XX ENANCIB
21 a 25 Outubro/2019 – Florianópolis
A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-8 – Informação e Tecnologia

MODELAGEM DE TÓPICOS: MAPEAMENTO CIENTÍFICO DO GT-8 DO ENANCIB

TOPIC MODELING: SCIENTIFIC MAPPING OF ENANCIB GT-8

Marcos de Souza - Universidade Federal de Minas Gerais (UFMG)
Antonio Izo Júnior – Faculdade São Camilo – Rio de Janeiro (FASCRJ)
Renato Rocha Souza - Universidade Federal de Minas Gerais (UFMG) e Fundação Getúlio Vargas (FGV)

Modalidade: Trabalho Completo

Resumo: Com crescente volume de informações, tem sido cada vez mais necessário o uso de ferramentas computacionais para organizar, recuperar e compreender tal quantitativo de informações. A modelagem de tópicos tem possibilitado realizar essas atividades, com algoritmos de *Machine Learning,* que utilizam métodos estatísticos, por meio de uma estrutura não supervisionada em documentos eletrônicos que constituem *corpora* de dados, buscando assim, analisar e descobrir temas e suas respectivas relações. A pesquisa possui como objetivo geral identificar os tópicos de maior relevância do *corpus* de dados constituído por artigos completos e resumos expandidos do grupo de trabalho número oito do Encontro Nacional de Pesquisa em Ciência da Informação, denominado GT-8: Informação e Tecnologia. Além disso, discutiu-se sobre o comportamento dos termos mais frequentes encontrado no *corpus* de dados entre o período analisado. A modelagem de tópicos foi realizada por meio de algoritmo de *Machine Learning* que utilizou o método probabilístico *Latent Dirichlet Allocation*. Como resultado, foi possível identificar 10 tópicos constituídos por um conjunto de palavras e pesos que melhor representam o *corpus* de dados, bem como as relações desses tópicos. O tópico 1 foi destacado como o de maior relevância de todo o *corpus*, apresentando assim, 40,5% dos tokens. O termo 'informação' apresentou maior frequência no *corpus* com 1327, 1630, 5052, 1606, 2254, 1962 e 2300 repetições entre os anos de 2012 a 2018. Tal mapeamento científico do comportamento dos termos encontrados no *corpus* possibilita criar ações para futuras contribuições práticas, metodológicas e científicas da pesquisa.

Palavras-Chave: Modelagem de tópicos; Mapeamento científico; Alocação de Dirichlet Latente; Informação e tecnologia.

Abstract: With a growing volume of information, it is increasingly necessary to use computational tools to organize, retrieve and understand quantitative information. Topic modeling can perform these activities with Machine Learning algorithms that use statistical methods, through an unsupervised structure in documents that use data corpora, searching, analyzing and discovering themes and their relationships. A research has as its general objective to identify the most relevant topics for corpus data consisting of complete articles and summaries of the working group number eight of the Encontro Nacional de Pesquisa em Ciência da Informação, called GT-8: Informação e Tecnologia. In addition, we discuss the behavior of terms most frequently found without data corpus between the analyzed period. An outline model was performed using the Machine Learning algorithm that used the probabilistic Latent Direct Allocation method. As a result, it was possible to identify 10 topics consisting of a set of words and weights that best represent the data corpus, as well as the relationships of these topics. Topic 1 was highlighted as the largest relevance of the entire corpus, thus presenting 40.5% of tokens. The term 'information' is more frequent in the corpus with 1327, 1630, 5052, 1606, 2254, 1962 and 2300 repetitions from 2012 to 2018. Such scientific mapping of the behavior of terms found in the corpus enables the use of actions for studied, methodological and scientific research practices.

Keywords: Topic modeling; Scientific mapping; Latent Dirichlet Allocation; Information and technology.

1 INTRODUÇÃO

Com crescente volume de informações disponibilizados no ciberespaço, tem sido cada vez mais necessário o uso de ferramentas computacionais para organizar, recuperar e compreender tal quantitativo de informações. A comunicação científica tem a sua parcela de contribuição na produção e disseminação de informações científicas por meio de canais formais ou informais. São exemplos desses canais: teses e dissertações publicadas em bibliotecas digitais; resumos; resumos expandidos; artigos completos publicados em anais de eventos ou periódicos científicos; livros disponibilizados em diferentes plataformas; atas de reuniões; relatórios de pesquisas; trabalho de conclusão de curso; dentre outros formatos.

O Grupo de Trabalho (GT) 8 do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) iniciou suas atividades no ano de 2008. O GT realiza estudos sobre Informações e Tecnologias, teóricos e práticos que envolvam o desenvolvimento de tecnologias de informação e comunicação. Entre os anos de 2012 e 2018 o GT-8 produziu 273 pesquisas científicas nos formatos de resumos expandidos e artigos completos. Esse quantitativo de trabalhos, mesmo fazendo parte de um único GT, constitui um leque amplo de informações, uma vez que, a Ciência da Informação é considerada uma área interdisciplinar.

Organizar e resumir informações de forma manual, mesmo em se tratado de um determinado período, de um único GT, é algo cansativo e refutável, na maioria das vezes para

a pesquisa científica. A modelagem de tópicos tem possibilitado realizar tal atividade, por intermédio de algoritmos de *Machine Learning*, que utilizam métodos estatísticos, por meio de uma estrutura não supervisionada em documentos eletrônicos que constituem *corpora* de dados, buscando assim, analisar e descobrir temas e suas respectivas relações.

Partindo desse princípio, questiona-se: de que forma tem-se apresentado os assuntos científicos de maior relevância produzidos junto ao 'GT-8 Informação e Tecnologia' do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB)?

A pesquisa possui como objetivo geral identificar os tópicos de maior relevância, as palavras e pesos que constituem os tópicos do *corpus* de documentos do GT-8. Além disso, busca discutir as respectivas relações entre os termos encontrados no *corpus* de dados e apresentar o comportamento dos termos mais frequentes. Pressupõe-se que o mapeamento científico do GT-8 referente a Informações e Tecnologias contribui para novas perspectivas e tomadas de decisões estratégicas do grupo, como: possíveis frentes de pesquisa; novos temas de interesse; e lacunas a serem preenchidas.

Justifica-se a importância da pesquisa, uma vez que a realização do mapeamento científico do grupo pode apresentar novos resultados e prospectar diferentes cenários e perspectivas para a área estudada, tais como contribuições práticas, metodológicas e científicas da pesquisa.

2 REFERENCIAL TEÓRICO

O capítulo 2 apresenta três seções: 1 - Ciência da Informação sob duas perspectivas: perspectiva tecnológica, abordando características como processamento, armazenamento, recuperação, interpretação, transmissão e transformação; e perspectiva interdisciplinar; 2 - Comunicação Científica como meio de debates e disseminação da informação; e 3 – descrição do *corpus* estudado. Ambos os assuntos correlacionam com o referencial empírico dessa pesquisa apresentados no capítulo 3.

2.1 Ciência da Informação

A Ciência da Informação traz em sua origem o movimento acelerado das Tecnologias da Informação e Comunicação, que, por sua vez, buscam soluções tecnológicas por meio de suas abordagens, natureza, manifestações e efeitos da informação e do conhecimento de

forma que seja possível garantir o fluxo e o uso da comunicação. Destaca-se nesse processo o ciclo informacional que possui as características de produção, organização, armazenamento, representação, disseminação, recuperação, acesso e o uso da informação (NHACUONGUE; FERNEDA, 2015).

