em questão
E-ISSN 1808-5245

Mapeamento de conhecimento científico: modelagem de tópicos das teses e dissertações do Programa de Pós-Graduação em Ciência da Informação da UFMG

Marcos de Souza

Doutor; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil;
marcosdesouza82@gmail.com

Renato Rocha Souza

Doutor; Universidade Federal de Minas Gerais / Fundação Getúlio Vargas, MG / RJ, Brasil;
rsouzaufmg@gmail.com

Resumo: O uso das ferramentas computacionais tem sido cada vez mais exigido para organizar, recuperar e compreender o crescente volume de dados. A comunicação científica tem contribuído, por meio de trabalhos formais e informais, para esse fenômeno; entretanto, a organização de uma grande coleção de documentos pode se tornar um processo lento e questionável quando realizado sem recursos tecnológicos. A modelagem de tópicos, por meio de algoritmos de aprendizagem de máquina, tem possibilitado organizar e resumir *corpora* de dados. A problemática da pesquisa é descobrir como se têm apresentado os temas das teses e dissertações produzidas pelo Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais. Busca-se identificar os tópicos de maior relevância do *corpus* de dados, constituído por documentos do tipo teses e dissertações desse programa de pós-graduação, assim como os termos de cada tópico e os pesos atribuídos a cada um desses termos. Na modelagem de tópicos, utilizou-se o modelo de alocação de Dirichlet latente, configurado para identificar 6, 8, 10, 12, 14, 16, 18 e 20 tópicos junto ao *corpus* de dados, o que permitiu realizar o mapeamento científico dos documentos analisados. Os resultados com 14 tópicos foram mais coesos e apresentaram menos ruídos e, por isso, permitiram inferir os nomes dos tópicos de maneira mais segura e estabelecer correlações com as linhas de pesquisa do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais.

Palavras-chave*:* Modelagem de tópicos. Aprendizagem de Máquina. Alocação de *Dirichlet Latente*. Mapeamento Científico. Ciência da Informação.

1 Introdução

O uso das ferramentas computacionais tem sido cada vez mais exigido para organizar, recuperar e compreender o crescente volume de dados que são

disponibilizados cotidianamente em diversos formatos no ciberespaço. A comunidade científica – pesquisadores, instituições, professores e estudantes – tem contribuído para esse fenômeno de volume de dados ao produzir e disseminar informações científicas. Esse processo envolve os canais formais e informais da comunicação científica.

Entre os canais formais de comunicação científica, tem-se teses e dissertações publicadas nas bases de dados das bibliotecas digitais de teses e dissertações das Instituições de Ensino Superior (IES); resumos simples, resumos expandidos e artigos completos publicados em anais de eventos ou periódicos científicos; e-books disponibilizados em diferentes plataformas da web. Já entre os canais informais de comunicação científica, relatórios de pesquisas; atas de reuniões; lista de discussões.

Ao longo de sua história, o Programa de Pós-graduação em Ciência da Informação (PPGCI) da Universidade Federal de Minas Gerais (UFMG), na modalidade *stricto sensu*, com os cursos de mestrado, desde 1976, e de doutorado, desde 1997, tem produzido, através de pesquisas, documentos de comunicação científica (PPGCI, 201?a). Através da Lei nº 12.527/2011 (BRASIL, 2011), dita Lei de Acesso à Informação (LAI), as IESs passaram a armazenar e disponibilizar teses e dissertações dos programas de pós-graduação por meio das bibliotecas digitais de teses e dissertações.

Organizar e resumir uma coleção de documentos com um grande volume de informações científicas, como a que tem sido produzida ao longo dos anos nos mais diferentes formatos pelo PPGCI, pode-se tornar uma tarefa exaustiva, duradoura para a pesquisa científica, mesmo quando se trata de uma amostragem reduzida. A modelagem de tópicos tem possibilitado, por meio de algoritmos de *machine learning*, que utilizam métodos estatísticos, realizar essas atividades através de uma estrutura não supervisionada em documentos eletrônicos que constituem os *corpora* de dados (BLEI, 2012). Dessa forma, torna-se possível analisar e descobrir temas e suas respectivas relações contidas em coleções de documentos. A partir desse princípio, questiona-se: de que forma se têm apresentado os temas, resultados da comunicação científica na

modalidade de teses e dissertações do PPGCI da Universidade Federal de Minas Gerais?

