que cada ponto de análise atende a mais de uma proposição e uma proposição é atendida por mais de um ponto de análise, conforme estabelecido na Tabela 8 - Ponto de Análise PA1 - Seleção do Processo, Tabela 9 - Ponto de Análise PA2 - Necessidade de Programação e Tabela 10 - Ponto de Análise PA3 - Motivo para adoção do processo. Assim, uma vez que todos os pontos de análise foram atendidos, as proposições também foram atendidas. Observando a Figura 26, tem-se as proposições derivadas dos objetivos específicos em uma conjugação N para N. Assim os objetivos específicos são atendidos à medida que as proposições a ele relacionadas são atendidas. A seguir as proposições são detalhadas apresentando a relação com os objetivos e os pontos de análise.

P1 - Em soluções de RSL que usam técnicas de PLN, existem lacunas quanto à melhor forma de implementar, atendendo aos requisitos de qualidade: Rapidez, Transparência, Confiabilidade e Rastreabilidade (PA1, PA3 - OE1, OE2, OE3)

Esta proposição foi atendida e preencheu essa lacuna atendendo os requisitos de qualidade de uma RSL. *Rapidez* - o processo demonstrou ser mais rápido do que o processo tradicional. *Transparência, confiabilidade e rastreabilidade* -atendidas pelo fato de manter o registro de todo o histórico de seleção

P2 - É possível encontrar uma solução que respeite a liberdade de condução por parte do pesquisador, contemplando a atuação em diferentes áreas de conhecimento e diferentes habilidades com o uso de tecnologias (PA1- OE2)

O processo desenvolvido permite que o pesquisador realize a RSL na sequência e na qualidade que melhor lhe aprouver, por exemplo existe a liberdade de definir o número de tópicos, o número de palavras-chave e o período a ser analisado.

P3 - Os Motivos para a comunidade acadêmica/ científica aderir à nova solução de RSL estão relacionados à simplicidade de execução, respeito à autonomia do pesquisador na condução da sua pesquisa e fornecimento de informações consistentes para o relato da RSL (PA1 PA3 - OE1, OE2, OE3).

Embora este processo de RSL não tenha ainda sido utilizado por grande quantidade de pesquisadores, as atividades do processo são simples e podem ser executadas em qualquer ordem, além de oferecer uma grande quantidade de informações para os relatórios de pesquisa.

P4 - É possível implementar uma solução que amplie a consulta às BDs existentes, que o acesso possa ser compartilhado por pesquisadores e que reduza e postergue a necessidade de analisar artigos completos. (PA2 - OE3).

O processo permite a utilização de diversos BDs e realiza a integração das informações. Também permite o compartilhamento do projeto entre os diversos pesquisadores. O trabalho realizado com títulos e resumos permite a postergação da necessidade de analisar artigos completos.

Retomando os objetivos específicos:

OE1. Analisar ferramentas disponíveis para revisão literária com uso de técnicas de PLN.

OE2. Analisar o processo de revisão literária de uma pesquisa científica da área de tecnologia

OE3. Avaliar técnicas de PLN quanto à usabilidade e aplicabilidade em revisão literária

Com relação às ferramentas existentes para RSL que usam PLN, foram encontradas poucas ferramentas com funcionalidades pontuais e nem todas estão disponíveis para uso pela comunidade. O processo de revisão literária escolhido foi o de Kitchenham (2013) por atender à engenharia de software. As técnicas de PLN selecionadas foram utilizadas com sucesso na construção do Sw3T.

Com estas colocações sobre as proposições, citando seu papel no atendimento aos Objetivos Específicos, tem-se que estes também foram atendidos.

Da mesma forma, sendo atendidos os objetivos específicos, o objetivo geral está atendido, pois foi a definição do Processo RSL com Sw3T aderente às três fases propostas por Kitchenham (2013). Finalmente a questão da pesquisa está respondida uma vez que técnicas de PLN foram utilizadas satisfatoriamente no Sw3T.

7.4 REFLEXÕES SOBRE A EXPERIÊNCIA VIVENCIADA

Alguns fatores interferiram na conclusão desta pesquisa, como a epidemia de COVID que impossibilitou o trabalho presencial com os bolsistas, o longo tempo de pesquisa levou à troca de bolsistas que se formaram nesse período e aspectos pessoais de saúde do orientador e da orientanda. Porém, a utilização do método de pesquisa-ação facilitou a retomada do projeto em 2021, uma vez que novos pontos de análise puderam ser identificados e o coorientador se propôs a tomar frente da orientação.

Uma característica positiva do Sw3T é oferecer um ambiente compartilhado para trabalhos em equipe com vários pesquisadores em um mesmo projeto. Também suporta a existência de diversos projetos isolados entre si.

Entendemos que o Sw3T poderá ser disponibilizado em plataforma gratuita para permitir que a comunidade acadêmica possa usufruir dessa ferramenta. Para tanto, serão tomadas as providências para registro do software para garantir a propriedade intelectual.

Esta pesquisa também foi valiosa para membros da equipe que participou do trabalho. Suas contribuições foram preciosas para o estudo uma vez que, também foram executores de determinados ciclos previstos na estrutura da pesquisa. Essa participação trouxe benefícios individuais a cada um deles, pois desfrutaram dos resultados das revisões literárias executadas, usando o novo processo em seus projetos de pesquisa vinculados à bolsa e a seus TCCs universitários. Vale dizer que um dos trabalhos de iniciação científica concorreu ao prêmio de melhor iniciação científica de 2018 promovido pelo CTI Renato Archer.

7.5 CONCLUSÕES E PESQUISAS FUTURAS

O resultado desta pesquisa deve contribuir com a instituição de pesquisa na área de Tecnologia da Informação, onde foi desenvolvida. A instituição está organizada em laboratórios e conta com equipes de tecnologistas, pesquisadores e bolsistas onde muitos projetos de pesquisa são realizados. Há uma certa rotatividade destes profissionais, e o novo processo de RSL com o Sw3T pode ser utilizado em seus diversos projetos, agilizando a execução e contribuindo para o rápido

aprendizado desta atividade pelos novos membros da equipe, resultando em revisões ágeis, consistentes, confiáveis e rastreáveis.

A pesquisa apresenta ainda como ponto forte o fato de que os ciclos foram realizados em diferentes áreas de conhecimento e com diferentes atores, enriquecendo a abrangência da aplicação do processo de revisão literária. A continuidade desta pesquisa deve incluir a conclusão das avaliações que estão sendo realizadas por pesquisadores de fora do projeto a fim de corroborar com a usabilidade da ferramenta.

Pretende-se disponibilizar o software Sw3T para a comunidade assim que os aspectos formais de propriedade forem concluídos. Com relação à evolução da ferramenta Sw3T, poderá ser refinada a interface com o usuário, acrescidas funcionalidades para aumentar a automação e incorporadas novas pesquisas de algoritmos e ferramentas que possam vir a apoiar os pesquisadores.

Novos estudos sobre tecnologias de PLN e de modelagem de tópicos que estão surgindo devem ser desenvolvidos buscando aprimorar as funções do software. A maioria dos estudos sobre PLN utilizou o modelo de tópicos LDA, mas cabe analisar tendências de pesquisa e comparar desempenhos de diferentes modelos de tópicos, como STM.

Um dos pontos identificados como trabalho futuro é a incorporação tecnológica de realidade aumentada para visualização da relação entre tópicos e entre clusters para auxiliar na análise e seleção das publicações em representações tridimensionais.

REFERÊNCIAS

- (Albanese, 2022) ALBANESE, N Cosimo. 2022. **Topic modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: a comparison**. Publicado em 19/09/2022. Disponível em: <https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5>. Visit em: 15/abr/2023.
- (Alnusyan et al., 2020) ALNUSYAN, R., ALMOTAIRI, R., ALMUFADHI, S., SHARGABI, A. A., & ALSHOBAILI, J. 2020. **A semi-supervised approach for user reviews topic modeling and classification**. In International Conference on Computing and Information Technology (ICCIT-1441) (pp. 1-5), IEEE.
- (Asmussen&Møller, 2019) ASMUSSEN, Claus Boye; MØLLER, Charles. 2019. **Smart literature review: a practical topic modelling approach to exploratory literature review**. *Journal of Big Data*, v. 6, n. 1, p. 1-18.
- (Bishop, 2006) BISHOP, C. (2006). **Pattern Recognition and Machine Learning**. [S. l.]: Springer. Available at: <https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/>.
- (Blei&Lafferty, 2007) BLEI, David M.; LAFFERTY, John D. (2007). **A correlated topic model of science**. 2007.
- (Bryman, 2006) Bryman, A. (2006). **Integrating quantitative and qualitative research: how is it done?** *Qualitative research*, 6 (1), 97-113.
- (Baskerville&Wood-Harper, 1996) BASKERVILLE, Richard & WOOD-HARPER, Trevor. 1996. **A Critical Perspective on Action Research as a Method for Information Systems Research**. *Journal of Information Technology*. 11. 235-246. 10.1080/026839696345289.
- (Bechhofer, 2004) BECHHOFFER, Frank. 2004. **Qualitative data and The Affluent Worker study: a missed opportunity?** *International Journal of Social Research Methodology*, v. 7, n. 1, p. 45-50.
- (Cardoso&Alarcão&Celorico, 2010) CARDOSO, Teresa; ALARCÃO, Isabel; CELORICO, Jacinto Antunes. 2010. **Revisão da literatura e sistematização do conhecimento**. Porto Editora.
- (Carrara, 2011) CARRARA, André Ramos. 2011. **Implantação de sistema BPMS para a gestão por processos: uma análise crítica**. PhD Thesis. Universidade de São Paulo.
- (Chakkarwar&Tamane, 2019). CHAKKARWAR, Vrishali; TAMANE, Sharvari C. 2019. **Quick insight of research literature using topic modeling**. In: *Smart Trends in Computing and Communications: Proceedings of SmartCom*. P189-197.
- (Coghlan&Brannick, 2005) COGHLAN, D; BRANNICK, T. 2005. **Doing Action Research in Your Own Organization**. [S. l.]: Second. Edition, SAGE Publications, Available at: <https://books.google.com.br/books?id=qrQBwAEACAAJ>.
- (Coughlan&Coughlan, 2002) COUGHLAN, Paul; COUGHLAN, David. 2002. **Action research for operations management**. *International journal of operations & production management*, 2002, v.22, n 2, p. 148-151.
- (Croom, 2002) CROOM, S. 2002. **Methodology Editorial**. *International Journal of Operations and Production Management*, v22, n.2 p. 148-151.