O conceito de Ciência da Informação, estabelecido por Borko a partir das ideias de Taylor, foi constituído durante duas reuniões realizadas em outubro de 1961 e abril de 1962, no *Georgia Institute of Technology* e na *National Science Foundation* – uma agência governamental dos Estados Unidos (BARRETO, 2007). É definido como:

> Ciência da Informação é a disciplina que investiga as propriedades e o comportamento informacional, as forças que governam os fluxos de informação visando à acessibilidade e a usabilidade ótima. A Ciência da Informação está preocupada com o corpo de conhecimentos relacionados à origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação, e utilização da informação. Isto inclui a pesquisa sobre a representação da informação em ambos os sistemas, naturais e artificiais, o uso de códigos para a transmissão eficiente da mensagem e o estudo do processamento de informações e de técnicas aplicadas aos computadores e seus sistemas de programação (BORKO, 1968, p.3, tradução nossa).

Mesmo sem ter tido contato com a microeletrônica e os avanços das telecomunicações após a invenção dos microcomputadores, Borko conceituou a Ciência da Informação prospectando campos como a análise linguística, demanda da informação, linguagens documentárias, análise de projeto de sistemas, tradução, sistemas especialistas, padrões de reconhecimento de imagens e voz e produção e reprodução de documentos (SOUZA, 2007).

Shera e Cleveland corroboram Borko e acrescentam que os processos de informação são constituídos por geração, disseminação, organização, armazenamento, recuperação e uso (SHERA; CLEVELAND, 1977). Os autores, Capurro e Hjorland (2007), destacam como processos da informação as etapas de geração, coleta, transformação, interpretação, armazenamento, recuperação, disseminação, transformação e uso para o domínio particular das tecnologias modernas da área. Trata-se de uma ciência que correlaciona as tecnologias da informação voltadas para o estudo científico do comportamento do homem pela busca e meios de processar a informação por meio de computadores (SOUZA, 2007). A Ciência da Informação, enquanto disciplina, procura estudar um corpo de conhecimento científico, tecnológico e de sistemas (CAPURRO; HJORLAND, 2007).

Para Borko (1968) a Ciência da Informação possui componentes da ciência dura que busca realizar investigações sem a necessidade da ciência aplicada e se caracteriza como ciência interdisciplinar por abranger áreas como artes gráficas, biblioteconomia, comunicações, gestão, linguística, lógica, matemática, psicologia, tecnologias computacionais. Le Coadic (1996) define Ciência da Informação como sendo um campo interdisciplinar que está relacionado às áreas da psicologia, filosofia, sociologia, linguística, matemática, lógica, informática, telecomunicações, economia, direito e política.

Sob um olhar mais tecnológico em relação às características da Ciência da Informação apresentado por Saracevic (1996), 28 anos após a conceituação de Borko (1968), destacam-se: 1) interdisciplinaridade. Relações entre as áreas da computação e da inteligência artificial de forma a produzir pesquisas teóricas e experimentos; 2) uso das tecnologias da informação para transformação da sociedade moderna em sociedade da informação considerando a quantidade e qualidade da informação e da comunicação; 3) evolução da sociedade da informação considerando o papel social e econômico da Ciência da Informação.

Saracevic (1996, p.47) destaca a Ciência da Informação como o campo de estudo voltado para prática científica e profissional que busca solucionar problemas por meio de tecnologias informacionais "[...] problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação". Le Coadic (1996) contribui ao relacionar os conceitos de informação e signo como meio de registro e transmissão da informação por meio do suporte tecnológico:

> Informação é um conhecimento inscrito (gravado) sob a forma escrita (impressa ou numérica), oral ou audiovisual. A informação comporta um elemento de sentido. É um significado transmitido a um ser consciente por meio de uma mensagem inscrita em um suporte espacial-temporal: impresso, sinal, elétrico, onda sonora, etc. Essa inscrição é feita graças a um sistema de signos (a linguagem), signo este que é um elemento da linguagem que associa um significante a um significado: signo alfabético, palavra, sinal, pontuação (LE COADIC, 1996, p. 5).

2.2 Comunicação científica

A pesquisa científica pode produzir uma série de publicações e ser disseminada em diversos formatos durante a sua construção e ou após o seu término. Exemplos da produção científica podem ser apresentados por meio de palestras, relatórios, congressos, artigos de periódicos, livros impressos e ou digitais. Durante esse processo, o autor deve expor sua

pesquisa ao julgamento realizado por pares, de forma que seja possível alcançar o consenso, confiabilidade e validação da pesquisa científica para que seja publicada (ZIMAN, 1979).

Disseminar a informação da pesquisa científica possibilita tornar público a produção do conhecimento (LARA; CONTI, 2003). A informação pode fluir por diferentes canais de comunicação e por diferentes formatos de acordo com o estágio da pesquisa. Nos canais informais destacam os relatórios de pesquisas, atas de reuniões, textos apresentados em seminários, ou mesmo, anais de eventos. Entretanto podem possuir difícil recuperação da informação, pois nem sempre são armazenados e disponibilizados para acesso. Já os canais formais possibilitam que as informações sejam armazenadas e coletadas (MUELLER, 2007).

As teleconferências e as listas de discussões realizadas por meio do suporte tecnológico têm potencializado o processo da comunicação científica, entretanto, os eventos presenciais acabam por apresentar possibilidades diferenciadas como o contato pessoal entre um maior número de pesquisadores que estão concentrados em um único lugar, podendo realizar, assim, a troca de informações com maior intensidade. Os eventos científicos acabam por agradar aos pesquisadores justamente por permitirem a exposição e a discussão de suas pesquisas de forma que possam ser avaliadas por outros pesquisadores (CAMPELLO, 2007).

Existem diferentes tipos de encontros científicos, dentre eles, colóquio, encontro, fórum, reunião, seminário e simpósio. Os documentos gerados por meio de encontros científicos são publicados no formato de anais que reúnem um conjunto de trabalhos apresentados, podendo ser palestras, conferências, resumos dos trabalhos ou pesquisas na íntegra. Além disso, podem ser publicados antes ou após a realização do evento (CAMPELLO, 2007).

2.3 ENANCIB e a produção científica do GT-8

O Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) é o evento nacional da Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (ANCIB). Sua primeira edição foi realizada no ano de 1994 na Universidade Federal de Minas Gerais (UFMG). O evento vem ocorrendo interruptamente a cada ano, desde 2005, congregando professores, pesquisadores e estudantes da Ciência da Informação e áreas afins, que buscam identificar o estado da arte, por meio de apresentações e discussões de pesquisas científicas organizadas por Grupos de Trabalhos (GT's) (ANCIB, 201?b).