O objetivo geral da pesquisa é identificar os tópicos de maior relevância do *corpus* constituído por documentos do tipo teses e dissertações do PPGCI, assim como os termos de cada tópico e os pesos atribuídos a cada termo. Dentre os objetivos específicos, estão: identificar o melhor conjunto de resultados extraídos por meio da modelagem de tópicos; correlacionar os tópicos emergentes com as linhas de pesquisa do PPGCI; discutir sobre tópicos cujos resultados apresentam ruídos; apresentar os termos mais frequentes nos diferentes tipos de n-gramas[1], sendo extraídos do *corpus* unigramas, bigramas e trigramas.

Pressupõe-se que o mapeamento científico do PPGCI possa contribuir de forma prática, metodológica e/ou científica para a tomada de decisões estratégicas na Escola da Ciência da Informação (ECI) da Universidade Federal de Minas Gerais. Espera-se sugerir novas frentes de pesquisa e novos temas de interesse, bem como apontar lacunas a serem preenchidas.

A pesquisa se justifica porque o mapeamento científico permite apresentar, com base nos dados, novos resultados e prospectar diferentes cenários para a ciência estudada, podendo contribuir, assim, de forma prática, metodológica ou científica para futuras decisões a serem tomadas pelo PPGCI. Além disso, este é um estudo que pode ser replicado em outras circunstâncias, comparando, por exemplo, os tópicos mais relevantes de *corpora* de dados contendo documentos do tipo teses e dissertações organizados por regiões do país.

2 Ciência da Informação

O surgimento da Ciência da Informação (CI) está alinhado ao movimento acelerado das tecnologias da informação e comunicação (TICs), que, por sua vez, busca soluções tecnológicas através da informação e do conhecimento, para garantir o fluxo de uso da comunicação. Fazem parte desse contexto processos como a produção, a organização, o armazenamento, a representação, a

disseminação, a recuperação, o acesso e o uso da informação (NHACUONGUE; FERNEDA, 2015).

A CI, enquanto disciplina, investiga três pontos fundamentais para uma acessibilidade e usabilidade da informação mais bem desenvolvida, sendo eles: (a) as propriedades e comportamentos informacionais; (b) as formas que governam os fluxos informacionais, e (c) os significados do processamento da informação (BORKO, 1968). Além disso, preocupa-se com:

> [...] corpo de conhecimentos relacionados à origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação. Isto inclui a investigação da 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).

Os autores Shera e Cleveland (1977) e Capurro e Hjorland (2007) corroboram Borko ao afirmarem que os processos da informação são constituídos por sua geração, sua coleta, sua transformação, sua interpretação, sua organização, sua disseminação, seu armazenamento, sua recuperação e seu uso. Todavia, esse processo da informação passa a ter destaque no domínio particular das tecnologias modernas da área. Enquanto disciplina, a CI busca estudar o campo do conhecimento científico, tecnológico e de sistemas (CAPURRO; HJORLAND, 2007).

A CI, em sua interface com outras disciplinas, estuda áreas da psicologia, filosofia, sociologia, lógica, matemática, informática, telecomunicações, economia, direito e política (LE COADIC, 1996). São características da CI: 1) a interdisciplinaridade, pois a CI estuda as relações entre as áreas da computação e da inteligência artificial, com o objetivo de produzir pesquisas teóricas e práticas; 2) a CI está inexoravelmente associada as tecnologias da informação, contribui para transformação da sociedade moderna em sociedade da informação, e 3) a CI, bem como outras disciplinas, contribui para evolução da sociedade da informação (SARACEVIC, 1996).