- (Davies, 2007) DAVIES, P. 2007. **Revisões sistemáticas e a Campbell Collaboration**. In: THOMAS, G. et al. Educação baseada em evidências: atualização dos achados científicos para a qualificação da prática pedagógica. Porto Alegre: Artmed, p. 31-43.
- (De Almeida, 2014) DE ALMEIDA, André Pinho. **Classificação e agregação automática de notícias desportivas**. 2014. PhD Thesis. Instituto Politécnico do Porto (Portugal).
- (Deerwester et al., 1990) Deerwester S, Ddumais S T, Furnas G W, et al. 1990. **Indexing by latent semantic analysis**. *Journal of the American Society for Information Science*, 41(6): 391-407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9)
- (Dudziak, 2021) DUDZIAK, Elisabeth. 2021. **O que é literatura cinzenta?** AGUIA Blog, 16 ago. 2021. Disponível em: <https://www.aguia.usp.br/noticias/o-que-e-literatura-cinzenta/> Acesso em: 15/março/2023.
- (Durkin&Durkin, 1998) DURKIN, J., & DURKIN, J. 1998. *Expert systems: design and development*. Prentice Hall PTR. (Echevarria, 2014) Echevarria IM, Walker S. 2014. **To make your case, start with a PICOT question**. *Nursing*; 44(2):18-9. Doi: 10.1097/01.NURSE.0000442594.00242.f9. PMID: 24430384.
- (Egger, 2022) EGGER, Roman. 2022. **Text Representations and Word Embeddings: Vectorizing Textual Data**. In: *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*. Cham: Springer International Publishing, 2022. p. 335-361.
- (Ellmann, 2018) ELLMANN, Mathias. 2018. **Natural language processing (NLP) applied on issue trackers**. In *Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering*. Association for Computing Machinery, New York, NY, USA, 38–41. <https://doi.org/10.1145/3283812.3283825>
- (Facchinetti et al, 2022) FACCHINETTI, Tullio; BENETTI, Guido; GIUFFRIDA, Davide; NOCERA, Antonino. 2022. **SLR-kit: A semi-supervised machine learning framework for systematic literature reviews**. *Knowledge-Based Systems*. Vol. 251, 2022,109266. ISSN 0950-7051.
- (Forza, 2002) **Erro! Fonte de referência não encontrada.**FORZA, C. 2002. **"Survey research in operations management: a process-based perspective"**, *International Journal of Operations & Production Management*, Vol. 22 No. 2, pp. 152-194. <https://doi.org/10.1108/01443570210414310>
- (Franco, 2005) FRANCO, Maria Amélia Santoro. 2005. **Pedagogia da pesquisa-ação**. Educação e pesquisa, v. 31, p. 483-502, 2005.
- (Gil, 2006) GIL, A C. 2006. **Como elaborar projetos de pesquisa**. 4 ed. São Paulo: Atlas, 175p.
- (Hofmann, 1999). HOFMANN, Thomas. 1999. **Probabilistic latent semantic analysis**. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI'99)*. Morgan Kaufmann Publisher Inc., San Francisco, CA, USA, 289–296.
- (Karunananda et al, 2021) KARUNANANDA, A.; SILVA, T; HANDAPANGODA, D.; SUMANAPALA, S. 2021. **Intelligent Personal Research Assistant**. 5th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI), pp. 1-7, Colombo, Sri Lanka, Doi: 10.1109/SLAAI-ICAI54477.2021.9664714.
- (Khanzode et al., 2020) KHANZODE, Ku. Chhaya A.; SARODE, Dr. Ravindra D. 2020. **Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A Literature Review**. *International Journal of Library & Information Science (IJLIS)*. Volume 9 Issue 1, January-April, 30-36. ISSN Print: 2277-3533 and ISSN Online: 2277-3584. Doi: <https://doi.org/10.17605/OSF.IO/GV5T4>.

- (Kitchenham, 2004) KITCHENHAM, Barbara. 2004. **Procedures for performing systematic reviews**. Keele, UK, Keele University, v. 33, n. p. 1-26, 2004.
- (Kitchenham et al, 2010) KITCHENHAM, B.A.; PRETORIUS, R.; BUDGEN, D.; BRERETON, P.; TURNER, M.; NIAZI, M. and LINKMAN, S. 2010. **Systematic Literature Reviews in Software Engineering — A Tertiary Study**. Information and Software Technology 52, 8 (2010), 792-805. Doi: <http://dx.doi.org/10.1016/j.infsof.2010.03.006>
- (Kitchenham&Brereton, 2013) KITCHENHAM, Barbara; BRERETON, Pearl, 2013. **A systematic review of systematic review process research in software engineering**. Information and software technology, v. 55, n. 12, p. 2049-2075.
- (Kometani&Nagaoka, 2016) KOMETANI, Y., NAGAOKA, K. 2016. **Construction of a literature review support system using Latent Dirichlet Allocation**. In Human Interface and the Management of Information: Applications and Services - 18th International Conference, HCI International, Proceedings (Vol. 9735, pp. 159-167). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 9735). Springer Verlag. https://doi.org/10.1007/978-3-319-40397-7_16.
- (Lakatos&Marconi, 2006) LAKATOS, E. M.; MARCONI, M. A. 2006. **Fundamentos da Metodologia Científica**. 6. ed. São Paulo: Atlas, 2006. 315 p.
- (Lin et al, 2020) LIN, H.J.; SHEU, P. C. Y.; TSAI, J. J. P. et al. 2020. **Text mining in a literature review of urothelial cancer using topic model**. BMC Cancer 20, 462. <https://doi.org/10.1186/s12885-020-06931>.
- (Lyu, 2022) LYU, Yuting. 2022. **Dockerized Knowledge-Oriented Multi-modal Social Event Detection System**. In: 2022 5th International Conference on Data Science and Information Technology (DSIT). IEEE, 2022. p. 1-6.
- (Mateen et al, 2021) MATEEN, Ahmed; YASIR, Muhammad; NAWAZ, Qamar; AFSAR, Salman; YASIN Qasim; YUNUSI, Muhammad. 2021. **An Analysis on Text Mining Techniques for Smart Literature Review**. International Journal of Advanced Trends in Computer Science and Engineering. Volume 10, No.2, March - April 2021
- (Miguel et al, 2010) MIGUEL, P. A. Cauchick; FLEURY, A.; MELLO, C. H. P., NAKANO, D. N., Turrioni, J. B., Ho, L. L., ... & Pureza, V. 2010. **Metodologia de pesquisa em engenharia de produção e gestão de operações**. Rio de Janeiro: Editora Elzevir. ISBN: 9788535235234
- (Miguel, 2011) Miguel, Paulo. 2011. **Qualitative research approach in production engineering - an assessment of a research project and a sample of Master of Science dissertations**. Exacta. 9. 197-206. 10.5585/exacta. v9i2.2503.
- (Monard&Baranauskas, 2003) MONARD, Maria Carolina; BARANAUSKAS, José Augusto. 2003. **Conceitos sobre aprendizado de máquina**. Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32.
- (Nature, 2022) NATURE. 2022. **Principal component analysis**. Springer Nat Rev Methods Primers 2, 101. <https://doi.org/10.1038/s43586-022-00192-w>
- (Newcastle, 2023) NEWCASTLE University, library guides. 2023. **What is grey literature?** Disponível em: <https://libguides.newcastle.edu.au/social-sciences-grey-lit/grey-lit> Acesso em: 15 Mar 2023
- (OpenAI, 2023) OPENAI. **ChatGPT (GPT-3.5) V3**. 2023. Disponível em: <https://openai.com/blog/chatgpt>. Acesso em 23/03/2023.