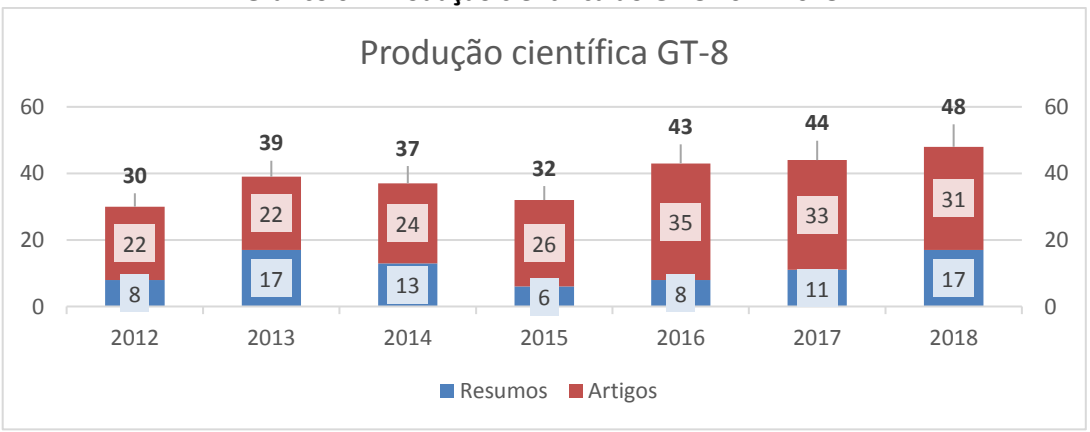
O XII ENANCIB, realizado no ano de 2011, na Universidade de Brasília (UnB) apresentou pela primeira vez 11 GT's (ENANCIB, 2011). Esse número de grupos vem se mantendo até a edição atual, com pequenas alterações de nomenclaturas. O ENANCIB apresenta os seguintes grupos: GT-1: Estudos Históricos e Epistemológicos da Ciência da Informação; GT-2: Organização e Representação do Conhecimento; GT-3: Mediação, Circulação e Apropriação da Informação; GT-4: Gestão da Informação e do Conhecimento nas Organizações; GT-5: Política e Economia da Informação; GT-6: Informação, Educação e Trabalho; GT-7: Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação; GT-8: Informação e Tecnologia; GT-9: Museu, Patrimônio e Informação; GT-10: Informação e Memória; e GT-11 - Informação & Saúde (ANCIB, 201?b).

O GT-8 iniciou suas atividades no ano de 2008, durante o IX ENANCIB, realizado na Universidade de São Paulo (USP). O 'GT-8: Informações e Tecnologia' possui como ementa (ANCIB, 201?a):

> Estudos e pesquisas teórico-práticos sobre e para o desenvolvimento de tecnologias de informação e comunicação que envolvam os processos de geração, representação, armazenamento, recuperação, disseminação, uso, gestão, segurança e preservação da informação em ambientes digitais (ANCIB, 201?a, *online*).

Entre o XIII ENANCIB, realizado no ano de 2012, na Fundação Oswaldo Cruz (Fiocruz), até o XIX ENANCIB, realizado no ano de 2018, na Universidade Estadual de Londrina (UEL), o GT-8 publicou 273 produções científicas nos anais dos eventos, sendo 80 resumos expandidos apresentados na modalidade de pôster e 193 artigos completos apresentados como comunicação oral. O gráfico 01 apresenta o quantitativo de publicações, separadas por tipo de produção científica e por cada ano em que o evento foi realizado - 2012 - 2018.

Gráfico 01: Produção científica do GT-8 2012-2018.



Fonte: Elaborado pelos autores.

Entre os tipos de produções científicas na modalidade de resumos e artigos publicados por meio dos anais do ENANCIB, destaca-se o ano de 2015, com o menor número de resumos expandidos publicados durante o intervalo analisado, sendo apenas 6 trabalhos. Já os anos de 2013 e 2018 tiveram o maior número de trabalhos aceitos nessa modalidade, sendo 17 trabalhos. Com relação aos artigos científicos, o ano de 2012 teve o menor quantitativo de publicações, sendo 22 trabalhos. O maior número de trabalhos aceitos nessa modalidade ocorreu no ano de 2017, sendo publicados 33 artigos científicos nos anais do evento.

3 REFERENCIAL EMPÍRICO

O capítulo 3 apresenta duas seções: 1 - conceitos de modelagem de tópicos, tais como o método probabilístico *Latent Dirichlet Allocation* (LDA); e 2 - conceitos de *corpus*, Processamento de Linguagem Natural e *Machine Learning* que compõe a parte prática da pesquisa.

3.1 Modelagem de tópicos

Com o crescente volume de informações disponibilizadas na internet, tem sido necessário cada vez mais o uso de ferramentas computacionais que sejam capazes de organizar, pesquisar e compreender essa quantidade expressiva de informações (HOFMANN, 1999b; BLEI, 2012). A modelagem de tópicos tem possibilitado, com algoritmos de *Machine Learning*, que utilizam métodos estatísticos, organizar e resumir, por meio de uma estrutura não supervisionada, arquivos eletrônicos que constituem grandes *corpora* de dados e informações, sendo que, em escalas elevadas, se tornaria humanamente impossível realizar análises manuais por meio de estruturas latentes que buscam analisar e descobrir temas, bem como suas respectivas relações e mudanças ao longo de um determinado período (BLEI, 2012; KASZUBOWSKI, 2016).

Os modelos de extração de tópicos probabilísticos partem da premissa que os documentos possuem um conjunto de tópicos misturados, sendo um tópico formado por uma distribuição probabilística de palavras. A sua formação é constituída por um modelo generativo probabilístico, que, quando aplicado em um determinado *corpus*, acaba por especificar um procedimento probabilístico pelo qual os documentos são gerados. Durante esse processo, uma nova distribuição de tópicos é estabelecida seguindo as etapas: 1 – cada palavra é inserida num documento; 2 – um tópico é escolhido aleatoriamente com base na

distribuição realizada anteriormente; e 3 – uma palavra tópica é selecionada. Torna-se possível inverter o processo utilizando técnicas estatísticas que inferem o conjunto de tópicos utilizados para gerar a coleção de documentos (STEYVERS; GRIFFITHS, 2007).

A construção de um documento utilizando a modelagem de extração de tópicos é realizada após a definição de um determinado número de assuntos ou tópicos que serão emergidos junto ao *corpus* de documentos. Esse quantitativo de tópicos é responsável por determinar os termos que serão utilizados no documento. Embora os termos ainda sejam desconhecidos, pode-se estimar o quantitativo de termos a partir dos documentos e dos termos por meio das variáveis observadas (SANTOS, 2015).

3.1.1 *Latent Dirichlet Allocation*

Os métodos probabilísticos utilizados para modelagem de tópicos são compostos por conceitos da computação e da estatística, sendo o *Latent Dirichlet Allocation* (LDA) um dos modelos generativos mais utilizados para organizar grandes coleções de documentos. Esse modelo utiliza uma abordagem *bayesiana* e parte do princípio de que os documentos contidos em um determinado *corpus* sejam representados como misturas aleatórias de tópicos latentes, sendo cada tópico caracterizado por uma distribuição de palavras que compreendem a cada um dos documentos (BLEI, 2012).

O modelo LDA utilizou do ferramental da álgebra linear, especificamente a decomposição *Singular Value Decomposition* (SVD) para realizar a decomposição de *corpus* nos seus temas constituintes e reduzir efeitos adversos como sinonímia e polissemia por meio da identificação de associações estatísticas entre os termos, desenvolvendo-se assim por meio de uma matriz que realiza a contagem de frequência dos termos contidos nos documentos de todo o *corpus* ou de fragmentos dos documentos (HOFMANN, 1999a; CHANG *et al*., 2009; AGGARWAL; ZHAI, 2012).

