XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

ISSN 2177-3688

GT 8 – Informação e Tecnologia

INFORMAÇÃO E TECNOLOGIA NO GT 8 DO ENANCIB: UMA ABORDAGEM DE PROCESSAMENTO DE LINGUAGEM NATURAL

INFORMATION AND TECHNOLOGY AT ENANCIB WG 8: NATURAL LANGUAGE PROCESSING APPROACH

Fernanda Farinelli - Universidade de Brasília (UnB)
Felipe Augusto Arakaki - Universidade de Brasília (UnB) e Universidade Federal de São Carlos (UFSCar)
Fernando de Assis Rodrigues - Universidade Federal do Pará (UFPA)
Ana Carolina Simionato Arakaki - Universidade Federal de São Carlos (UFSCar) e Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)

Modalidade: Trabalho Completo

Resumo: Analisa a produção acadêmica do GT 8 do Enancib, focando em temas de Informação e Tecnologia de 2008 a 2023. O objetivo é compreender a complexidade do domínio do GT 8 e estabelecer procedimentos para construção semi automatizada de um vocabulário para seus pesquisadores, contribuindo para o entendimento dos temas recorrentes e suas inter-relações no contexto da informação e tecnologia. Metodologicamente, adotou-se uma abordagem quali-quantitativa, dividida em quatro etapas: coleta, higienização, processamento e análise de dados textuais usando técnicas de Processamento de Linguagem Natural (PLN). Os resultados mostram que os n-gramas mais frequentes são o unigrama 'informação', seguido pelos bigramas 'ciência informação', 'recuperação informação' e 'tecnologia informação'. Como considerações, destaca-se os desafios linguísticos e metodológicos enfrentados e sugerem aprofundamento em técnicas de PLN para futuras pesquisas.

Palavras-chave: Informação e Tecnologia; GT 8 - ENANCIB; Processamento de Linguagem Natural (PLN).

Abstract: Analyzes the academic production of WG 8 of Enancib, focusing on Information and Technology themes from 2008 to 2023. The aim is to understand the complexity of the WG 8 domain and establish procedures for the semi-automated construction of a vocabulary for its researchers, contributing to the understanding of recurring themes and their interrelationships in the context of information and technology. Methodologically, a qualitative-quantitative approach was adopted, divided into four stages: collection, sanitization, processing and analysis of textual data using Natural Language Processing (NLP) techniques. The results show that the most frequent n-grams are the 'information' unigram, followed by the 'information science', 'information retrieval' and 'information

technology' bigrams. As considerations, we highlight the linguistic and methodological challenges faced and suggest further study of NLP techniques for future research.

Keywords: Information and Technology; WG 8 - ENANCIB; Natural Language Processing (NLP).

1 INTRODUÇÃO

Desde 2008, o Grupo de Trabalho Informação e Tecnologia (GT 8) do Encontro Nacional de Pesquisa em Ciência da Informação (Enancib) se individualiza pelos

> [...] estudos e pesquisas teórico-práticos sobre e para o desenvolvimento de tecnologias de informação e comunicação que envolvam os processos de coleta, geração, representação, armazenamento, recuperação, disseminação, uso, reuso, gestão, análise, processamento, tratamento, governança, visualização, segurança e preservação de dados e informação em ambientes informacionais (ANCIB, 2024).

Neste escopo, a interdisciplinaridade da Ciência da Informação para outras áreas e disciplinas acontece de forma orgânica e com vários desafios, tais como os desafios terminológicos. Pontua-se que o conceito de interdisciplinaridade dentro da Ciência da Informação é explicitado por diversos autores, em especial por Saracevic (1995). Complementarmente, estudos como Santos *et al.* (2013) têm comprovado as relações entre Informação e Tecnologia em diferentes modalidades de comunicação científica, mostrando tanto as relações existentes quanto as tendências temáticas.

Os desafios terminológicos decorrentes pela interdisciplinaridade são constituídos pela diversidade de vocabulários entre as áreas para representar os mesmos conceitos. Isso pode ser um problema, pois, em alguns casos, utilizam-se termos similares para representar coisas diferentes ou termos diferentes para representar conceitos similares. Baker *et al.* (2011) já destacaram a necessidade de uma discussão conceitual da terminologia empregada em áreas interdisciplinares, tais como a Ciência da Informação e a Ciência da Computação. Pontua-se como problemática de pesquisa: como individualizar os termos e conceitos utilizados pelos pesquisadores do GT 8 de forma a refletir a complexidade e as inter-relações temáticas no domínio da informação e tecnologia?

A partir desse contexto, o objetivo do trabalho é compreender a complexidade do domínio do GT 8 e construir um vocabulário de termos semi-automatizado para que contribua

para o entendimento dos temas recorrentes e suas inter-relações no contexto da informação e tecnologia.

O GT 8 já foi tema de diversas pesquisas, como Santos *et al.* (2016), que realizaram um levantamento dos trabalhos apresentados durante o período de 2008-2015. A importância dessa análise leva a compreensão da trajetória do GT 8, no intuito de identificar as palavras-chave mais utilizadas, incidência do termo 'tecnologia', e mapeamento da distribuição geográfica da produção. O estudo caracterizou as afiliações institucionais e formação acadêmica dos pesquisadores, além de detectar a rede de citações para mapear aproximações teóricas e impactos dos trabalhos, corroborando com as atividades que visam fornecer uma visão abrangente sobre tendências, colaborações e impactos das pesquisas.

Em sequência, Santos *et al.* (2017) categorizam as temáticas das pesquisas que abordaram o termo tecnologia. Os autores (Santos *et al*., 2017) realizaram a análise de domínio das pesquisas publicadas no âmbito do GT 8 do ENANCIB partindo de sete categorias: teoria, desenvolvimento, uso, avaliação, políticas, ética e ensino, o que permitiu o mapeamento dos seguintes aspectos por categoria: autores, instituições mais produtivas e a quantidade de comunicações científicas em distribuição temporal.