- (Reinehr, 2008) REINEHR, S. 2008. **Reuso sistemático de software e linhas de produto de software no setor financeiro: estudo de caso no Brasil**. Tese (Doutorado) – Departamento de Engenharia de Produção da Escola Politécnica de USP, 327p, São Paulo.
- (Rodrigues&Oliveira, 2012) RODRIGUES, Rosângela Schwarz; OLIVEIRA, Aline Borges de. 2012. **Periódicos científicos na América Latina: títulos em acesso aberto indexados no ISI e SCOPUS**. *Perspectivas em Ciência da Informação*, Volume: 17, Número: 4, pág.: 77-99.
- (Rong, 2014) RONG, Xin. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.
- (Ros et al., 2017) ROS, Rasmus; BJARNASON, Elizabeth; RUNESON, Per. 2017. **A machine learning approach for semi-automated search and selection in literature studies**. In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. p. 118-127.
- (Santos, 1999) SANTOS, Antônio Raimundo dos. 1999. **Metodologia científica: a construção do conhecimento**. Editora DP&A p.139.
- (Santos&Pimenta&Nobre, 2007). SANTOS, Cristina Mamédio da Costa; PIMENTA, Cibele Andrucio de Mattos; NOBRE, Moacyr Roberto Cuce. 2007. **A estratégia PICO para a construção da pergunta de pesquisa e busca de evidências**. *Revista latino-americana de enfermagem*, 15: 508-511.
- (Sarkis-Onofre et al, 2021) SARKIS-ONOFRE, R., CATALÁ-LÓPEZ, F., AROMATARIS, E. et al. 2021. **How to properly use the PRISMA Statement**. *Syst Rev* 10, 117. <https://doi.org/10.1186/s13643-021-01671-z>
- (Sergio et al, 2019) SERGIO, M. P., de Souza COSTA, T., de Paula PESSÔA, M. S., & PEDRO, P. S. 2019, June. **A semantic approach to support the analysis of abstracts in a bibliographical review**. In *2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (pp. 259-264). IEEE.
- (Schopf et al., 2021) SCHOPFEL, Joachim; FARACE, Dominic; FRANTZEN, Jerry. 2021. Data from **"OpenGrey, System for Information on Grey Literature in Europe"**.
- (Sievert&Shirley, 2014) SIEVERT, Carson; SHIRLEY, Kenneth. 2014. **LDavis: A method for visualizing and interpreting topics**. In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. p. 63-70.
- (Silva, 2021) SILVA, Sherlon Almeida da. 2021. **Avaliação de um sistema de analítica visual para recuperação de informação em coleções de documentos**. 2021. PhD Thesis. Universidade de São Paulo.
- (Tanhaei et al, 2022) TANHAEI, Ghazaleh, et al. 2022. **Designing a Topic-Based Literature Exploration Tool in AR-An exploratory study for neuroscience**. In: *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. p. 471-476.
- (Tauchert et al, 2020) TAUCHERT, Christoph, et al. 2020. **Towards an integrative approach for automated literature reviews using machine learning**. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.
- (Thiollent, 2018) THIOLLENT, M. 2018. **Metodologia da pesquisa-ação**. 18ª Edição. Editora Cortez, São Paulo. 132p.
- (Tiba et al, 2018) TIBA, Sarah; VAN RIJNSOEVER, Frank J.; HEKKERT, Marko P. 2018. **Firms with benefits: A systematic review of responsible entrepreneurship and corporate**

- social responsibility literature.** *Corporate Social Responsibility and Environmental Management*, 26. 2: 265-284. Doi:10.1002/csr.1682
- (Thomas, 2007) THOMAS, G. 2007. **Introdução: evidência e prática.** In: THOMAS, G. et al. *Educação baseada em evidências: atualização dos achados científicos para a qualificação da prática pedagógica.* Porto Alegre: Artmed, p. 9-27.
- (Vosgerau&Romanowski, 2014) VOSGERAU, Dilmeire Sant Anna Ramos; ROMANOWSKI, Joana Paulin. 2014. **Estudos de revisão: implicações conceituais e metodológicas.** *Rev. Diálogo Educ.* p. 165-190.
- (Waal&Barnard, 2008) WAAL, A.D., Venter, J., & BARNARD, E. 2008. **Applying Topic Modeling to Forensic Data.** *IFIP Int. Conf. Digital Forensics. International Federation for Information Processing Digital Library; Advances in Digital Forensics IV;* 285. DOI: 10.1007/978-0-387-84927-0_10, Corpus ID: 18135631
- (Weber, 1999) WEBER, R. 1999, June. **Intelligent Jurisprudence Research: a new concept.** In *Proceedings of the 7th international conference on Artificial intelligence and law* (pp. 164-172)
- (Westbrook, 1995) WESTBROOK, R. 1995. **Action research: a new paradigm for research in production and operations management.** *International Journal of Operations & Production Management*, v. 15, n. 12, p. 6-20. <http://dx.doi.org/10.1108/01443579510104466>
- (Wood-Harper, 1995) WOOD-HARPER, Trevor. 1995. **Research methods in information systems: using action research.** *Research methods in information systems*, v. 169, p. 191.
- (Wu et al., 2008) Ho Chung WU, Robert Wing Pong LUK, Kam Fai WONG, and Kui Lam KWOK. 2008. **Interpreting TF-IDF term weights as making relevance decisions.** *ACM Trans. Inf. Syst.* 26, 3, Article 13, 37 pages. <https://doi.org/10.1145/1361684.1361686>.
- (Yin, 2005) YIN, R. K. 2005. **Estudo de Casos: Planejamento e Métodos.** ISBN: 8536304626. Url: <https://ria.ufrn.br/123456789/943>, making relevance decisions, 212p. 3 ed. Porto Alegre: Bookman.
- (YueKang et al, 2020) Yue Kang, Zhao Cai, Chee-Wee Tan, Qian Huang & Hefu Liu. 2020. **Natural language processing (NLP) in management research: A literature review,** *Journal of Management Analytics*, 7:2, 139-172, Doi: 10.1080/23270012.2020.1756939
- (Yu&Egger, 2021) YU, Joanne; EGGER, Roman. 2021. **Tourist experiences at overcrowded attractions: A text analytics approach.** In: **Information and Communication Technologies in Tourism 2021: Proceedings of the ENTER 2021 e Tourism Conference.** Springer International Publishing, p. 231-243.

APÊNDICE I – MANUAL DO SW3T

Este apêndice descreve como usar o aplicativo Sw3T. O aplicativo está disponível no site “www.sw3t.com.br” em um servidor temporário até o final de 2023.

(i) Entrada no aplicativo

Para entrar no aplicativo o usuário deve se cadastrar e criar sua senha. A Figura 27 exibe a tela inicial.

Figura 27 - Tela inicial no Sw3T

A tela de geral, Figura 28, mostra todos os projetos do usuário. Ela também permite criar novos projetos.

#	Nome	Descrição	Fim Projeto
2	LDA - Revisao Literaria	desc	31/07/2020
24	Data maning and literature review		31/12/2020
49	whatsapp 2022		31/12/2030
69	Sw3T Ensueto Total	atualizar descrição	31/12/2023
70	LDA TOTAL COM BIGRAM		31/12/2023
75	Revisao Literaria aplicando Modelo de Topico LDA		31/12/2023
76	teste		31/12/2023
79	conteúdo de artigos		31/12/2023

Figura 28 - Tela Geral do Sw3T - todos os projetos de um usuário

A Tela Geral do aplicativo contém um índice do lado esquerdo que se mantém em todas as telas. Somente o espaço à direita apresenta novas informações de acordo com a função que está sendo executada.

A seguir são apresentadas as funcionalidades do Sw3T sugerindo uma sequência de execução das funções.

(ii) Criar um novo projeto

Para criar um novo projeto deve-se clicar no ícone com sinal (“+”) da Tela Geral e ir para tela de Projeto Novo como na Figura 29.

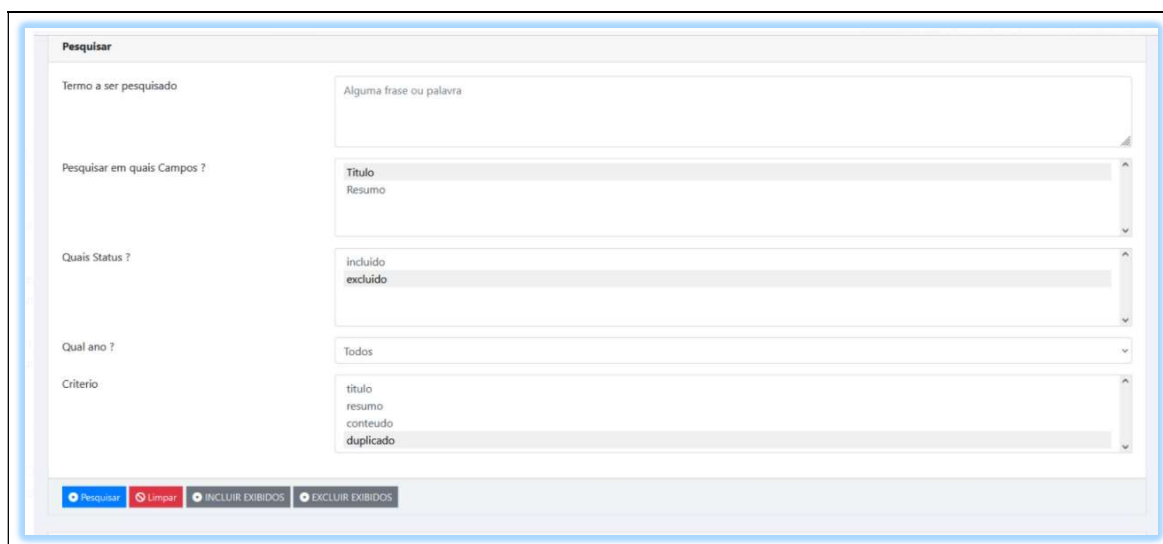


Figura 29 - Tela do Sw3T para incluir novo projeto

Para concretizar a inclusão do novo projeto, deve-se clicar no botão “cadastrar”. Esta tela também alterar o conteúdo e/ou excluir o projeto.

(iii) Acessando projeto já cadastrado

Na Tela Geral, ao lado de cada projeto existem dois botões: ver as informações do projeto (retorna a tela incluir projeto) ou entrar no projeto.

Ao entrar no projeto se visualiza a tela do Projeto conforme Figura 30, com as várias funções.