Quando aplicado o modelo LDA em um *corpus* de documentos, os tópicos são interpretáveis como temas na coleção e as representações do documento remetem aos seus temas. Nesse processo, são destacados três pontos: 1 - as variáveis aleatórias ocultas codificam a estrutura temática; 2 - os tópicos aprendidos resumem a coleção e as representações dos documentos; e 3 - os *corpora* em grupos sobrepostos são organizados pela representação do documento (CHANEY; BLEI, 2012). Cada documento contido em um *corpus* possui sua destruição própria de tópicos. Dessa forma, cada documento pode conter vários

tópicos e cada um deles contém a sua proporção de relevância. Tal distribuição de tópicos para cada documento está relacionada à distribuição multivariada de *Direchlet* (SANTOS, 2015).

3.2 Conceitos de *corpus*, Processamento de Linguagem Natural e *Machine Leraning*

Corpus é considerado um artefato produzido para fins específicos de pesquisa e é constituído por uma coleção de textos produzidos pelo homem em um ambiente natural de comunicação. A coleção não é criada com propósito de compor um *corpus*, entretanto, seu conteúdo é definido por fenômenos linguísticos que representam uma variedade de linguagem e devem ser legíveis por computadores (SARDINHA, 2000; PUSTEJOVSKY; STUBBS, 2012).

Processamento de Linguagem Natural (PLN) é um campo da engenharia e da ciência da computação que se desenvolveu a partir do estudo linguístico computacional na área da inteligência artificial e o seu objetivo está em estreitar a comunicação humana com máquinas por meio da linguagem natural (PUSTEJOVSKY; STUBBS, 2012). O PLN trata computacionalmente diversos aspectos da comunicação humana de forma que seja possível realizar a extração de significados de conteúdo, esses, podendo ser sons, palavras ou sentenças (GONZALEZ; LIMA, 2003; MARQUESONE, 2016).

Machine Learning é um tópico de investigação da área da inteligência artificial que possui como objetivo realizar a aprendizagem de sistemas computacionais a partir de algoritmos que aprendem interativamente, por meio de processo repetitivo a partir dos dados fornecidos. Esses algoritmos possuem como intenção descobrir *insights* ocultos nos dados de forma que se possam encontrar informações específicas (VASCONCELOS; BARÃO, 2017; MACHADO, 2018). Os algoritmos não supervisionados são utilizados em dados que não possuem rótulos históricos ou resultados conhecidos previamente (AYODELE, 2010; PUSTEJOVSKY; STUBBS, 2012).

4 PROCEDIMENTOS METODOLÓGICOS

A fase empírica da pesquisa que diz respeito à modelagem de tópicos foi realizada num *corpus* de dados constituídos de 273 documentos científicos do GT-8 do ENANCIB, publicados entre os anos de 2012 e 2018.

Para esta fase, foram utilizadas as seguintes etapas adaptadas de McKinley (2018): 1 - interação com o mundo externo: coleta e constituição do *corpus* de dados formado por documentos textuais, resumos expandidos e artigos completos, do XIII ao XIX ENANCIB; 2 - pré-processamento e preparação: limpeza, manipulação, combinação, normalização, tratamento e transformação dos dados para realização da análise descritiva; 3 - transformação: operações matemáticas e estatísticas aplicadas em grupos de conjuntos de dados a fim de obter novos conjuntos de dados; 4 - modelagem e processamento: conectar os dados já tratados a modelos estatísticos e algoritmo de *Machine Learning*, sendo utilizado o modelo *Latent Dirichlet Allocation* (LDA); 5 - apresentação: visualizações gráficas ou sínteses textuais; 6 - documentação: análise e discussão dos resultados. Foram utilizados para esse procedimento o *framework* Jupyter Notebook, a linguagem de programação Python e bibliotecas as Pdfminer, Gensim, NLTK, Numpy, Matplotlib, Plotly e pyLDAvis.

A pesquisa é classificada, de acordo com Gil (2010), em: quanto a finalidade/natureza - como aplicada, buscando solucionar um problema específico e sugerir novas questões a serem investigadas; quanto a abordagem do problema - como quali-quantitativo, perpassando pela interpretação dos fenômenos e a atribuição de significados por meio com dados quantificáveis; quanto aos objetivos - como exploratório, possibilitando utilizar um conjunto de procedimentos técnicos buscando identificar melhor o fato ou fenômeno no âmbito da pesquisa.

5 RESULTADOS E DISCUSSÕES

Durante a análise de pré-processamento foi possível identificar os termos mais frequentes do *corpus* de dados por meio dos n-gramas[1], tais como os unigramas com 940.925, bigramas com 940.652 e tigramas com 940.379 termos. O quadro 01 apresenta a lista dos 35 termos mais frequentes de cada categoria.

[1] Um n-grama consiste em um pedaço de n-caracteres, extraído de uma cadeia de caracteres maior sendo uma palavra ou frase. Usualmente, n assume o valor 1, 2 ou 3 respectivamente para unigrama, bigrama e trigrama (Sukkarieh et al 2003).

Quadro 01: Termos e frequência do *corpus* de dados.