Oliveira e Xavier (2017) analisaram a evolução das principais temáticas discutidas no GT 8 no período de 2008 a 2015, e traçaram tendências de pesquisa deste GT. Utilizando como método a modelagem de tópicos, os autores Souza, Izo Júnior e Souza (2019) identificaram os tópicos de maior relevância, as palavras e os pesos que constituem os tópicos do *corpus* de documentos do GT 8, além de discutir as respectivas relações entre os termos encontrados no *corpus* de dados e apresentar o comportamento dos termos mais frequentes. Complementar aos estudos anteriores, Yamane e Castro (2023) realizaram um mapeamento e análise dos trabalhos apresentados nos Enancib no período que abrange 2016 a 2022.

Contudo, nenhum dos trabalhos correlatos contemplam o objetivo aqui almejado, que é a construção de um vocabulário de termos semi-automatizado, refletindo a complexidade e as inter-relações dos conceitos no domínio da Informação e Tecnologia. Essa lacuna evidencia a necessidade de uma investigação mais aprofundada que não apenas analise as intersecções terminológicas, mas que também proponha soluções para facilitar a comunicação e a colaboração entre pesquisadores de diferentes áreas. Assim, na seção 2, será apresentada a fundamentação teórica sobre linguagem natural e na sequência, a metodologia proposta para a construção desse vocabulário, bem como as estratégias para a identificação e unificação dos

termos utilizados pelos integrantes do GT 8, ao decorrer dos resultados e considerações ao estudo.

2 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) é uma área interdisciplinar que envolve as áreas da Computação e da Linguística, visando desenvolver métodos para o processamento computacional da linguagem humana. Esse campo busca criar modelos e recursos linguísticos que automatizam o processamento da linguagem natural, permitindo que computadores entendam, gerem e extraiam conhecimento útil da linguagem humana (Caseli; Nunes; Pagano, 2024; Caseli; Freitas; Viola, 2022; Gonzalez; Lima, 2003; Pereira, 2011).

Entre os objetivos do PLN estão a recuperação de informação, tradução automática, interpretação de textos e realização de inferências. (Liddy, 2001; Vieira; Lopes, 2010). Na Recuperação de Informação, o PLN é fundamental para melhorar a eficiência dos sistemas de busca e análise de texto, permitindo que computadores compreendam e manipulem a linguagem natural, extraindo informações relevantes e facilitando a recuperação de dados úteis em grandes volumes de texto (Gonzalez; Lima, 2003; Moreira, 2024; Vieira; Lopes, 2010). O principal tipo de dado utilizado é um *corpus*, que é uma coleção de textos, como, por exemplo, o conjunto de textos do GT 8 de todas as edições do Enancib. Um *corpus* pode ser comparável (textos sobre o mesmo assunto), paralelo (versões em diferentes línguas) ou alinhado (*corpus* em paralelo com indicações de traduções correspondentes). Além disso, pode ser monolíngue, bilíngue ou multilíngue (Caseli; Freitas; Viola, 2022).

O texto é inicialmente visto como uma sequência de caracteres, que formam *tokens*, unidades linguísticas como palavras, delimitadas por espaços. A tokenização transforma a sequência de caracteres em *tokens*. Em português, palavras são geralmente delimitadas por espaços ou pontuação, facilitando o processo. *Tokens* se referem a cada ocorrência, enquanto os *types* contabilizam *tokens* únicos, indicando o tamanho do vocabulário de um *corpus* e sua cobertura linguística (Caseli; Freitas; Viola, 2022). Os unigramas representam palavras individuais ou *tokens*, os bigramas são formados por pares de palavras consecutivas, e os trigramas consistem em sequências de três palavras consecutivas, e assim sucessivamente.

Após a tokenização, ocorrem processamentos adicionais como lematização e radicalização. A lematização transforma as palavras em sua forma canônica, por exemplo, as

flexões "bonita" e "bonitos" são transformadas em "bonito". Já a radicalização, por outro lado, extrai a raiz da palavra, resultando em "bonit" para os mesmos exemplos. A relevância desses processamentos varia conforme a aplicação: algumas preferem formas canônicas ou raízes para reduzir o tamanho do vocabulário, enquanto outras se beneficiam de considerar todas as formas flexionadas do *corpus* (Caseli; Freitas; Viola, 2022).

A remoção de caracteres especiais como números e pontuação e das *stopwords* é um passo importante em muitas aplicações de PLN. *Stopwords* são palavras que pouco acrescentam ao conteúdo do texto, como preposições, determinantes e conjunções, como exemplo: para, aquela, mas - respectivamente. Desconsiderar essas palavras e caracteres especiais ajuda a focar nas partes mais relevantes para a análise, melhorando a eficiência dos algoritmos de processamento (Caseli; Freitas; Viola, 2022).

Outro aspecto importante do PLN é modelagem de n-gramas, uma técnica estatística que estima a probabilidade de ocorrência de uma palavra com base nas n-1 palavras que a precedem, formando sequências denominadas n-gramas, como por exemplo unigramas (n=1), bigramas (n=2), trigrams (n=3) e quadrigramas (n=4) (Jurafsky; Martin, 2009). Modelos n-gram são úteis porque as relações lexicais, semânticas e sintáticas em contextos locais são bons preditores da próxima palavra, mesmo sem considerar a estrutura completa da frase (Manning; Schütze, 1999).

Nesse contexto, a modelagem de tópicos é uma técnica estatística usada para organizar e resumir grandes coleções de documentos eletrônicos, aplicando métodos de *Machine Learning* para identificar padrões e agrupamentos semânticos dentro de textos. Isso facilita a compreensão e a organização de grandes volumes de dados textuais, identificando temas principais sem a necessidade de leitura manual de todos os documentos. Kherwa e Bansal (2019, p. 1) destacam que a modelagem "[...] inclui a hierarquia de classificação, métodos de modelagem de tópicos, técnicas de inferência posterior, diferentes modelos de evolução de *Latent Dirichlet Allocation* (LDA) e suas aplicações em diferentes áreas da tecnologia". Souza e Souza (2021, p. 7) afirmam que "[...] a modelagem de tópicos permite organizar e resumir grandes coleções de arquivos eletrônicos por meio de métodos estatísticos e algoritmos de *machine learning*".