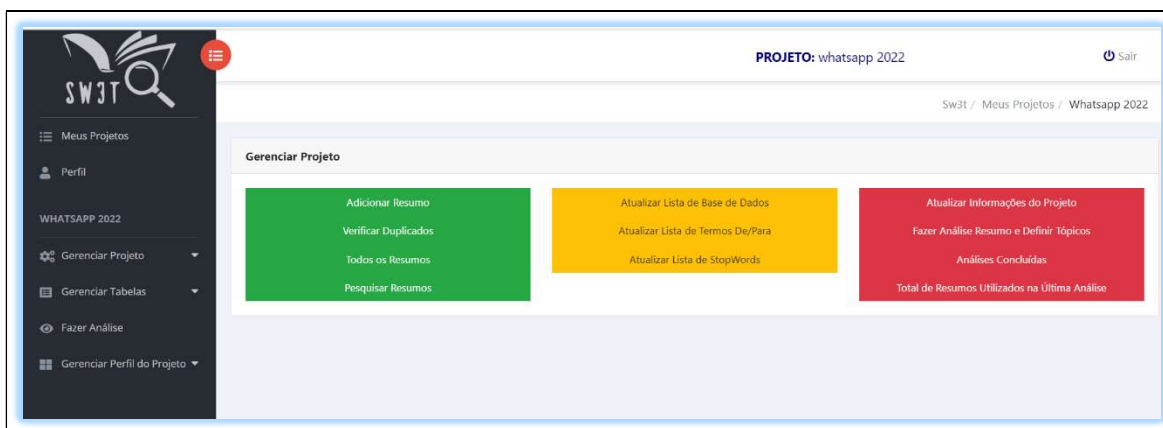


Figura 30 - Tela com as funções do Sw3T num Projeto

No **retângulo verde**, estão as funções de incluir metadados, verificar duplicado, ver todos os metadados do projeto e pesquisar termos por título ou resumo.

No **retângulo amarelo** estão as funções de atualizar lista de BDs, atualiza lista de termos de/para e atualizar lista de Stopwords.

No **retângulo vermelho** estão as funções de atualizar dados do projeto - retornando para tela incluir projeto e as funções de tratamento e análise dos resumos do projeto.

(iv) Tratar metadados do Projeto

A. Incluindo metadados de artigos

Na função incluir arquivo, na Figura 31 está a tela de inclusão do projeto. o usuário localiza o arquivo XML com metadados a serem incluídos. Informa fonte, área de conhecimento e inclui. Hoje estão previstas três classificações (exatas, biológicas e humanas). No caso de arquivo sem área de conhecimento definida, o usuário tem a opção “não definida”.

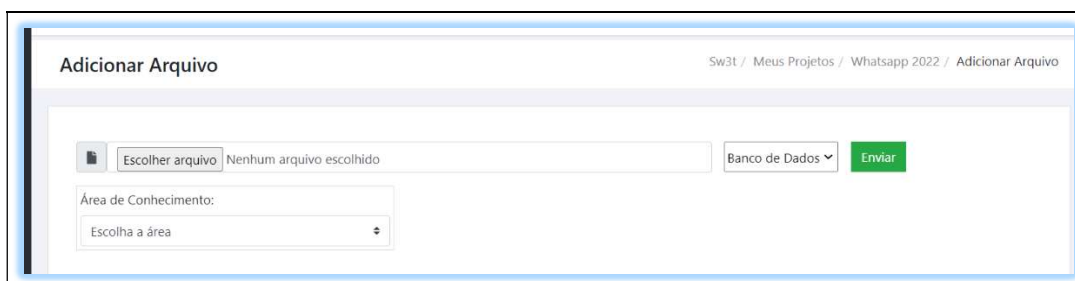


Figura 31 - Tela do Sw3T para Incluir arquivo XML de metadados

É feita a verificação da existência de conteúdo nos campos: Título, autores, resumo, data publicação. Aparece a mensagem informando quantos artigos foram incluídos e uma lista dos metadados incompletos.

B. Eliminar duplicados

A Figura 32 mostra a tela onde são verificadas se há metadados duplicados por título, ano e resumo.

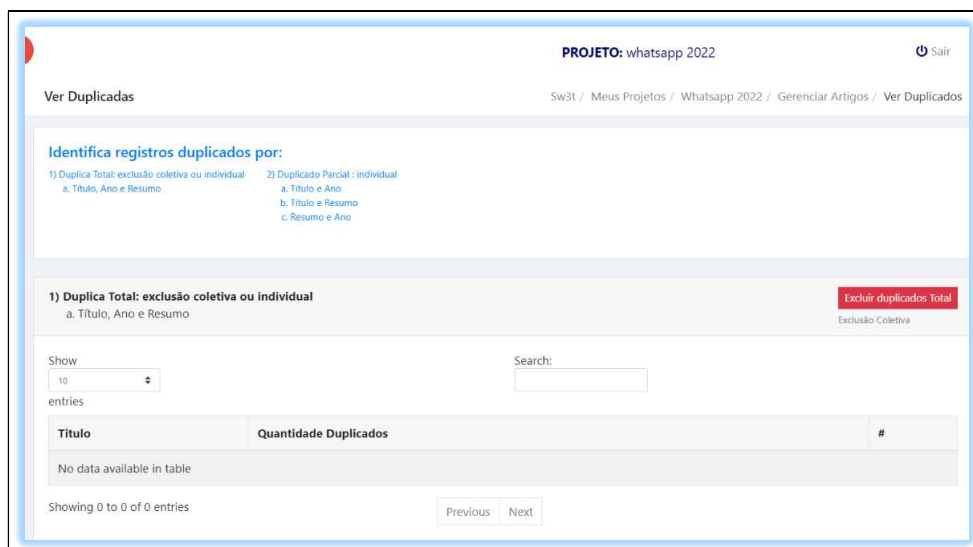


Figura 32 - Tela do Sw3T para tratar metadados duplicados

C. Visualizar todos os metadados

Para visualizar todos os metadados do projeto, tem-se a tela ver todos os metadados do projeto Figura 33 e pode-se exportar todos os metadados com o histórico da seleção.

Na tela pesquisar dados, pode-se listar exibir os metadados incluídos e/ou excluídos, assim como mudar seu status. O ícone “deletar todos os resumos” – limpa toda a base de dados do projeto.

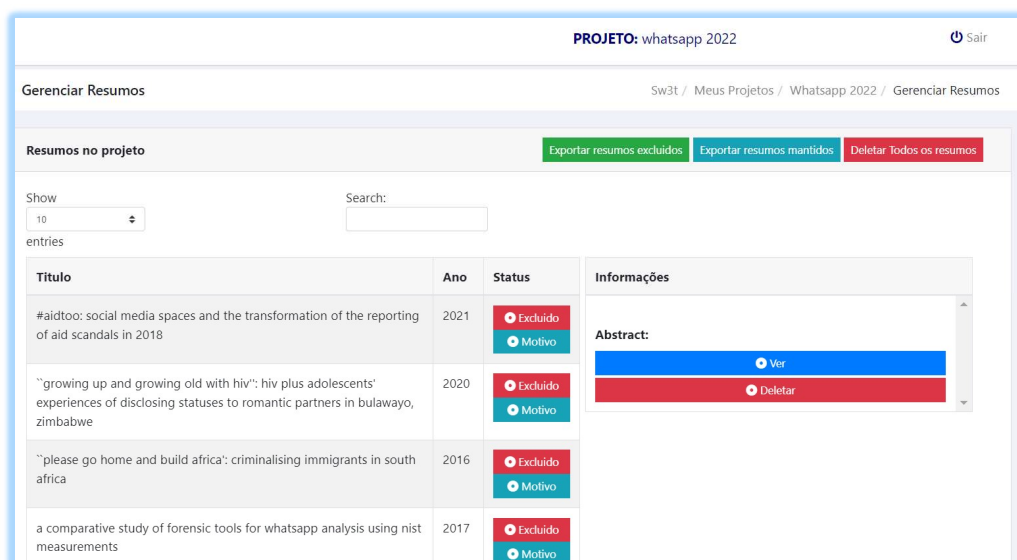


Figura 33 - Tela do Sw3T para ver todos os metadados de um projeto

Outras opções disponíveis nesta tela é a de exportar para um arquivo Excel todos os metadados, só os mantidos ou só os excluídos. Este recurso

permite que todas as informações sejam disponibilizadas em forma de planilha. Possibilitando seu uso de outras ferramentas como geração de gráficos.

D. Pesquisar metadados

Para realizar pesquisa pôr termo os metadados do projeto, tem-se as possibilidades de ser no resumo, no título, por ano e escolher se quer visualizar os excluídos e os incluídos. Veja a tela de pesquisa na Figura 34.

Figura 34 - Tela do Sw3T de pesquisa nos resumos

(v) Gerenciar tabelas do Projeto

O aplicativo conta com três tabelas de apoio, elas são: Troca de termos (de/para), Stopwords do usuário, e fontes de dados. Elas podem ser atualizadas a qualquer tempo dentro do projeto.

E. Gerenciar tabela de BDs.

Veja na Figura 35 a tabela atualizar a Lista de Base de Dados.

Figura 35 - Tela do Sw3T para atualizar a tabela de BDs fonte

F. Atualizar lista de termos (de/para): A Figura 36 mostra a tela para incluir, excluir e alterar dados da tabela de/para.