A CI busca solucionar problemas de comunicação do conhecimento e de seus registros no contexto social, institucional ou individual por meio de estudos

práticos e profissionais que envolvem tecnologias informacionais (SARACEVIC, 1996). O conceito de informação e o de signo, por meio de registro e transmissão da informação num suporte tecnológico, é assim definido por Le Coadic (1996) como:

> 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).

No Brasil, a CI foi introduzida na década de 1970 com um curso de mestrado *stricto sensu* pelo Instituto Brasileiro de Bibliografia e Documentação (IBBD), atual Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). No ano de 1972, foi criado o periódico *Ciência da Informação* (PINHEIRO, 1997; RUSSO, 2010).

A área de concentração do PPGCI da UFMG, que oferece cursos de mestrado e doutorado na modalidade *stricto sensu*, denomina-se *Informação, Mediação e Cultura* e se estrutura em três linhas de pesquisa: 1) Memória social, patrimônio e produção do conhecimento; 2) Políticas públicas e organização da informação, e 3) Usuários, gestão do conhecimento e práticas informacionais. Os cursos buscam trabalhar de maneira transversal, avançada e aprofundada, por meio de um plano teórico e técnico sobre assuntos, documentos e informação, e têm como base os cursos de Arquivologia, Biblioteconomia e Museologia, aproveitando, assim, a especificidade de cada curso (PPGCI, 201?a).

São destacados no PPGCI da UFMG os seguintes pontos cronológicos (PPGCI, 201?b):

a) criação do curso *stricto sensu* na modalidade de mestrado em Biblioteconomia, com áreas de concentração em Educação, Biblioteconomia e Informação Especializada, no ano de 1976;

b) alteração do nome do curso para Ciência da Informação, no ano de 1991. Com isso, o curso deixa de concentrar-se nos suportes de registros da biblioteca para se concentrar na informação contida neles. Com essa alteração, foram integradas as linhas de pesquisas Informação Gerencial, Informação Científica e Tecnológica, Informação Social e Informação Histórica;

c) criação dos cursos de graduação de Arquivologia, em 2008, e de Museologia, em 2009, e consequente ampliação do âmbito em que a ECI atuava na modalidade de graduação, circunscrito, desde 1963, ao curso de Biblioteconomia;

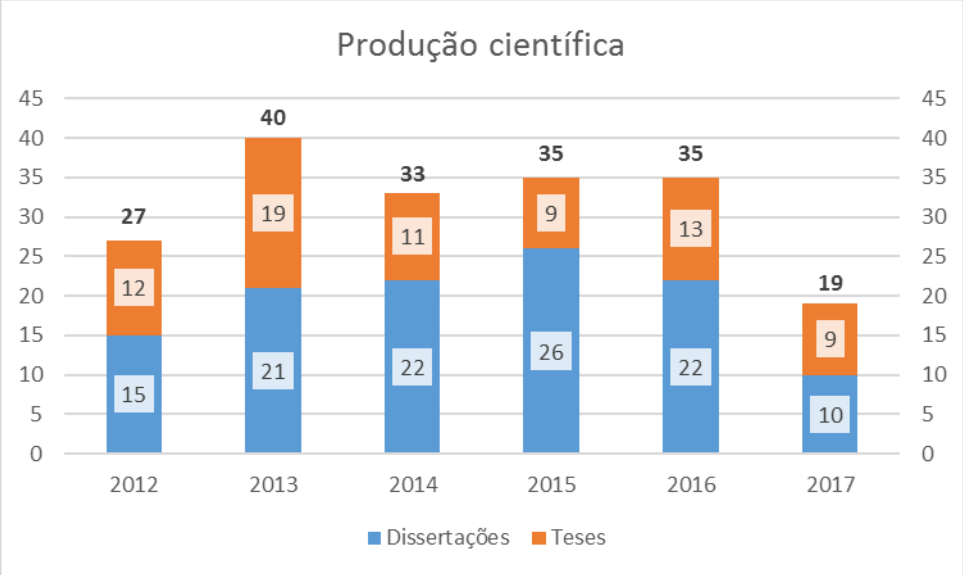
d) implantação, em 2011, de uma nova proposta curricular para o PPGCI, buscando atender a formação dos discentes dos três cursos de graduação, além das alterações nas linhas de pesquisa de Informação, Cultura e Sociedade, Gestão da Informação e do Conhecimento e Organização e Uso da Informação;