Esse método, que pode ser realizado de forma não supervisionada, extrai automaticamente palavras importantes dos documentos, sem conhecimento prévio dos assuntos abordados ou rótulos predefinidos (Blei; Ng; Jordan, 2003; Blei, 2012). O processo

envolve o treinamento de algoritmos com um conjunto de dados para fazer previsões e explorar estruturas semânticas, e os modelos probabilísticos como o LDA representam documentos como combinações de tópicos, onde cada tópico é uma distribuição probabilística de palavras. Esse processo inclui três etapas principais: inserção de cada palavra em um documento, escolha aleatória de um tópico com base na distribuição de tópicos e seleção de uma palavra tópica (Steyvers; Griffiths, 2007). Na modelagem de tópicos, os n-gramas facilitam a identificação de padrões frequentes e o agrupamento de termos relacionados, melhorando a precisão de técnicas como LDA (Blei, Ng; Jordan, 2003).

Após definir o número de tópicos a serem abordados, a modelagem de tópicos determina os termos relevantes para cada documento, utilizando um modelo generativo para estimar a quantidade de termos a partir das variáveis observadas nos documentos (Santos, 2015). O LDA adota uma abordagem *bayesiana*, os documentos são representados por misturas de tópicos latentes. Cada tópico é caracterizado por uma distribuição de palavras presentes nos documentos. Ao aplicar o LDA, os tópicos são interpretados como temas da coleção de documentos, e as representações dos documentos refletem esses temas, organizando os *corpus* de maneira eficiente (Blei, 2012; Chaney; Blei, 2012).

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa adotou uma abordagem quantitativa-qualitativa, de natureza aplicada, de objetivo exploratório e com procedimento experimental, com foco na análise de conteúdo textual para identificar e interpretar padrões e temas subjacentes nos dados coletados. A metodologia foi estruturada em quatro etapas: identificação e coleta; higienização; processamento automático; e análise.

Etapa 1. Identificação e Coleta do *corpus*: A primeira etapa envolveu a identificação e coleta dos textos das comunicações científicas publicadas pelo GT 8. Dados de pesquisas anteriores, como os levantamentos realizados por Santos (2015, 2016) e por Yamane e Castro (2023), foram utilizados para facilitar essa operacionalização. Os textos foram coletados em formato *Portable Document Format* (PDF) e, subsequentemente, convertidos para formato *Text File Format* (TXT), por meio do comando *pdftotext* para *GNU/Linux*, parte da biblioteca *Poppler*[1].

[1] FreeDesktop.org. **Poppler.** [S.l], 2024. Disponível em: https://poppler.freedesktop.org. Acesso em: 12 jul. 2024.

Etapa 2. Higienização preliminar do *corpus* coletados: Para otimizar o tratamento dos dados, foi realizada uma higienização inicial dos textos. Informações não essenciais foram removidas, tais como autoria, e-mails, referências bibliográficas, cabeçalhos, resumos e textos integralmente em língua estrangeira. Esse processo garantiu que o *corpus* estivesse preparado para a próxima etapa, focando apenas no conteúdo relevante para esta pesquisa.

Etapa 3. Processamento automático do *corpus*: A técnica de Processamento de Linguagem Natural (PLN) foi empregada na sistematização, isolamento e identificação de palavras contidas no *corpus*. Uma das tarefas do PLN envolveu o desenvolvimento do Projeto de Extração de n-grams, no qual se gerou a relação de unigramas, bigramas e trigramas, por meio de um algoritmo em Python (versão 3.10). As bibliotecas e ferramentas empregadas foram: *pt-core-news-lg* (versão 3.7.x); *SQLAlchemy* (versão 2.x); *spacy* (versão 3.7.x); *spacy-ngram* (versão 0.0.3) e *tqdm* (versão 4.66.x). O algoritmo foi configurado para utilizar *SQLite3* como Sistema de Gerenciamento de Banco de Dados, com uma instância pré-configurada para facilitar a execução (Figura 1). A integração dos dados com o algoritmo foi automatizada por meio do dialeto *SQLAlchemy*[2]. Neste primeiro momento, optou-se pela remoção de números e caracteres especiais e a não remoção de *stopwords* no curso do processamento. Não foram aplicadas às tarefas de *lematização e stemming*. A aplicação no algoritmo possibilitou a extração de n-gramas relevantes, gerando como saída três arquivos em formato *Comma-Separated Values* (CSV) para a análise: cada um contendo, respectivamente unigramas, bigramas e trigramas, segmentados pela ocorrência dos mesmos por ano e total.

Etapa 4. Análise dos n-gramas: Para avaliar os n-gramas e suas respectivas frequências foi utilizado a combinação das ferramentas *Microsoft Excel* e *Power BI*. Adicionalmente, foi necessário fazer nova rodada de higienização, com a finalidade de realizar um tratamento manual para remoção de n-gramas frequentes e não relevantes para análise como, por exemplo, *apud*, comunicação oral, encontro nacional pesquisa e do autor.