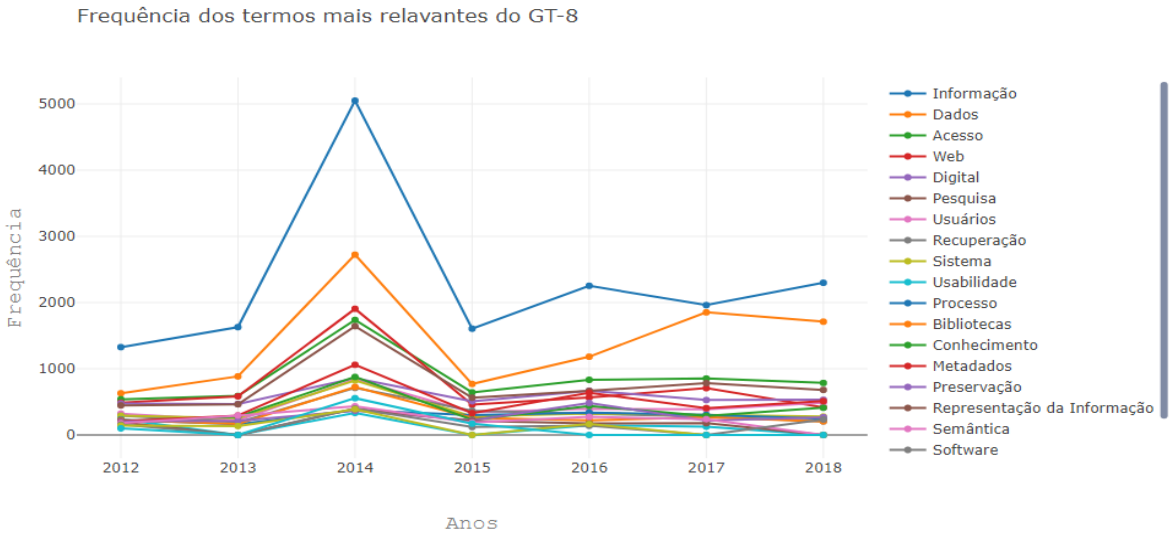
Nº/ GRAMA	UNIGRAMAS	BIGRAMAS	TRIGRAMAS
1	informação,16061	recuperação_informação,1359	resource_description_framework,493
2	dados,9927	arquitetura_informação,1264	ambientes_informacionais_digitais,359
3	pesquisa,5190	informação_tecnologia,1039	international_organization_standardization,356
4	web,5114	web_semântica,923	machine_readable_cataloging,349
5	digital,4025	big_data,690	extensible_markup_language,325
6	informações,3887	preservação_digital,686	tecnologias_informação_comunicação,322
7	uso,3575	redes_sociais,628	arquitetura_informação_pervasiva,218
8	metadados,3447	ambientes_informacionais,589	fonte_elaborado_autores,197
9	usuários,3416	linked_data,577	livros_didáticos_digitais,192
10	forma,3410	resource_description,525	sistemas_recuperação_informação,158
11	digitais,3091	description_framework,495	tese_doutorado_informação,148
12	data,2982	information_science,488	linked_open_data,148
13	information,2911	dados_pesquisa,481	world_wide_web,146
14	sistemas,2760	bases_dados,458	gestão_arquivística_documentos,146
15	conhecimento,2727	informação_comunicação,454	sistema_recuperação_informação,139
16	sistema,2586	sistemas_informação,454	conselho_nacional_arquivos,126
17	busca,2539	encontrabilidade_informação,406	uniform_resource_identifier,121
18	usuário,2535	base_dados,403	dissertação_mestrado_informação,119
19	processo,2503	tecnologias_informação,392	nacional_pesquisa_informação,116
20	documentos,2409	informacionais_digitais,388	fonte_dados_pesquisa,116
21	modelo,2404	repositórios_digitais,385	encontro_nacional_pesquisa,115
22	recuperação,2365	dublin_core,378	modalidade_apresentação_comunicação,113
23	partir,2309	padrões_metadados,371	apresentação_comunicação_oral,113
24	meio,2291	international_organization,361	informação_faculdade_filosofia,104
25	recursos,2277	markup_language,358	souza_santos_silva,99
26	tecnologia,2198	organization_standardization,356	comunicação_oral_resumo,97
27	conteúdo,2119	repositórios_institucionais,356	informação_comunicação_tic,82
28	podem,2068	machine_readable,355	web_ontology_language,81
29	cada,2067	cada_vez,352	doutorado_informação_faculdade,80
30	bibliotecas,2061	readable_cataloging,349	ciclo_vida_dados,80
31	preservação,1975	dados_abertos,332	gt-8_informação_tecnologia,75
32	comunicação,1895	ciclo_vida,332	society_information_science,73
33	desenvolvimento,1840	recursos_informacionais,331	american_society_information,71
34	figura,1829	extensible_markup,328	requisitos_funcionais_registros,70
35	contexto,1774	objetos_digitais,324	informação_belo_horizonte,70

Fonte: Elaborado pelo autor.

Torna-se possível perceber que os termos mais frequentes não necessariamente são os mais importantes ou estão relacionados aos termos utilizados na área estudada. Alguns desses termos podem ser tratados junto as *stopwords* – palavras de parada – para que não sejam exibidas, como por exemplo: 'partir', 'podem', 'cada', 'figura', 'trabalho', 'cada vez', 'nesse sentido', 'fonte_elaborado', 'apresentação comunicação oral'. Além disso, é possível notar um distanciamento entre a frequência dos unigramas, bigramas e trigramas. Enquanto o termo 'informação' possui 16.061 repetições no *corpus*, o bigrama mais frequente 'recuperação_informação' ocupa a posição de número 59 em uma ordem de frequência de termos de todo o *corpus*, obtendo assim 1.359 repetições ou equivalente a 91,53%. Esse valor é menor que o unigrama mais encontrado no *corpus*. Embora os números dos termos bigramas e trigramas sejam menores quando comparados aos unigramas, pode-se perceber que tais termos estão amplamente conectados a ementa do GT-8.

Entre os 150 termos de maior ocorrência do *corpus*, foram selecionados 20 termos de relevância de forma que fosse possível identificar seus respectivos comportamentos entre os anos de 2012 a 2018, conforme apresentado no gráfico 02[2]. Os termos são: 'informação', 'dados', 'acesso', 'web', 'digital', 'pesquisa', 'usuários', 'recuperação', 'sistema', 'usabilidade', 'processo', 'bibliotecas', 'conhecimento', 'metadados', 'preservação', 'recuperação da informação', 'semântica', 'software' e 'ontologia'.

Gráfico 02: Distribuição de termos frequentes do GT-8 2012-2018.



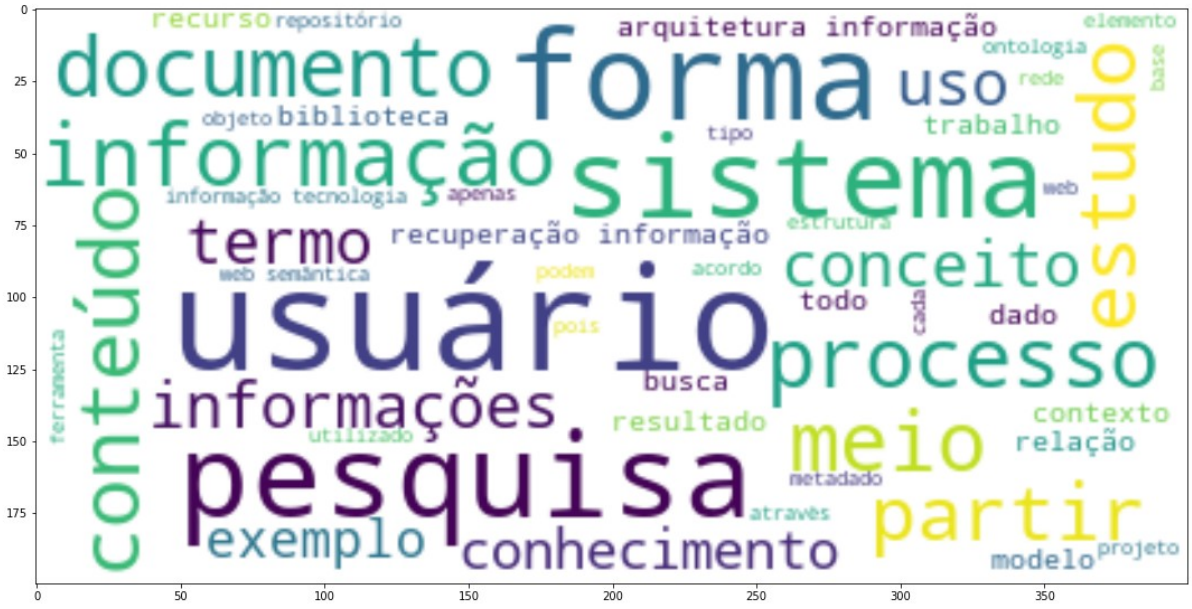
Fonte: Elaborado pelos autores.