Figura 36 - Tela do Sw3T para atualizar tabela de/para

G. Atualizar lista de Stopwords do usuário. A Figura 37 mostra a tela para incluir, excluir e alterar dados da tabela

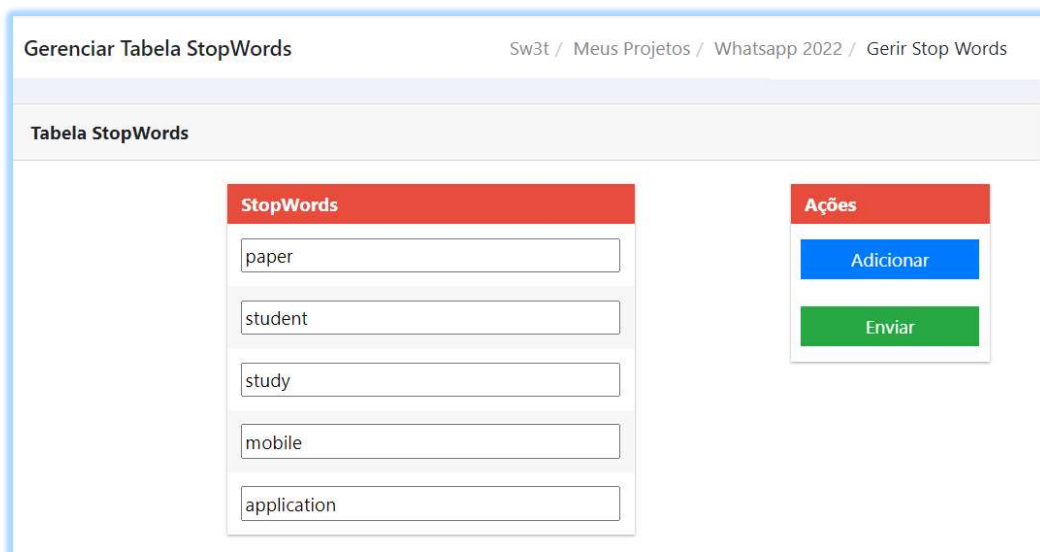


Figura 37 - Tela do Sw3T para atualizar tabela *Stopwords* do usuário

Outra forma de incluir novos termos desta tabela *Stopwords* é através tela de tópicos na lista de termos disponibilizados. Também existe a função de importar tabelas de *Stopwords* e de/para par outro projeto do usuário. Veja a Figura 38.

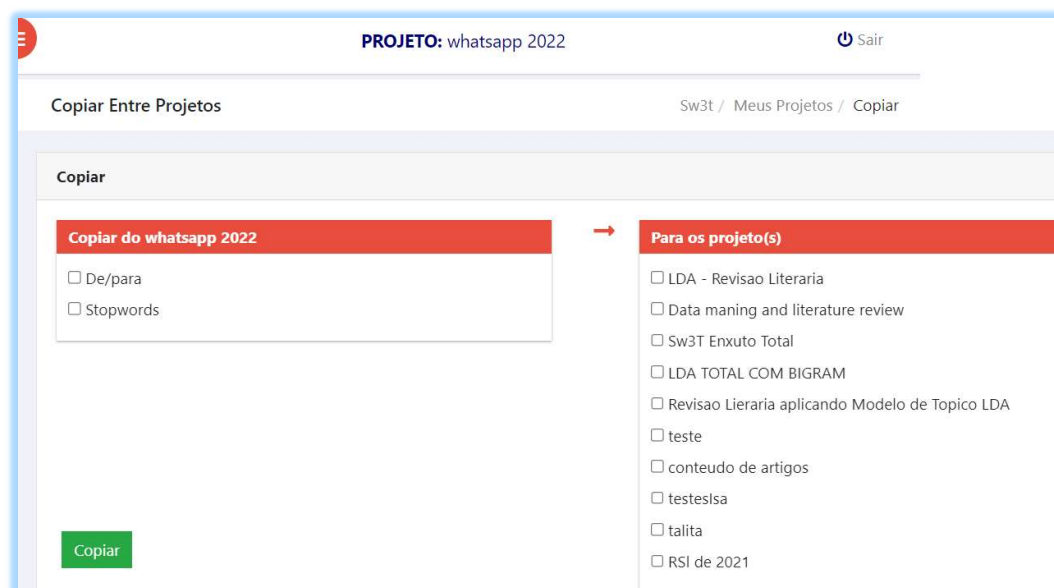


Figura 38 - Tela do Sw3T para exportar tabelas para outro projeto

(vi) Tratar resumos e fazer a análise de modelagem de tópicos

Na tela de fazer análise, a execução dos botões deve seguir uma ordem pré-definida. E o preenchimento dos parâmetros da análise.

Figura 39 - Tela do Sw3T para preparar resumos e fazer análise

A sequência de botões da tela na Figura 39 deve ser:

- Transforma em letra minúscula e limpar pontuação.
- Tratar tabelas de *StopWords* e de/para, retirar *StopWords* padrão e lematiza os termos.
- Por fim, preencher os parâmetros e gerar tópicos.

Os parâmetros a serem preenchidos são:

- Número e tópicos a serem gerados;
- Número de palavras-chave serem utilizadas para definir os tópicos,
- Ciclos de aprendizado do modelo,
- Intervalo de anos a serem analisados
- Para qual área de conhecimento
- Se deseja ver o gráfico. ^

Ao final da análise irá aparecer na tela o botão “ver análise”.

(vii) Ver análise concluída e realizar o processo de seleção de resumos

A Figura 40 exibe os tópicos da análise e as palavras-chave identificadas durante a análise. A quantidade de resumos por tópico é exibida.

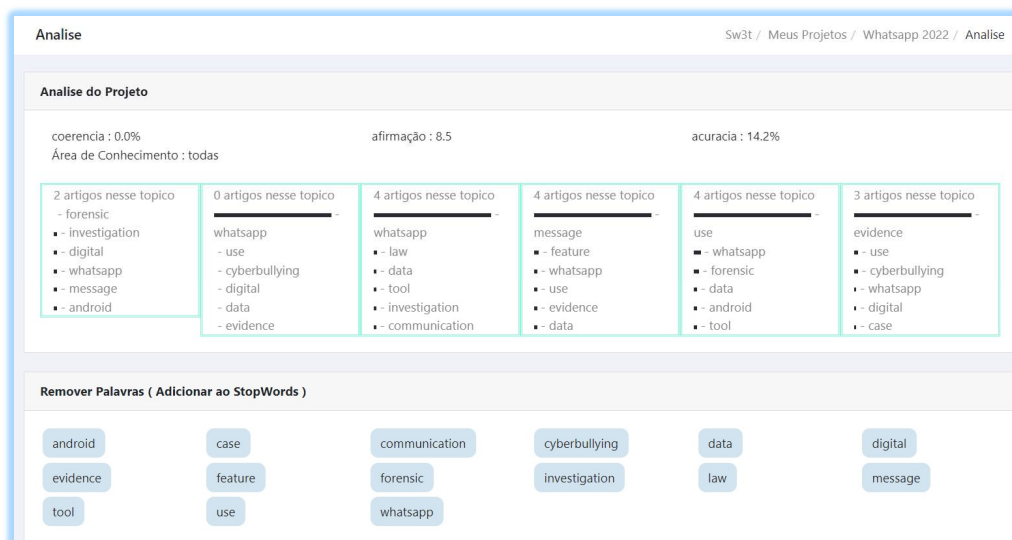


Figura 40 - Tela do Sw3T para visualizar uma análise

(viii) Visualizar Gráfico de dispersão de cluster.

Exibe a disposição dos tópicos considerando a aproximação entre os clusters. Quanto mais distantes, menos tópicos incomuns há entre os clusters, como pode ser observado na Figura 41 (Sievert&Shirley, 2014). A afirmação deve estar alta, mas a acurácia deve estar baixa.

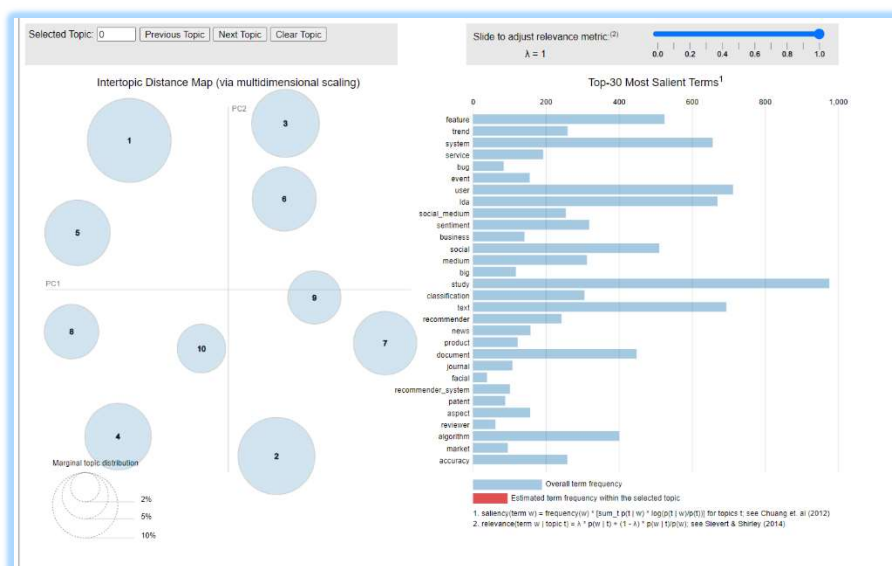


Figura 41 - Tela do gráfico do Sw3T que mostra proximidade entre clusters

(ix) Visualizar de um cluster.

Na Figura 42 é apresentado todos os resumos que compõem o cluster e o indicador de dominância que apresenta a probabilidade de o resumo pertencer ao tópico em relação aos outros tópicos.

Relação de Artigos por tópico

artifact
 └─ evidence
 └─ whatsapp
 └─ forensic
 └─ message
 └─ data
 └─ digital
 └─ feature

[Voltar](#)

[Excluir todos os resumos deste resumo](#)

Resumos nesse Tópico: 4

Todos Mantidos Excluídos

Título	Ano	% Dominancia	Status	#
browser forensic investigations of whatsapp web utilizing indexeddb persistent storage	2020	99.55%	Mantido Motivo	Ver Deletar
framework design for the retrieval of instant messaging in social media as electronic evidence	2020	99.54%	Mantido Motivo	Ver Deletar
forensic analysis of encrypted instant messaging applications on android	2018	99.63%	Mantido Motivo	Ver Deletar

Figura 42 - Tela do Sw3T com resumos de um cluster

(x) Visualizar de um resumo de um cluster.