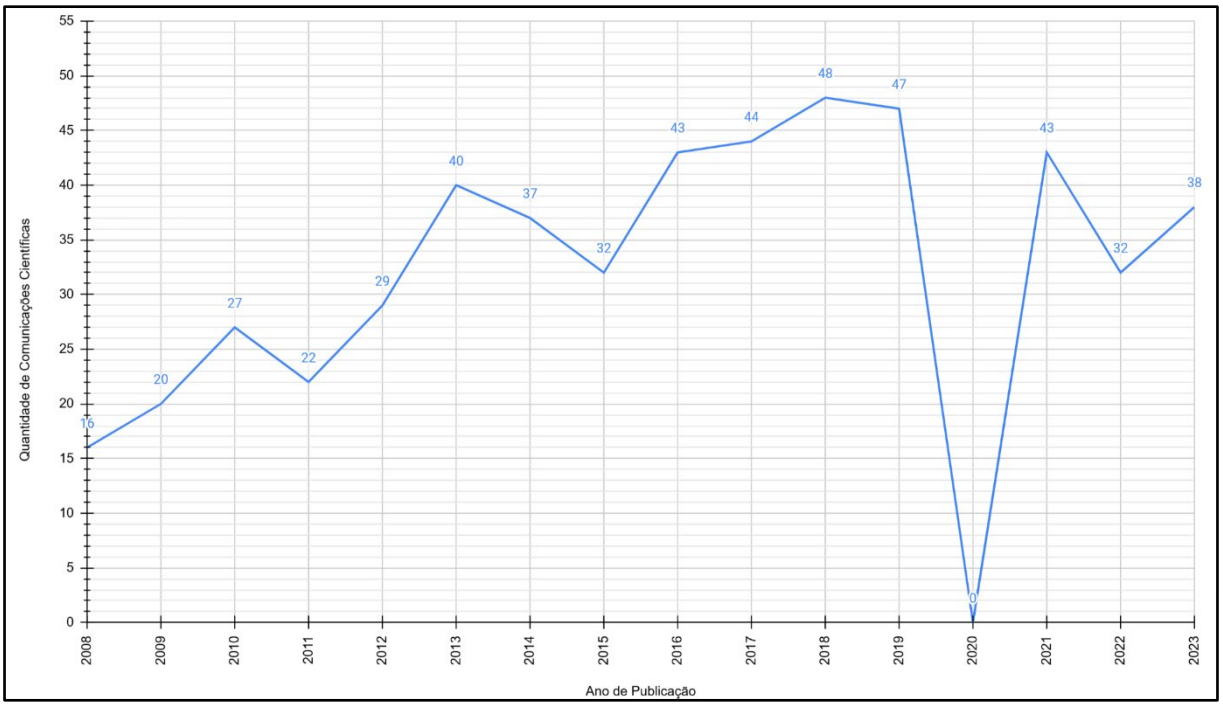
[2] O algoritmo está disponível para acesso no *GitHub*, pelo endereço eletrônico https://github.com/rodriguesprobr/ngram_oracao_extrator. Acesso em: 19 de set. 2024.

4 DESCRIÇÃO E ANÁLISE DOS RESULTADOS

A identificação de unigramas, bigramas e trigramas nas comunicações científicas demonstra a complexidade e a riqueza dos dados analisados, permitindo uma compreensão mais detalhada das tendências e padrões presentes nos textos acadêmicos.

A distribuição anual das comunicações científicas ao longo dos anos, indica uma variação na produção acadêmica: em 2008, foram apresentados 16 trabalhos; em 2009, 20 trabalhos; em 2010, 27 trabalhos; em 2011, 22 trabalhos; em 2012, 29 trabalhos; em 2013, 40 trabalhos; em 2014, 37 trabalhos; em 2015, 32 trabalhos; em 2016, 43 trabalhos; em 2017, 44 trabalhos; em 2018, 48 trabalhos; em 2019, 47 trabalhos; em 2021, 43 trabalhos; em 2022, 32 trabalhos; e em 2023, 38 trabalhos. Destaca-se que não houve a realização do Enancib no ano de 2020 devido à pandemia de Covid-19. Essa análise evidencia a crescente produção científica ao longo dos anos, com um aumento significativo no número de comunicações a partir de 2013, conforme apresentado pela Figura 1.

Figura 1 - Distribuição anual das comunicações científicas do GT 8



Fonte: Elaborado pelos autores (2024).

O levantamento identificou um total de 518 comunicações científicas, nas quais foi possível identificar 396.894 vezes com os uni-gramas, 913.334 vezes com os bigramas e

1.002.102 vezes com os trigramas, demonstrando a frequência e a relevância desses elementos textuais nos documentos analisados.

Ressalta-se que para o estudo e análise de dados, foram ignorados os n-gramas que não atingiram a frequência mínima total de 20 ocorrências. Essa abordagem foi adotada para garantir que apenas os termos e suas combinações mais relevantes e representativos fossem considerados na análise. Ao estabelecer esse critério de inclusão, foi possível focar nos n-gramas que têm um impacto no *corpus*, evitando a inclusão de termos menos relevantes ou esporádicos que poderiam distorcer os resultados e as conclusões do estudo. A Tabela 1 exibe as ocorrências dos cinco unigramas, bigramas e trigramas mais frequentes, segmentados por ano de publicação. O processo de contabilização dos n-gramas, foi considerado para essa análise, a ocorrência de pelo menos uma vez no trabalho.