[2] Acesse ao gráfico dinâmico para realizar outras interações através do link: https://github.com/marcosdesouza82/topic-model-enancib/blob/master/grafico_frequencia_gt8_enancib.ipynb

Destaca-se nesses resultados o termo ‘informação’ com 1327, 1630, 5052, 1606, 2254, 1962 e 2300 repetições nos anos de 2012, 2013, 2014, 2015, 2016, 2017 e 2018, respectivamente. O termo ‘dados’ apresentou a frequência de 630, 885, 2724, 771, 1183, 1853, 1712. O termo ‘conhecimento’ apresentou frequência de 210, 275, 877, 231, 439, 287, 412. Os resultados de ambas as frequências mantêm um padrão de crescimento e pouca variação, para mais ou para menos, entre os anos, exceto para o pico elevado de valores ocorrentes ano de 2014. O mesmo ocorre para os termos ‘web’ com frequência de 487, 584, 1907, 459, 566, 708, 417, ‘digital’ 477, 469, 858, 510, 669, 528, 532 e ‘pesquisa’ 447, 464, 1644, 563, 667, 784, 681. Entretanto, os termos ‘representação da informação’ 191, 0, 398, 212, 176, 178, 0, ‘semântica’ 180, 294, 433, 188, 274, 239, 0 e ‘*software*’ 157, 0, 400, 120, 140, 0, 234 possuem valores iguais a 0. Nesse caso, a frequência dos termos no respectivo ano está abaixo dos 150 termos listados. Outro exemplo está no termo ‘acessibilidade’, onde os valores somente aparecem em 3 anos - 100, 0, 556, 171, 0, 0, 0.

A visualização de frequência de todo o *corpus*, sem distinção de documentos ou separação por ano, apresentada na figura 01, em forma de nuvem de palavras, que contém os 50 termos mais frequentes. Faz-se necessário ressaltar que o primeiro bigrama ocupa a posição número 59 dos termos mais frequentes. O primeiro trigrama está alocado na posição 303. Dessa forma, bigramas e trigramas não são exibidos.

Figura 01: Nuvem de palavras dos termos frequentes do GT-8 2012-2018.



Fonte: Elaborado pelos autores.

O algoritmo de *Machine Learning Latent Dirichlet Allocation* (LDA) foi configurado para realizar a identificação de 6, 8, 10, 12, 14 e 16 tópicos junto ao *corpus* de dados. Cada tópico de possui o idx no valor igual 10, o que significa que cada tópico possui 10 palavras que melhor lhe representam juntamente com o peso da contribuição da palavra para esse tópico. Além disso, foram configurados chunksize = 1000, que se refere ao número de documentos a serem usados em cada bloco de treinamento, passes = 40, referente ao número de passagens de treinamento pelos documentos e iterations = 600, referente ao número máximo de iterações no *corpus* ao inferir a distribuição de tópico de um *corpus*.

Quadro 02: Tópicos, palavras e pesos do modelo LDA aplicado ao corpus de dados.

TÓPICO	PALAVRAS E PESOS
1	0.005*"informação" + 0.001*"information" + 0.001*"forma" + 0.001*"informações" + 0.001*"design" + 0.001*"processo" + 0.001*"digital" + 0.001*"indexação" + 0.001*"redes" + 0.001*"uso"
2	0.005*"informação" + 0.005*"dados" + 0.002*"pesquisa" + 0.002*"web" + 0.001*"data" + 0.001*"metadados" + 0.001*"informações" + 0.001*"uso" + 0.001*"digital" + 0.001*"digitais"
3	0.002*"web" + 0.002*"informação" + 0.002*"usuários" + 0.001*"dados" + 0.001*"modelo" + 0.001*"pesquisa" + 0.001*"busca" + 0.001*"uso" + 0.001*"forma" + 0.001*"usuário"
4	0.003*"informação" + 0.002*"digital" + 0.001*"dados" + 0.001*"metadados" + 0.001*"preservação" + 0.001*"pesquisa" + 0.001*"repositórios" + 0.001*"conhecimento" + 0.001*"uso" + 0.001*"processo"
5	0.004*"informação" + 0.003*"dados" + 0.002*"web" + 0.002*"digital" + 0.001*"informações" + 0.001*"uso" + 0.001*"pesquisa" + 0.001*"conhecimento" + 0.001*"data" + 0.001*"forma"
6	0.007*"informação" + 0.001*"pesquisa" + 0.001*"recuperação" + 0.001*"busca" + 0.001*"sistema" + 0.001*"information" + 0.001*"usuário" + 0.001*"arquitetura" + 0.001*"informações" + 0.001*"arquitetura_informação"
7	0.004*"informação" + 0.001*"dados" + 0.001*"digital" + 0.001*"pesquisa" + 0.001*"usuário" + 0.001*"registros" + 0.001*"forma" + 0.001*"metadados" + 0.001*"obra" + 0.001*"informações"
8	0.003*"informação" + 0.001*"pesquisa" + 0.001*"dados" + 0.001*"inovação" + 0.001*"conhecimento" + 0.001*"informações" + 0.001*"tecnologia" + 0.001*"digitais" + 0.001*"uso" + 0.001*"digital"
9	0.001*"informação" + 0.000*"descrição" + 0.000*"usuários" + 0.000*"leitura" + 0.000*"description" + 0.000*"organização" + 0.000*"sistema" + 0.000*"usuário" + 0.000*"type" + 0.000*"e-books"
10	0.002*"dados" + 0.002*"metadados" + 0.001*"preservação" + 0.001*"digital" + 0.001*"bibframe" + 0.001*"modelo" + 0.001*"informação" + 0.001*"documentos" + 0.001*"padrões" + 0.001*"usuários"

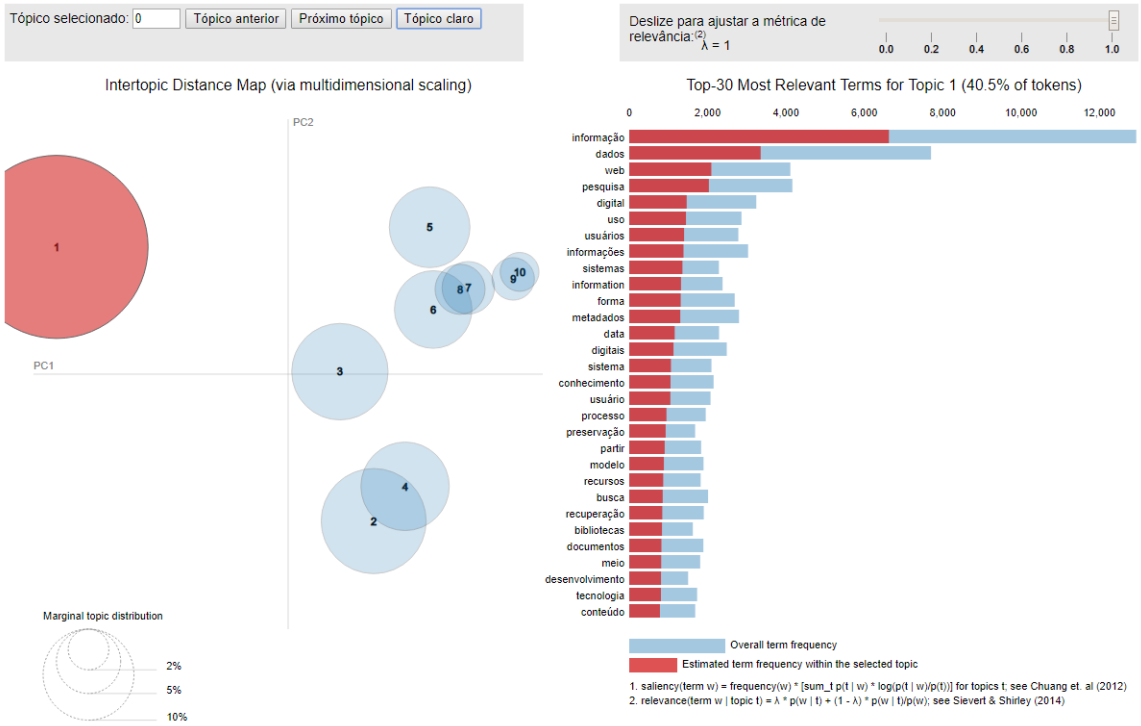
Fonte: Elaborado pelos autores.