Na Figura 43 pode-se visualizar uma a uma das etapas de preparação dos resumos.

PROJETO: whatsapp 2022

Sair

Ver Resumo

Sw3t / Meus Projetos / Whatsapp 2022 / Gerenciar Resumos / Ver Resumo

Título do resumo: browser forensic investigations of whatsapp web utilizing indexeddb persistent storage

Resumo Original

digital evidence is becoming an indispensable factor in most legal cases. however, technological advancements that lead to artifact complexity, are forcing investigators to create sophisticated connections between the findings and the suspects for admissibility of evidence in court. this paper scrutinizes whether indexeddb, an emerging browser technology, can be a source of digital evidence to provide additional and correlating support for traditional investigation methods. it particularly focuses on the artifacts of the worldwide popular application, whatsapp. a single case pretest-posttest quasi experiment is applied with whatsapp messenger and web application to populate and investigate artifacts in indexeddb storage of google chrome. the findings are characterized and presented with their potential to be utilized in forensic investigation verifications. the storage locations of the artifacts are laid out and operations of extraction, conversion and presentation are systematized. additionally, a proof of concept tool is developed for demonstration. the results show that whatsapp web indexeddb storage can be employed for time frame analysis, demonstrating its value in evidence verification.

1211 Caracteres

Resumo Enxuto

Cum troca de termos e eliminação de stopwords do usuário digital evidence is becoming an indispensable factor in most legal cases however technological advancements that lead to artifact complexity are forcing investigators to create sophisticated connections between the findings and the suspects for admissibility of evidence in court this paper scrutinizes whether indexeddb an emerging browser technology can be a source of digital evidence to provide additional and correlating support for traditional investigation methods it particularly focuses on the artifacts of the worldwide popular application whatsapp a single case pretest posttest quasi experiment is applied with whatsapp messenger and web application to populate and investigate artifacts in indexeddb storage of google chrome the findings are characterized and presented with their potential to be utilized in forensic investigation verifications the storage locations of the artifacts are laid out and operations of extraction conversion and presentation are systematized additionally a proof of concept tool is developed for demonstration the results show that whatsapp web indexeddb storage can be employed for time frame analysis demonstrating its value in evidence verification

1211 Caracteres

Resumo Lematizado

Sem stopwords do sistema e termos simplificados digital evidence become indispensable factor most legal case however technological advancement that lead artifact complexity force investigator create sophisticated connection between the finding and the suspect for admissibility evidence court this paper scrutinize whether indexeddb emerge browser technology can source digital evidence provide additional and correlate support for traditional investigation method particularly focus the artifact the worldwide popular application whatsapp single case pretest posttest quasi experiment apply with whatsapp messenger and web application populate and investigate artifact indexeddb storage google chrome the finding characterize and present with their potential utilize forensic investigation verification the storage location the artifact lay out and operation extraction conversion and presentation systematize additionally proof concept tool develop for demonstration the result show that whatsapp web indexeddb storage can employ for time frame analysis demonstrate value evidence verification

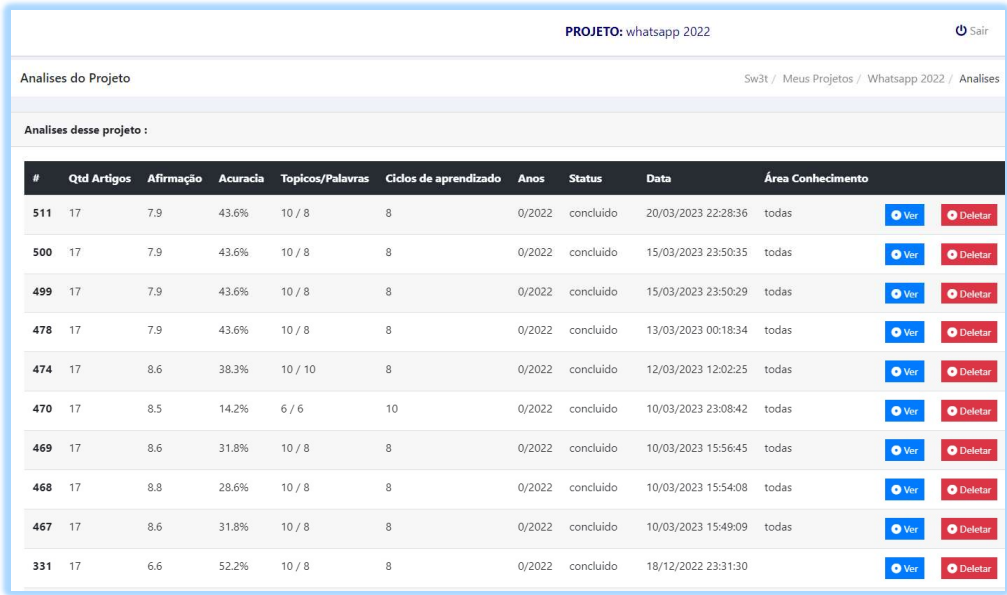
[Gerar Resumos Enxutos](#)

[Deletar Artigo](#)

Figura 43 - Tela do Sw3T das etapas do tratamento dos resumos

(xi) Acompanhar o andamento do projeto

A tela de visualização de todas as análises, está na Figura 44. Permite a visualização de todas as análises do projeto facilitando comparar os resultados das métricas de cada análise pelos indicadores de afirmação e de acurácia e assim decidir qual a análise traz a melhor combinação entre os parâmetros número de tópicos, número de palavras e ciclos de treinamento.



The screenshot shows a web interface for a project named 'PROJETO: whatsapp 2022'. It displays a table titled 'Análises desse projeto:' with the following columns: #, Qtd Artigos, Afirmação, Acurácia, Tópicos/Palavras, Ciclos de aprendizado, Anos, Status, Data, Área Conhecimento, and buttons for 'Ver' and 'Deletar'. The table lists 11 analysis entries with their respective metrics.

#	Qtd Artigos	Afirmação	Acurácia	Tópicos/Palavras	Ciclos de aprendizado	Anos	Status	Data	Área Conhecimento	Ver	Deletar
511	17	7.9	43.6%	10 / 8	8	0/2022	concluido	20/03/2023 22:28:36	todas	Ver	Deletar
500	17	7.9	43.6%	10 / 8	8	0/2022	concluido	15/03/2023 23:50:35	todas	Ver	Deletar
499	17	7.9	43.6%	10 / 8	8	0/2022	concluido	15/03/2023 23:50:29	todas	Ver	Deletar
478	17	7.9	43.6%	10 / 8	8	0/2022	concluido	13/03/2023 00:18:34	todas	Ver	Deletar
474	17	8.6	38.3%	10 / 10	8	0/2022	concluido	12/03/2023 12:02:25	todas	Ver	Deletar
470	17	8.5	14.2%	6 / 6	10	0/2022	concluido	10/03/2023 23:08:42	todas	Ver	Deletar
469	17	8.6	31.8%	10 / 8	8	0/2022	concluido	10/03/2023 15:56:45	todas	Ver	Deletar
468	17	8.8	28.6%	10 / 8	8	0/2022	concluido	10/03/2023 15:54:08	todas	Ver	Deletar
467	17	8.6	31.8%	10 / 8	8	0/2022	concluido	10/03/2023 15:49:09	todas	Ver	Deletar
331	17	6.6	52.2%	10 / 8	8	0/2022	concluido	18/12/2022 23:31:30	todas	Ver	Deletar

Figura 44 - Tela do Sw3T com todas as análises de um projeto

Afirmação: indica (de 0 A 10), o quanto a quantidade de tópicos definido é adequado sendo quanto maior mais adequado onde quanto maior seu valor mais adequado é a quantidade de tópicos.

Acurácia: Indica a qualidade da distribuição das tems - medida da capacidade do modelo de classificar corretamente os termos em seus respectivos tópicos. Essa medida é geralmente calculada através da taxa de acerto, que é a proporção de termos corretamente classificados em relação ao total de termos.

(xii) Visualizar o histórico de inclusão e exclusão dos metadados do projeto

A tela da Figura 45 fornece o histórico de inclusão e exclusão do projeto.