Tabela 1 - Ocorrências dos Top 5 n-gramas, segmentados por ano de publicação

n-grama	Ocorrências por Ano															Total
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2021	2022	2023	
informação	16	20	27	22	29	40	37	32	43	44	48	47	43	32	38	**518**
dado	15	20	26	22	28	37	37	32	42	44	45	47	42	31	38	**506**
palavra	16	19	26	22	28	39	37	32	42	43	45	44	41	32	37	**503**
objetivo	16	20	26	22	28	37	35	31	43	44	44	46	40	32	37	**501**
uso	16	18	24	22	29	39	36	30	42	44	47	42	40	31	38	**498**
Outros unigramas	*10903*	*13768*	*15539*	*13961*	*18775*	*22235*	*21939*	*21155*	*26710*	*27504*	*27801*	*28234*	*23187*	*17821*	*22228*	***311760***
ciência informação	15	18	18	17	24	32	27	25	33	33	32	37	34	27	28	**400**
base dado	9	11	10	11	13	18	22	22	24	23	26	29	28	25	23	**294**
recuperação informação	10	11	14	18	21	21	21	16	25	23	20	24	12	18	20	**274**
tecnologia informação	7	13	14	14	17	27	20	17	22	26	20	24	17	16	16	**270**
acesso informação	9	12	13	13	16	16	13	16	17	22	20	20	19	12	17	**235**
Outros bigramas	*2419*	*2991*	*3565*	*3442*	*4672*	*5416*	*5502*	*5149*	*7292*	*7271*	*7022*	*7529*	*5770*	*4617*	*5708*	***78365***
tecnologia informação comunicação	4	9	8	12	12	18	14	11	16	18	14	16	9	11	11	**183**
ambiente informacional digital	3	2	5	7	10	7	6	6	9	11	12	13	6	5	9	**111**
world wide web	2	3	4	4	9	4	8	11	10	12	6	12	7	5	5	**102**
informação ambiente digital	0	1	2	3	4	4	2	3	7	2	7	8	6	6	7	**62**
resource description framework	0	1	1	1	5	5	4	5	7	9	7	5	4	1	4	**59**
Outros trigramas	*123*	*142*	*202*	*152*	*283*	*309*	*298*	*317*	*590*	*645*	*542*	*558*	*485*	*406*	*429*	***5481***
Total	**13583**	**17079**	**19524**	**17765**	**24003**	**28304**	**28058**	**26910**	**34974**	**35818**	**35758**	**36735**	**29790**	**23128**	**28693**	**400122**

Fonte: Elaborado pelos autores (2024).

Na Tabela 1, também é possível visualizar um mapa no estilo de calor com a distribuição de ocorrências anuais dos n-gramas mais frequentes no GT 8 da Enancib. Um mapa de calor em uma análise de Processamento de Linguagem Natural (PLN) foi empregado para visualizar padrões de frequência e co-ocorrência de termos em um *corpus*, para facilitar a identificação de termos-chave, revelar tendências temporais, e destacar áreas de interesse e sub-temas emergentes. No caso, a coloração varia de acordo com a frequência de ocorrência dos termos: os valores mais elevados são representados por tons de vermelho, enquanto os valores menores são indicados por tons de azul. Essa representação visual facilita a identificação de padrões e tendências ao longo do tempo.

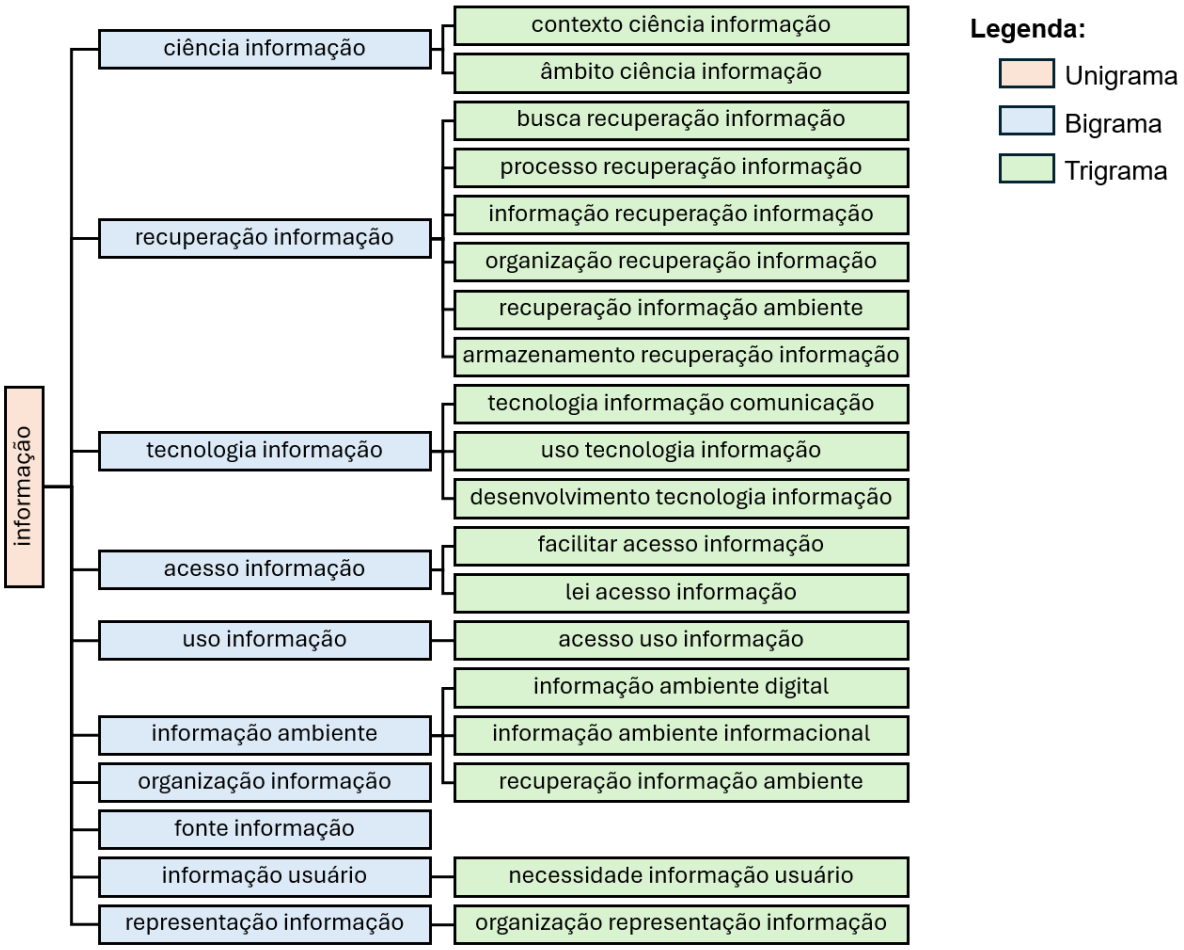
A correlação entre o mapa de calor da Tabela 1 e a quantidade de comunicações científicas aceitas por ano da Figura 1, pode ser observada ao analisar as tendências nos dois conjuntos de dados. A Figura 1 mostra a série histórica sobre a evolução do número de artigos aceitos no Enancib, enquanto o mapa de calor revela a frequência dos unigramas mais comuns nos artigos aceitos durante o mesmo período. A partir da Figura 1, entre 2016 e 2019, observa-se uma tendência de maior ocorrência de termos, o que também indica um aumento no número de submissões ao GT 8. Este período é marcado por um pico na quantidade de comunicações científicas aceitas.