O modelo com 6, 8, 10 e 12 tópicos apresentaram resultados relevantes e coesos com o *corpus* estudado. Já os modelos com 14 e 16 tópicos apresentaram ruídos em seus resultados, como, por exemplo, o tópico 0 do modelo 16, que apresenta palavras fora do contexto do GT-8, como 'abolição' e 'escravidão' com pesos igual a zero e sujeiras como xviii e sessões: 0.001*"jogos" + 0.001*"jogo" + 0.000*"abolição" + 0.000*"escravidão" + 0.000*"escravidão_abolição" + 0.000*"educativo" + 0.000*"sessões" + 0.000*"jogadores" +

0.000*"xviii" + 0.000*"monitoramento". O quadro 02 apresenta os resultados alcançados da modelagem de tópicos utilizando o modelo LDA com 10 tópicos e suas respectivas palavras e pesos. O tempo de processamento para realizar a modelagem de tópicos de 273 documentos foi de 32 minutos e 16 segundos em uma máquina com CPU Intel Core i7 - 2630QM 2.00GHz e memória de 8GB.

O gráfico 03 apresenta a distribuição dos tópicos por meio de um mapa de distância intertópico[3], com as respectivas relações e os 30 termos mais relevantes de cada tópico. O tópico 1 é o mais representativo dentre todos os tópicos listados, apresentando assim, 40,5% dos *tokens*. O tópico 2 apresenta 13,3% dos *tokens* e possui relações entre termos com o tópico 4 que apresenta 9,4% dos *tokens*. O tópico 3 possui 11,2% dos *tokens*. Os tópicos 6 com 7,3% dos *tokens*, 7 com 3,3% dos *tokens* e 8 com 3% dos *tokens* também possuem relações entre os termos, entretanto, os termos 7 e 8 apresentam individualmente menos da metade de relevância do tópico 6. O tópico 5 apresenta 7,9% dos *tokens*. Já os tópicos 9 e 10 apresentam respectivamente 2,2% e 1,8% dos *tokens*, apresentando assim, uma baixa representatividade, comparados aos tópicos de 1 a 6.

Gráfico 03: Distribuição dos tópicos utilizando pyLDAvis.



Fonte: Elaborado pelos autores.

[3] Acesse ao gráfico dinâmico para visualizar outras interações através do link: https://github.com/marcosdesouza82/topic-model-enancib/blob/master/topic_model_gt8_enancib.ipynb

6 CONSIDERAÇÕES FINAIS

A Ciência da Informação, enquanto campo interdisciplinar voltada para as práticas científicas e profissionais, tem produzido diversas soluções que envolvem processos de geração, representação, armazenamento, recuperação, disseminação, uso, gestão, segurança e preservação da informação. Esses processos podem ser oficializados por meio da comunicação científica, que, por sua vez, tem se tornado crescente a cada ano, conforme apresentado no quantitativo anual de produção científica do GT-8.

Através da modelagem de tópicos, foi possível identificar, junto ao *corpus* de documentos analisados, os principais tópicos e suas respectivas palavras/pesos que os representam. O modelo de *Machine Learning* LDA utilizado no *corpus* de dados apresentou um melhor resultado quando configurado para realizar a aprendizagem a partir de 10 tópicos. Os modelos com 14 e 16 tópicos apresentaram ruídos não condizentes com a ementa do GT-8. Os modelos com 6 e 8 são insuficientes para cobrir toda a temática do grupo.

O tópico 1 foi o de maior relevância de todo o *corpus*, apresentando, assim, 40,5% dos *tokens* e tendo associado alguns dos termos como ‘informação’, ‘dados’, ‘web’, ‘pesquisa’, ‘digital’, ‘uso’, ‘usuários’, ‘metadados’, ‘sistema’, ‘conhecimento’, ‘recuperação’, ‘documentos’ e ‘tecnologia’. Todos os termos são condizentes com a proposta do grupo apresentada no seu ementário. As relações entre os tópicos ocorrem mediante palavras que constituem mais de um tópico. Dessa forma, foi observada tal relação entre os tópicos 2 e 4, 6 a 8 e 9 e 10.

A modelagem de tópicos não determina o nome exato de cada tópico. Assim, faz-se necessário realizar uma análise por um especialista da linguagem de domínio que fará a suposição do possível nome do tópico, com base nas palavras e pesos que os compõe. Dessa forma, sugere-se como pesquisas futuras que essa análise seja realizada por membros da comissão científica do GT-8 do ENANCIB. Outra perspectiva está em aplicar a mesma metodologia com outros modelos de *Machine Learning* como o *Latent Semantic Analysis* (LSA) ou *Probabilistic Latent Semantic Analysis* (pLSA) para uma comparação entre os resultados.

Além disso, o termo ‘informação’ foi o mais recorrente de todo o *corpus*, obtendo média anual de 2304 repetições. Já termo ‘dados’ foi o segundo mais recorrente e teve sua média anual de 1394 repetições. Termos constituídos por bigramas e trigramas possuem menos frequência, entretanto, muitos dos termos fazem parte da temática do GT-8. Com esse mapeamento científico, pode-se explorar através de temáticas específicas para futuros

ENANCIB's, os temas de pouca recorrência, entretanto, de muita relevância para o GT-8 e para a Ciência da Informação, tais como: a frequência na posição 64 'arquitetura_informação' com 1264 ocorrências; posição 154 'interoperabilidade' com 791 ocorrências; posição 187 'big_data' com 690 ocorrências; posição 190 'preservação_digital' com 686 ocorrências; posição 212 'redes_sociais' com 628 ocorrências; posição 293 'ambientes_informacionais' com 589 ocorrências; posição 567 'tecnologias_informação_comunicação' com 322 ocorrências; posição 1211 'sistemas_recuperação_informação' com 158 ocorrências; e a posição 1278 'curadoria_digital' com 151 ocorrências. Esse mapeamento e exploração das temáticas pode resultar em contribuições práticas, metodológicas e científicas da pesquisa.

REFERÊNCIAS

AGGARWAL, Charu C.; ZHAI, ChengXiang (Eds.). **Mining text data**. New York: Springer-Verlag, 2012.

ASSOCIAÇÃO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO (ANCIB). **GT-8: Informação e Tecnologia.** 201?a. Disponível em: <http://gtancib.fci.unb.br/index.php/gt-08>. Acesso em: 03 mar. 2019.

ASSOCIAÇÃO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO (ANCIB). **Portal de eventos da ANCIB.** 201?b. Disponível em: <http://enancib.ibict.br/index.php/enancib/index/schedConfs/archive>. Acesso em: 03 mar. 2019.

AYODELE, Taiwo Oladipupo. Types of machine learning algorithms. *In*: ZHANG, Yagang (Ed.). **New advances in machine learning**. Rijeka: IntechOpen, 2010. p. 19–48. Disponível em: <http://cdn.intechopen.com/pdfs-wm/10694.pdf>. Acesso em: 18 jun. 2019.