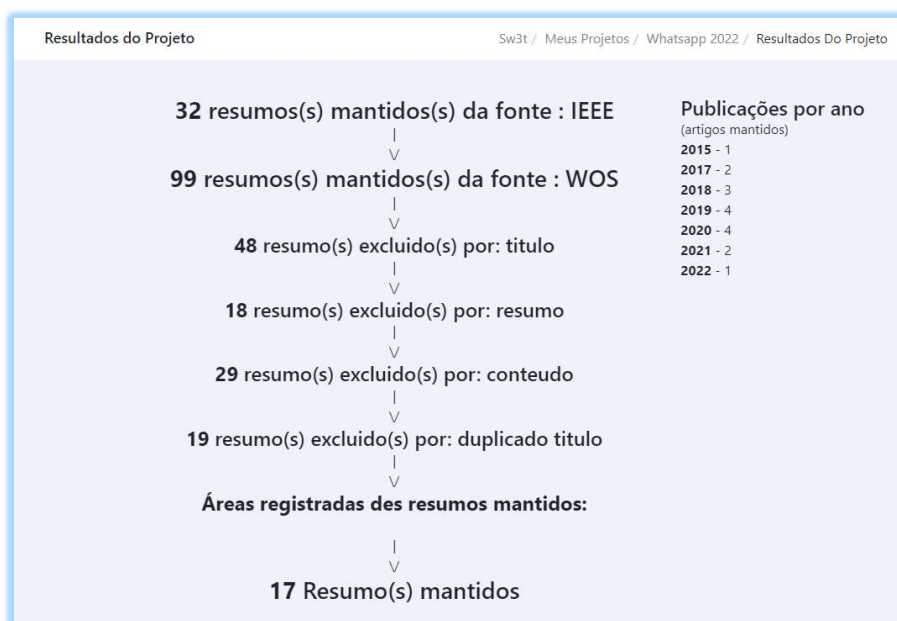


Figura 45 - Tela do Sw3T com histórico de seleção dos metadados do projeto

APÊNDICE 2 – POSTER PIBIC E INDICAÇÃO AO PRÊMIO IC 2018



**PIBIC/CNPQ/CTI
JICC 2018**

XX Jornada de Iniciação Científica do CTI



ABORDAGEM SEMÂNTICA NA REVISÃO LITERÁRIA PARA ANÁLISE DE ABSTRACTS

COSTA, Talita.S.¹; SILVA, Rafael.L.²; SERGIO, Marbília P.²

¹Centro de Tecnologia da Informação Renato Archer – CTI

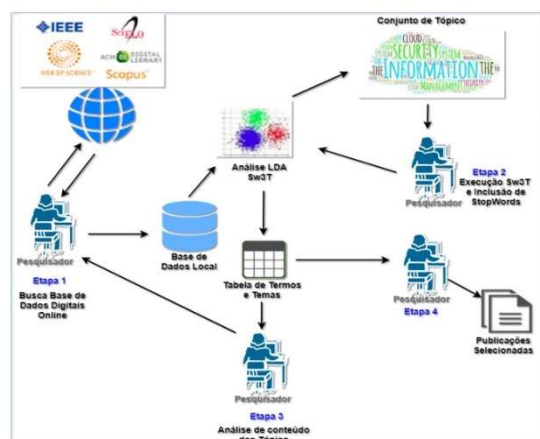
²Fundação de Apoio à Capacitação em TI – FACTI

[tscosta@cti.gov.br, rafael.lucas@facti.com.br, marbilia.sergio@cti.gov.br]

Introdução

A revisão literária é indispensável, não somente para definir bem o problema de pesquisa, mas também para obter uma ideia precisa sobre o estado da arte de um determinado tema [1]. Hoje em dia, há uma grande quantidade de conhecimento científico disponível em Bases de Dados Digitais (BDDs). No entanto, a busca por publicações relevantes leva o pesquisador a executar tarefas manuais e exaustivas, sujeitas a falhas.

Este trabalho de pesquisa se propôs a estudar os conceitos de análise semântica visando sua aplicação na atividade de revisão literária. Através do desenvolvimento de um aplicativo, pretendeu-se criar um mecanismo para identificação de termos complementares de busca



Resultados Obtidos

Para se estabelecer o estado da arte no tema foram observadas duas contribuições relevantes do aplicativo Sw3T, descritas e esquematizada na Figura 1 - Visão Geral das 4 etapas da Revisão Literária utilizando o Sw3T.

Etapa 1 – Busca por publicações nas BDDs. Utilizando-se Termos: principal e complementares. As bases definidas foram: ACM, IEEE, Scopus, WoS e Scielo.

Etapa 2 – Tratamento do conteúdo dos abstracts o que inclui também a identificação de termos comuns a serem incluídos no arquivo *StopWords*.

Esta etapa deve ser executada várias vezes até que o pesquisador avalie como satisfatório para identificar os temas. No caso estudado, foram identificadas os termos individuais “*Information*” e “*Security*” complementando o termo principal de busca. A tabela 1 - Exemplo de termos incluídos no arquivo *Stop Words*.

order	new	service	important	order	software	2013	paper
developed	Empty	have	feedback	study	based	section	get
however	proposed	chain	critical	type	around	chapter	the

Observa-se que os termos incluídos no *StopWords* são palavras comuns não relacionadas ao Tema.

Etapa 3 – Identificar termos complementares tornando a busca mais seletiva e viabilizando a ampliação para as demais BDDs. Observou-se que utilizando apenas de único termo composto “*Information Security Management*” obtinha-se um volume expressivo de publicações. Assim, analisou-se o resultado do SCOPUS para identificar termos complementares.

Etapa 4 – Os temas e termos resultantes da execução do Sw3T também puderam auxiliar o pesquisador na análise detalhada das publicações selecionadas uma vez que o Temas agrupavam as publicações semelhantes orientando a sequência de leituras.

Conclusão

O aplicativo trouxe agilidade no processo de revisão literária trazendo efeitos replicáveis e consistentes para uma pesquisa. Há necessidade de novos testes variado temas afim de corroborar com este resultado de que Tópicos e Termos gerados pelo Sw3T são efetivos nesta sua aplicação. Trabalhos futuros podem incluir a automação da busca nas diversas BDDs.

Referências

- [1] KITCHENHAM, Barbara et al. Systematic literature reviews in software engineering – a systematic literature review. Information and software technology, 2009.
- [2] FALEIROS, T. P.. Modelos probabilísticos de tópicos: desvendando o Latent Dirichlet Allocation. 2016.



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA,
INOVAÇÕES E COMUNICAÇÕES

Carta de Recomendação

Campinas, 22 de fevereiro de 2018.

Prezada Comissão do 16º Prêmio Destaque de Iniciação Científica e Tecnológica - 2018

Eu, servidora pública federal, MsC. Marbília Possagnolo Sergio, Pesquisadora do Centro de Pesquisa Renato Archer - CTI/MCTIC e coordenadora do Núcleo de Qualidade de Software – NQS, e a Dr^a Regina Paiva Melo Marin, Professora do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - IFSP - Campus Campinas, venho recomendar a minha bolsista de iniciação científica a Srta. **Talita de Souza Costa** para o prêmio Destaque de ICT de 2018.

Em janeiro de 2018 deu-se o início da bolsista no NQS com a sua integração ao projeto de pesquisa deste núcleo denominado “Desenvolvimento de modelos de gestão de Segurança da Informação para pequenas organizações governamentais” e onde está até o momento.

Estando o projeto na fase de revisão bibliográfica, a bolsista contribuiu com o processo de coleta e recuperação de publicações de diversas bases de dados. Como o volume de publicações a serem analisadas era consideravelmente grande, fez-se necessário o uso de algoritmos para apoiar o processo de seleção através da análise semântica dos resumos dos artigos.

Para atender a este desafio, a bolsista teve que aprender diversas tecnologias como a linguagem Python, uso do algoritmo LDA, e soluções para integrar arquivos Excel com Python entre outros aprendizados como o do aplicativo mendeley para extrair e tratar o conteúdo das publicações. A divulgação do resultado deste trabalho de IC já ocorreu em três oportunidades, pode-se citar:

- ✓ Publicação de um artigo nos anais da Jornada de Iniciação do CTI - JICC 2018;
- ✓ Exposição de um pôster durante a semana nacional de ciência e tecnologia do CTI Renato Archer quando recebeu o prêmio de “melhor pôster do “Grupo C”;
- ✓ Publicação de um resumo e a exposição de um pôster no XVIII Workshop de Computação Aplicada do INPE - WORCAP 2018, quando ganhou o prêmio de “melhor Pôster de Graduação”.

A bolsista continua a desenvolver seu trabalho de pesquisa neste núcleo e vem demonstrando sua competência técnica e capacidade de aprendizado, dedicação e envolvimento profissional. Sua assiduidade corrobora com seu comprometimento com o trabalho e com a equipe. Demonstra capacidade de trabalho em equipe e iniciativa para buscar soluções para as questões que se apresentavam.

Aproveito o ensejo para registrar meus protestos de estima e consideração aos senhores.

Atenciosamente,



Coordenadora de núcleo NQS
Orientadora e MsC Marbília P. Sergio



Professora do IFSP - Campinas
Coorientadora Dr^a. Regina P. M. Marin

APÊNDICE 3 – ARTIGO PUBLICADO EM 2019 NO IEEE

O artigo completo pode ser encontrado na biblioteca do IEEE

2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)

A Semantic Approach to Support the Analysis of Abstracts in a Bibliographical Review

Marbília P. Sergio and Talita S. Costa
CTI - MCTIC
Campinas, SP, 13069-901, Brazil
Email: {marbilia.sergio, lscosta}@cti.gov.br

Marcelo S. P. Pessoa
Politécnico - USP
Sao Paulo, SP, 05508-900, Brazil
Email: mpessoa@usp.br

Paulo S. Martins
FT - Unicamp
Limeira, SP, 13484-332, Brazil
Email: paulo@ft.unicamp.br

Abstract—A large amount of scientific information is found in digital databases. This fact turns the search for the state of the art an increasingly challenging task. This work has the goal of developing a technological solution to support researchers in bibliographic review using the Latent Dirichlet Allocation. The Sw3T software, developed in Python, analyses the abstracts in scientific publications that interest the researcher. Through the use of semantic analysis, Sw3t supplies themes, which are defined by a group of terms. Themes can subsidize the researcher in the process of identification of complementary terms through the restriction on the number of publications of interest and it contributes to the analysis of the publications selected by their classification by theme as well. Sw3t has shown to be efficient and increased reliability in the search process regarding complementary terms, as well it contributes to the promptness of the detailed analysis of the publications. The results of the bibliographical review, regarding the application of Sw3T, have shown to be consistent and replicable. Future work will address the use of Sw3T for other bibliographic reviews as well as improvements and broadening of its functionalities such as search in DDB automation and previous content processing in abstracts.