Ao visualizar as relações entre unigramas, bigramas e trigramas, é possível obter uma compreensão mais profunda das áreas temáticas e dos contextos em que a palavra *informação* é central. A Figura 2 demonstra uma hierarquia de termos derivados do unigrama *informação* que aparece como o mais frequente no *corpus* estudado. A estrutura hierárquica desdobra-se a partir de *informação* para identificar os dez bigramas mais recorrentes que utilizam este termo, bem como os trigramas associados a cada um desses bigramas. Especificamente, para os bigramas *organização informação* e *fonte informação*, nenhum dos trigramas derivados atingiu a frequência mínima de 20 ocorrências totais. Esta organização hierárquica permite a visualização de como os termos estão relacionados entre si e como a complexidade semântica aumenta conforme se adicionam mais palavras ao termo original.

A organização dos n-gramas em uma estrutura hierárquica, conforme apresentado, pode ser aplicada em metodologias de análise de tópicos, ajudando na identificação de termos-chave e suas co-ocorrências, mapeando tematicamente áreas de interesse do campo de Ciência da Informação. Essa abordagem facilita a compreensão das relações semânticas

entre termos, destacando subtemas e áreas específicas de pesquisa, promovendo uma análise mais estruturada e detalhada do *corpus* estudado.

Figura 2 - Hierarquia de Ocorrências de Termos Relacionados a "Informação" em artigos do Enancib



Fonte: Elaborado pelos autores (2024).

A Tabela 2 detalha as tendências e padrões na pesquisa de n-gramas ao longo dos anos, com *informação* e suas variações dominando o campo, refletindo seu papel central. A Tabela apresenta a ocorrência dos diferentes n-gramas da Figura 2 ao longo dos anos. A coluna à esquerda lista os n-gramas (e os categoriza de acordo com a legenda), e as colunas seguintes representam a frequência anual de cada n-grama, culminando em um total de ocorrências na última coluna.

Tabela 2 - Distribuição de ocorrências por ano dos n-gramas mais frequentes

N-gram[1]		Ocorrência por Ano															Total
		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2021	2022	2023	
U	informação	16	20	27	22	29	40	37	32	43	44	48	47	43	32	38	518
B	ciência informação	15	18	18	17	24	32	27	25	33	33	32	37	34	27	28	400
T	contexto ciência informação	1	3	2	1	1	2	2	1	3	5	7	6	5	2	7	48
T	âmbito ciência informação	0	4	2	2	2	5	4	3	5	3	2	4	3	4	2	45
B	recuperação informação	10	11	14	18	21	21	21	16	25	23	20	24	12	18	20	274
T	busca recuperação informação	0	0	2	3	4	3	3	2	7	5	4	2	3	5	7	50
T	processo recuperação informação	1	0	4	2	5	2	2	3	7	5	0	1	1	5	3	41
T	informação recuperação informação	2	2	2	1	3	0	5	2	3	1	2	1	1	2	6	33
T	organização recuperação informação	0	2	5	1	3	5	2	2	3	0	1	0	2	5	2	33
T	recuperação informação ambiente	0	2	1	1	3	2	4	1	3	2	0	0	1	0	3	23
T	armazenamento recuperação informação	0	2	1	0	2	5	3	1	1	1	2	1	1	0	1	21
B	tecnologia informação	7	13	14	14	17	27	20	17	22	26	20	24	17	16	16	270
T	tecnologia informação comunicação	4	9	8	12	12	18	14	11	16	18	14	16	9	11	11	183
T	uso tecnologia informação	1	4	1	2	3	5	1	2	2	6	3	5	2	2	1	40
T	desenvolvimento tecnologia informação	2	1	0	2	1	1	3	0	2	3	3	2	0	2	2	24
B	acesso informação	9	12	13	13	16	16	13	16	17	22	20	20	19	12	17	235
T	facilitar acesso informação	1	1	4	0	3	1	3	2	0	4	2	1	0	1	1	24
T	lei acesso informação	0	0	0	0	2	3	0	4	2	1	1	2	3	1	4	23
B	uso informação	1	13	7	10	11	10	15	11	11	12	13	16	7	8	10	155
T	acesso uso informação	0	4	3	0	5	5	3	3	1	4	2	6	2	3	2	43
B	informação ambiente	4	6	6	8	8	9	10	5	13	11	14	17	7	9	11	138
T	informação ambiente digital	0	1	2	3	4	4	2	3	7	2	7	8	6	6	7	62
T	informação ambiente informacional	4	1	1	1	3	4	4	2	4	4	2	3	3	2	3	41
T	recuperação informação ambiente	0	2	1	1	3	2	4	1	3	2	0	0	1	0	3	23
B[2]	organização informação	3	8	7	11	10	12	10	7	14	8	10	13	7	10	8	138
B[2]	fonte informação	5	7	9	7	7	7	10	3	14	9	11	15	9	5	13	131
B	informação usuário	6	8	13	7	9	7	9	7	11	7	8	14	9	10	4	129
T	necessidade informação usuário	3	1	3	3	2	2	3	1	3	2	2	2	1	3	0	31
B	representação informação	4	4	6	6	8	16	8	9	11	10	12	10	8	11	5	128
T	organização representação informação	0	0	0	0	2	1	3	2	1	3	2	1	1	6	1	23

Nota: [1] U - Unigrama, B - Bigrama, T - Tigrama. [2] O bigrama não aparece em nenhum trigrama em que a frequência do trigrama é, no mínimo, 20 ocorrências.
Fonte: Elaborado pelos autores (2024).