BARRETO, Aldo de Albuquerque. Uma história da Ciência da Informação. *In*: TOUTAIN, LÍDIA MARIA BATISTA BRANDÃO (Org.). **Para entender a Ciência da Informação**. Salvador: EDUFBA, 2007. p. 13–34. Disponível em: <http://pt.scribd.com/doc/32536278/Para-entender- a-Ciencia-da-Informacao>. Acesso em: 27 fev. 2019.

BLEI, David M. Probabilistic topic models. **Communications of the ACM**, v. 55, n. 4, p. 77–84, 2012. Disponível em: <http://dl.acm.org/citation.cfm?doid=2133806.2133826>. Acesso em: 27 fev. 2019.

BORKO, Harold. Information science: whats is it? **American Documentation**, v. 19, n. 1, p. 3-5, 1968.

CAMPELLO, Bernadete Santos. Encontros científicos. *In*: CAMPELLO, BERNADETE SANTOS; CENDÓN, BEATRIZ VALADARES; KRENMER, JANNETTE MARGUERITE (Orgs.). **Profissionais,**

Fontes de informação para pesquisadores e profissionais. Belo Horizonte: Editora UFMG, 2007. p. 55–71.

CAPURRO, Rafael; HJORLAND, Birger. O conceito de informação. **Perspectivas em Ciência da Informação**, v. 12, n. 1, 2007. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/54/47>. Acesso em: 26 fev. 2019.

CHANEY, Allison June-Barlow; BLEI, David M. Visualizing Topic Models. *In:* INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 6., 2012*,* Dublin. **Proceedings…** Palo Alto: AAAI Press, 2012. Disponível em: <https://pdfs.semanticscholar.org/59d7/d8415dacd300eb4d98b0da3cb32d27503b36.pdf>. Acesso em: 9 mar. 2019.

CHANG, Jonathan *et al*. Reading tea leaves: how humans interpret topic models. *In*: ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 23., 2009, Vancouver. **Advances in Neural Information Processing Systems 22**. Vancouver: NIPS, 2009. p. 288–296. Disponível em: <http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf>. Acesso em: 9 mar. 2019.

ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 12., 2011, Brasília. **XXI ENANCIB**: políticas de informação para a sociedade. Brasília: ANCIB**,** 2011.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010.

GONZALEZ, Marco; LIMA, Vera Lúcia Strube. Recuperação de informação e processamento da linguagem natural. *In*: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas. **Anais da III Jornada de Mini-Cursos de Inteligência Artificial**. Campinas: SBC, 2003. v. 3, p. 347–395.

HOFMANN, Thomas. Probabilistic latent semantic analysis. *In:* CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 15., 1999, Stockholm. **Proceedings…** San Francisco: Morgan Kaufmann Publishers, 1999a. p. 289–296. Disponível em: <http://www.iro.umontreal.ca/~nie/IFT6255/Hofmann-UAI99.pdf>. Acesso em: 6 mar. 2019.

HOFMANN, Thomas. Probabilistic latent semantic indexing. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 22., 1999, Berkeley. **Proceedings...** New York: ACM, 1999b. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.3584&rep=rep1&type=pdf>. Acesso em: 1 mar. 2019.

KASZUBOWSKI, Erikson. **Modelo de tópicos para associações livres**. 2016. 213f. Tese (Doutorado em Psicologia) - Universidade Federal de Santa Catarina (UFSC), Florianópolis, 2016. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/172577/343427.pdf?sequence=1>. Acesso em: 1 mar. 2019.

LARA, Marilda Lopes Gines de; CONTI, Vivaldo Luiz. Disseminação da informação e usuários. **São Paulo em Perspectiva**, v. 17, p. 26–34, 2003.

LE COADIC, Yves-François. **A ciência da informação**. Brasília: Briquet de lemos Livros, 1996.

MACHADO, Felipe Nery Rodrigues. **Big Data:** o futuro dos dados e aplicações. São Paulo: Editora Érica, 2018.

MARQUESONE, Rosangela. **Big Data:** técnicas e tecnologias para extração de valor dos dados. São Paulo: Casa do Código, 2016.

MCKINNEY, Wes. **Python para análise de dados:** tratamento de dados com pandas, numpy e ipython. São Paulo: Novatec, 2018.

MUELLER, Suzana Pinheiro Machado. A ciência, o sistema de comunicação e a literatura científica. *In*: CAMPELLO, BERNADETE SANTOS; CENDÓN, BEATRIZ VALADARES; KRENMER,

JANNETTE MARGUERITE (Orgs.). **Fontes de informação para pesquisadores e profissionais**. Belo Horizonte: Editora UFMG, 2007. p. 21–34.

NHACUONGUE, Januário Albino; FERNEDA, Edberto. O campo da ciência da informação: contribuições, desafios e perspectivas. **Perspectivas em Ciência da Informação**, v. 20, n. 2, p. 3–18, 2015. Disponível em: <http://dx.doi.org/10.1590/1981-5344/1932>. Acesso em: 7 mar. 2019.

PUSTEJOVSKY, James; STUBBS, Amber. **Natural language annotation for machine learning:** a guide to corpus-building for applications. Beijing: O'Reilly Media, 2012.

SANTOS, Fabiano Fernandes dos. **Extração de tópicos baseado em agrupamento de regras de associação**. 2015. 129f. Tese (Doutorado em Ciência da Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos, 2015. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-02122015-161054/>. Acesso em: 28 fev. 2019.

SARDINHA, Tony Berber. Lingüística de corpus: histórico e problemática. **DELTA: Documentação e Estudos em Linguística Teórica e Aplicada**, v. 16, n. 2, p. 323–367, 2000. Disponível em: <http://www.scielo.br/pdf/delta/v16n2/a05v16n2.pdf>. Acesso em: 13 mar. 2019.

SHERA, Jesse Hauk; CLEVELAND, Donald B. History and foundations of Information Science. **Annual Review of Information Science and Technology**, v. 12, p. 249–275, 1977.

SOUZA, Maria da Paixão Neres. Abordagem inter e transdiciplinar em ciência da informação. *In*: TOUTAIN, LÍDIA MARIA BRANDÃO (Org.). **Para entender a ciência da informação**. Salvador: **EDUFBA,** 2007. p. 75-90.

STEYVERS, Mark; GRIFFITHS, Tom. Probabilistic topic models. *In*: LANDAUER, Thomas K et al (Eds.). **Handbook of latent semantic analysis**. Mahwah: Lawrence Erlbaum Associates, 2007. p. 427-448.

SUKKARIEH, Jana Z.; PULMAN, Stephen G.; RAIKES, Nicholas. Auto-marking: using computational linguistics to score short, free text responses. *In*: ANNUAL CONFERENCE OF THE INTERNATIONAL ASSOCIATION FOR EDUCATIONAL ASSESSMENT, 29., 2003, Manchester. **Proceedings…** Manchester: IAEA, 2003.

VASCONCELOS, José Braga de; BARÃO, Alexandre. **Ciência dos dados nas organizações:** aplicações em python. Lisboa: FCA, 2017.

VIEIRA, Renata; LIMA, Vera Lúcia Strube. Lingüística computacional: princípios e aplicações. *In*: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 21., 2001, São Leopoldo. **Anais...** São Leopoldo: SBC, 2001. p. 47-86.

ZIMAN, John. **Conhecimento público**. Belo Horizonte: Itatiaia, 1979.