I. INTRODUCTION

In this day and age, researchers rely on a significant amount of knowledge available in Digital Databases (DDBs). They allow different search criteria and some even provide previous classification of publications, such as author, area of knowledge, date of publication and the bibliometry information. However, the pursuit for relevant publications for a specific research drives the researcher to perform tasks that are prone to flaws and subject to manual and exhaustive activities.

In order to guarantee that the result of a bibliographic review is indeed consistent for the software engineering area, Kitchenham [1] proposes a systematic bibliographic review methodology. It states that in order to extract relevant publications from the DDBs, it is crucial to apply well defined search terms. A flaw in choosing such terms may lead to a bibliographic review of poor quality. The first step is to build a set of descriptors or list of appropriate keywords to the research theme to be employed in the search in the databases [1].

We observe the exponential growth of publications in the DDBs. Mello [2] points out that the advantage of these bases is the ease of obtaining information. This fact has led governments, universities and research entities to provide their

academic community with access to these DDBs. But the disadvantage is the difficulty of identifying relevant and good-quality publications. The researcher in the search for the state of the art is faced with the increasingly complex and time-consuming work of selecting relevant publications for a given research theme.

Griffiths [3] applied the Markov chain Monte Carlo algorithm to analyze the abstracts of articles in the PNAS database (Proceedings of the National Academy of Sciences of the United States). He used the Bayesian model selection to establish for a large area (e.g. social sciences, biological sciences and exact sciences) the set of topics covered in the PNAS database, which returns the terms that better apply and the intensity that they appear in several sub-themes.

Our work proposes the use of semantic analysis to support the bibliographical review until a set of publications, which represent the state of the art, meets the demand of the researcher. The question to be answered by this research is how software based on the Latent Dirichlet Allocation (LDA) algorithm can automate tasks of the bibliographical review process.

The remainder of this paper is organized as follows: Section II introduces the concepts that are relevant for the understanding of this work. Section III highlights the methodology employed and Section IV reports the main results. Section V presents the limitations of the work, future work and the main conclusions.

II. BACKGROUND

In this section, we address basic elements of the bibliographical review process and the Latent Dirichlet Allocation, which are two key concepts upon which this work is based.

A. Bibliographical Review

A bibliographical review is essential for any scientific research initiative. It not only defines the research problem, but also establishes the state of the art of a particular theme. It involves locating, analyzing, synthesizing, and interpreting existing research on the researcher's area of interest. As a result of the review, the researcher acquires a subset of publications that meet their needs and are relevant to sustaining ongoing research [4]. The state of the art allows the identification of knowledge gaps and the determination of what would be the

APÊNDICE 4 – TRECHO DO TCC COM USO DO SW3T EM 2022

Estudo sobre segurança e integridade do conteúdo do WhatsApp e sua como evidência em Processos judiciais onde fui convidada para participar da banca conforme certificado também neste apêndice. O TCC está depositado na biblioteca do Faculdade de Tecnologia - FT da UNICAMP.

	UNIVERSIDADE ESTADUAL DE CAMPINAS Faculdade de Tecnologia						
Comissão de Graduação do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas							
Certificado de Participação							
Certifico que a Profa. Mestra Marbília Passagnolo Sergio participou da Comissão Examinadora do TCC/Monografia do(a) aluno(a) Yuri Tadeu Covo Zwaig, RA 245950, intitulada "Estudo sobre segurança e integridade do conteúdo das conversas do WhatsApp e o uso como evidência de processos judiciais", realizada em 13/07/2022.							
A Comissão Examinadora foi assim constituída:							
<table border="1"> <tr> <th data-bbox="400 857 1286 880">Presidente</th> </tr> <tr> <td data-bbox="400 880 1286 925"> Profa. Dra. Marli de Freitas Gomes Hernandez FT/ Unicamp </td> </tr> <tr> <th data-bbox="400 925 1286 947">Membros Titulares</th> </tr> <tr> <td data-bbox="400 947 1286 992"> Prof. Dr. Luiz Ariovaldo Fabri Junior FT/ Unicamp </td> </tr> <tr> <td data-bbox="400 992 1286 1039"> Profa. Mestra Marbília Passagnolo Sergio FT/ Unicamp </td> </tr> </table>			Presidente	Profa. Dra. Marli de Freitas Gomes Hernandez FT/ Unicamp	Membros Titulares	Prof. Dr. Luiz Ariovaldo Fabri Junior FT/ Unicamp	Profa. Mestra Marbília Passagnolo Sergio FT/ Unicamp
Presidente							
Profa. Dra. Marli de Freitas Gomes Hernandez FT/ Unicamp							
Membros Titulares							
Prof. Dr. Luiz Ariovaldo Fabri Junior FT/ Unicamp							
Profa. Mestra Marbília Passagnolo Sergio FT/ Unicamp							
Limeira, 27 de março de 2023							
Prof. Dr. Plínio Roberto Souza Vilela Coordenador do Curso							
<table border="1"> <tr> <th data-bbox="400 1285 1286 1319">CÓDIGO DE AUTENTICIDADE</th> </tr> <tr> <td data-bbox="400 1319 1286 1375"> Verifique a autenticidade deste documento na página www.dac.unicamp.br Código: 51dc96e174dee6e3ece3211e1893ee562ac3e219 </td> </tr> </table>			CÓDIGO DE AUTENTICIDADE	Verifique a autenticidade deste documento na página www.dac.unicamp.br Código: 51dc96e174dee6e3ece3211e1893ee562ac3e219			
CÓDIGO DE AUTENTICIDADE							
Verifique a autenticidade deste documento na página www.dac.unicamp.br Código: 51dc96e174dee6e3ece3211e1893ee562ac3e219							
UNICAMP - Universidade Estadual de Campinas							
Este documento foi emitido em 27/03/2023 12:10	Página 1 de 1						

Yuri Tadeu Covo Zwaig

**Estudo sobre segurança e integridade do conteúdo das conversas
do WhatsApp e o uso como evidência de processos judiciais**

Monografia apresentada à Faculdade de
Tecnologia da Universidade Estadual de
Campinas como parte dos requisitos
exigidos para a obtenção do título de
Bacharel em Sistemas de Informação.

Orientadora: Profa. Dra. Marli De Freitas
Gomes Hernandez.

Limeira
2022

2 Trabalhos Correlatos

Para o desenvolvimento dessa pesquisa, foi realizada a revisão literária nas bases de dados da IEEE e da WOS. Estas bases de dados foram escolhidas por representarem um amplo conjunto de publicações com conteúdos relacionados à tecnologia da informação e segurança digital. Utilizou-se como argumento de pesquisa nos títulos, resumos e palavras-chave, duas combinações de termos em inglês: [i] "WhatsApp" e "Forensic" e [ii] "WhatsApp" e "Legal", restringindo-se apenas a publicações de 2015 até 2022 obtendo-se um total de 32 do IEEE e 99 do WOS.

Os arquivos com informações das publicações selecionadas foram baixados das bases de dados e importados para o aplicativo SW3T versão 2022 [4], software de apoio a revisão bibliográfica desenvolvido no projeto de pesquisa da FT e ainda em fase de homologação e teste.

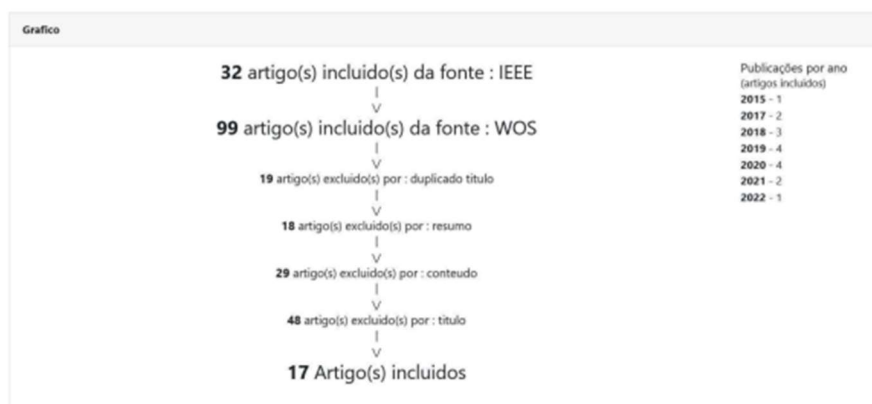


Figura 1 - Gráfico do SW3T com o histórico e justificativa dos arquivos excluídos

O resumo com resultado da análise das publicações e justificativas de exclusão no processo de revisão bibliográfica encontram-se na Figura 1 - Gráfico do SW3T com histórico e justificativa dos arquivos excluídos. Obteve-se no final da revisão 17 publicações de interesse para a pesquisa, onde foram excluídas: 19 publicações duplicadas, 48 publicações após a leitura do título, 18 publicações após leitura dos resumos e 29 publicações após análise do conteúdo.

A Tabela 1 - Relação final de artigos incluídos na pesquisa, apresenta o ano, autores e título da publicação, assim como a classificação de cada uma quanto ao motivo principal e sua contribuição. Foram identificados 4 motivos sendo: Artefatos e Ferramentas os artigos que tratam dos dados importantes para investigações forenses; Metodologia que descrevem sobre como são processados os artefatos; e Validade da Evidência os que discutem a validade das evidências do WhatsApp para efeito jurídico.

Entre estas publicações, destacam-se o de Umar et al [3] que utilizam 3 ferramentas forenses para a análise de artefatos do WhatsApp com base nos parâmetros NIST. Entre elas, a ferramenta "WhatsApp Key/DB Extractor" [5] utilizada nos testes com a versão de 2017 do WhatsApp mas que, não funciona nas versões de 2020 e 2022. As demais ferramentas deste artigo não são de fácil acesso inviabilizando sua utilização.