Os picos de publicação entre 2018 e 2020 destacam períodos de intensa atividade e interesse, seguidos por uma leve queda, possivelmente indicando mudanças nas áreas de foco e ao analisar esses dados auxilia na compreensão das dinâmicas e evoluções na área de Ciência da Informação. Os n-gramas mais frequentes são o unigrama *informação*, com 518 ocorrências, seguido pelos bigramas *ciência informação*, *recuperação informação* e *tecnologia informação*, respectivamente com 400, 274 e 270 ocorrências, e entre os trigramas, *tecnologia informação comunicação* é o mais frequente com 183 ocorrências.

A análise da tabela 2 revela que, em 2013, o bigrama *representação informação* apresenta um pico, sugerindo como tópico de pesquisa e depois tornou-se mais estável sua frequência. Já o bigrama *recuperação informação* mostra um aumento gradativo ao longo dos anos, indicando uma tendência a estar sempre presente nas publicações do GT. Esses padrões destacam como alguns tópicos emergem rapidamente e depois se estabilizam, enquanto outros mantêm uma presença constante, refletindo mudanças nas áreas de foco e interesse contínuo nas pesquisas.

A tabela 2 revela que o trigrama *lei acesso informação* só é citado a partir de 2012, coincidindo com a promulgação da Lei nº 12.527 (18 de novembro de 2011) que popularmente ficou conhecida como Lei de Acesso à Informação. Apesar dessa correlação sugerir que a introdução da legislação impulsionou pesquisas e discussões sobre o tema, observa-se que ele ainda não é uma temática amplamente discutida no GT, indicando que há espaço para um aprofundamento maior nas pesquisas sobre essa lei.

A popularidade dos termos mostra que *informação* é claramente um termo central na área, dado o seu alto número de ocorrências em várias combinações. No entanto, a alta frequência do termo *informação* pode ser influenciada pelo bigrama *ciência informação*, que é o mais frequente na tabela. Considerando que o *corpus* analisado é de um evento específico da área, há uma tendência natural para que um bigrama que represente a Ciência da Informação esteja sempre presente. Portanto, ao desconsiderar esse bigrama nas análises futuras, os resultados poderiam ser diferentes, revelando outros termos e n-gramas que possam indicar novas tendências e mudanças específicas no campo de estudo.

5 CONSIDERAÇÕES FINAIS

A análise textual em português no contexto dos trabalhos apresentados no GT 8 do Enancib enfrenta diversos desafios, relacionados tanto à complexidade linguística do idioma português quanto às especificidades do ambiente acadêmico. A preparação do texto é uma etapa que requer atenção a vários aspectos.

A primeira necessidade é a remoção de elementos que podem interferir nos algoritmos de PLN, como *hyperlinks*, numeração, acentuação e caracteres especiais. Além disso, é essencial eliminar as *stopwords*, que são palavras comuns e geralmente não carregam significado relevante para a análise. No entanto, o ambiente acadêmico requer um dicionário adicional de palavras específicas a serem removidas, como *et al.*, *apud*, *citado por*, *grifo nosso*, *tabela*, *figura*, *gráfico*, nomes de autores, citações e siglas das instituições de vínculo dos autores. Adicionalmente, o padrão acadêmico de uso de siglas em vez de expressões completas requer atenção especial. Por exemplo, *SOC* para *Sistema de Organização do Conhecimento*, e casos semelhantes, devem ser corretamente identificados e associados à expressão completa para evitar a fragmentação dos dados analisados.

Outro desafio significativo é a padronização das palavras para singular e plural, masculino e feminino, e flexões verbais. Termos como *técnica* e *técnicas*, *informação* e *informações*, *ontologia* e *ontologias* devem ser unificados. Flexões verbais, como *recuperação*, *recupera*, *recuperar* e *recuperam*, também precisam ser padronizadas, assim como variações de gênero, por exemplo, *autor* e *autora*. Essa padronização é essencial para garantir a consistência e a precisão na análise de tópicos.

Além disso, a remoção de elementos que podem ser relevantes para a análise apresenta um desafio. Por exemplo, a remoção de numeração pode afetar partes do léxico MARC21, um padrão de catalogação amplamente utilizado. Outro desafio é lidar com diferentes formas de dizer a mesma coisa. É comum encontrar trigramas semelhantes, como *padrão dublin core*, *metadado dublin core*, ou o bigrama *padrão dc*, todos se referindo ao mesmo tópico. Para enfrentar esse desafio, é necessário implementar um sistema de reconhecimento e unificação de termos semelhantes. A sequência de passos usados na preparação de dados deve ser cuidadosamente analisada, pois um passo pode impactar diretamente no outro. Por exemplo, a remoção de pontuação antes da eliminação de

hyperlinks pode afetar a remoção correta dos *hyperlinks*, que comumente possuem elementos de pontuação na sua composição.

Como trabalhos futuros pretende-se realizar um aprofundamento das técnicas de PNL e a realização da modelagem de tópicos para o estabelecimento dos conceitos apresentados no âmbito do GT 8.

REFERÊNCIAS

Associação de Pesquisa e Pós-graduação em Ciência da Informação. **GTs - XXIV ENANCIB.** Espírito Santo, 2024. Disponível em: https://ancib.org/sites/enancib2024/index.php/gts/. Acesso em: 12 jul 2024.

BAKER, T. *et al*. **Library Linked Data Incubator Group Final Report**. [*S.l.*]: W3C Incubator Group Report, 2011. Disponível em: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/. Acesso em: 8 set. 2019.

BLEI, David M. Probabilistic topic models. **Communications of the ACM**, v. 55, n. 4, p. 77-84, 2012.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine learning research**, v. 3, n. Jan, p. 993-1022, 2003.

CASELI, H. de M.; NUNES, M. das G. V.; PAGANO. A. **O que é PLN?**. *In*: CASELI, H. de M.; NUNES, M. das G. V. (org.). **Processamento de linguagem natural**: conceitos, técnicas e aplicações em português. 2.ed. São Carlos: BPLN, 2024. p. 10–16. Disponível em: https:// brasileiraspln.com/livro-pln/2a-edicao/parte-introducao/cap-introducao/cap-introducao.html. Acesso em: 10 jul. 2024.

CASELI, H.; FREITAS, C.; VIOLA, R. Processamento de linguagem natural. *In*: BRAZILIAN SYMPOSIUM ON DATA BASES, 37., 2022, Búzios. **Short courses of the […]**. Búzios: SBC, 2022. Disponível em: https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/103/460/732?inline=1. Acesso em: 10 jul. 2024.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de informação e processamento da linguagem natural. *In*: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas, SP. **Anais do [...]**. Campinas, SP: SBC, 2003. p. 347–395. Disponível em: https://www.academia.edu/download/43022678/minicurso-jaia2003.pdf. Acesso em: 10 jul. 2024.

JURAFSKY, D.; MARTIN, J. H. N-gram Language Models. *In*: SPEECH and language processing: an introduction to natural language processing, computational linguistics, and speech recognition (draft of february 3, 2024). 3.ed. [*S. l.*]: Stanford University, 2024. Cap. 3, p. 32–59.

KHERWA P., BANSAL P. Topic Modeling: a comprehensive review. **EAI endorsed transactions on scalable information systems,** n. 7, v. 24, p. 1–16, 2019. Disponível em: https://eudl.eu/pdf/10.4108/eai.13-7-2018.159623. Acesso em: 10 jul. 2024.

LIDDY, E. D. Natural Language Processing. In: ENCYCLOPEDIA of Library and Information Science. 2. ed. New York: Marcel Decker, Inc., 2003. Disponível em: https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub. Acesso em: 10 jul. 2024.

MANNING, C. D.; HINRICH SCHÜTZE. **Foundations of statistical natural language processing**. Massachusetts: The MIT Press, 1999.

MOREIRA, V. P. Recuperação de Informação. *In*: CASELI, H. de M.; NUNES, M. das G. V. (org.). **Processamento de linguagem natural**: conceitos, técnicas e aplicações em português. 2.ed. São Carlos: BPLN, 2024. cap. 19, p. 454–475. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao/parte-introducao/cap-introducao/cap-introducao.html. Acesso em: 10 jul. 2024.

OLIVEIRA, H. P. C. DE; XAVIER, T. N. Tendências de pesquisa em informação e tecnologia: análise do GT 8 no Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação. **Folha de Rosto**, Juazeiro do Norte, v. 3, n. 2, p. 76-87, 27 dez. 2017.

PEREIRA, S. do L. **Processamento de linguagem natural**. [*S. l.: s. n.*], 2011. Disponível em: https://www.ime.usp.br/~slago/pl-12.pdf. Acesso em: 10 jul. 2024.

SANTOS, P. L. V. A. C. ; ROMANETTO, L. M. ; ARAKAKI, F. A. ; CONEGLIAN, C. S. ; GONÇALEZ, P. R. V. A. ; SIMIONATO, A. C. ; RODRIGUES, F. A. Informação e Tecnologia: percurso temático do GT 08. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais [...]**. Marília: ANCIB, 2017.

SANTOS, P. L. V. A. C. *et al.* Informação e tecnologia no ENANCIB: percurso do GT 8 no período de 2008-2015. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais** [...]. Salvador: ANCIB, 2016. Disponível em: http://www.ufpb.br/evento/index.php/enancib2016/enancib2016/paper/viewFile/3618/2583. Acesso em: 9 jun. 2023.

SANTOS, P. L. V. A. C. *et al.* Mapeamento do termo Tecnologia em periódicos da CI no escopo do GT-8: Informação e Tecnologia. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Florianópolis. **Anais [...].** Rio de Janeiro: ANCIB, 2013. p. 1–19. Disponível em: http://enancib.ibict.br/index.php/enancib/xivenancib/paper/viewFile/4372/3495. Acesso em: 1 jan. 2017.

SARACEVIC, T. A natureza interdisciplinar da ciência da informação. **Ciência da Informação**, Brasília, v. 24, n. 1, p. 36–41, 1995.

SOUZA, M.; IZO JÚNIOR, A.; SOUZA, R. Modelagem de Tópicos: mapeamento científico do GT-8 do Enancib. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 20., 2019, Florianópolis. **Anais [...].** Florianópolis: ANCIB, 2019. p. 1–19. Disponível em: https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/490. Acesso em: 12 Jul. 2024.

SOUZA, M.; SOUZA, R. R. Mapeamento de conhecimento científico: modelagem de tópicos das teses e dissertações do Programa de Pós-Graduação em Ciência da Informação da

UFMG. **Em Questão**, Porto Alegre, v. 27, n. 3, p. 228–250, 2021. DOI: 10.19132/1808-5245273.228-250. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/104211. Acesso em: 13 jul. 2024.

VIEIRA, R.; LOPES, L. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. *In*: PERNA, Cristina Lopes; DELGADO, Heloísa Koch; Finatto, Maria José (org.). **Linguagens especializadas em corpora**: modos de dizer e interfaces de pesquisa. 1.ed. Porto Alegre: EDIPUCRS, 2010. p. 183–201. Disponível em: https://www.academia.edu/download/50033978/linguagensespecializadasemcorpora.pdf#page=184. Acesso em: 10 jul. 2024.