e) mudança, em 2017, no escopo e na formação do PPGCI; as áreas de concentração foram alteradas para Informação, Mediações e Cultura, sendo as linhas de pesquisa: Memória Social, Patrimônio e Produção do Conhecimento; Políticas Públicas e Organização da Informação; Usuários, Gestão do Conhecimento e Práticas Informacionais.

O PPGCI da UFMG tem produzido pesquisas nas mais diversas áreas que contemplam suas linhas de pesquisa. No ano de 2012, foram defendidas 15 dissertações e 12 teses, totalizando, assim, 27 pesquisas. No ano de 2013, ocorreu um aumento representativo de trabalhos defendidos referente ao ano anterior em ambas as modalidades; foram 21 dissertações (29%) e 19 teses (37%), ou seja, um total de 40 (33%) trabalhos defendidos. No ano de 2014, houve um aumento para 22 dissertações (5%), uma queda nas teses para 11 (-73%). Um total de 33 pesquisas defendidas (-21%). No ano de 2015, foram produzidas 26 dissertações (15%) e nove teses (-22%); logo, o total de trabalhos defendidos foi de 35 (6%). No ano de 2016, foram defendidas 22 dissertações (-18%) e 13 teses (31%). O total de trabalhos defendidos no ano de 2016 foi o mesmo do ano de 2015. Já no ano de 2017, houve uma queda representativa nos

trabalhos defendidos, uma vez que a coleta dos dados foi realizada no mês de abril. Dessa forma, foram coletadas dez dissertações (-120%), nove teses (-44%) e um total de 19 trabalhos (-84%), conforme apresentado no Gráfico 1. Os percentuais foram calculados com o objetivo apresentar a taxa de crescimento ou redução da produção científica, comparando-as ao ano anterior.

Gráfico 1 – Teses e dissertações defendidas no PPGCI da UFMG entre 2012 e abril de 2017.



Fonte: Elaborado pelos autores.

A próxima seção apresenta os conceitos sobre modelagem de tópicos, pela qual se pode organizar e resumir um *corpus* por meio de modelos generativos probabilísticos, e destaca o modelo *latent Dirichlet allocation* (LDA).

3 Modelagem de tópicos

A modelagem de tópicos permite organizar e resumir grandes coleções de arquivos eletrônicos por meio de métodos estatísticos e algoritmos de *machine learning*. Esse processo, bem como descobrir e analisar temas, identificar padrões e suas respectivas relações entre termos de um determinado *corpus*, acompanhar as mudanças dos termos ao longo do tempo por meio de suas estruturas latentes, pode se tornar uma atividade lenta e factível de erros quando realizado sem recursos tecnológicos. As estruturas latentes são conjuntos formados por tópicos que reúnem palavras semanticamente próximas e que são

determinantes para a escolha das palavras que formaram o documento, de acordo com o modelo idealizado (BLEI, 2012; KASZUBOWSKI, 2016).

Os tópicos dos modelos são constituídos por um conjunto de palavras importantes extraídas automaticamente e não supervisionadas dos documentos contidos em um ou mais *corpora*. Os algoritmos não possuem informações sobre os assuntos contidos nos documentos ou rótulos predefinidos, como, por exemplo, títulos, palavras-chave ou áreas de concentração. Para isso, são necessárias medidas de coerência para analisar tópicos bons e ruins, uma vez que não existe garantia de interpretações dos resultados (BLEI; NG; JORDAN, 2003; BLEI, 2012).

Os algoritmos perpassam por um processo de treinamento a partir de um conjunto de entrada de dados; a partir desse conjunto, busca-se fazer previsões até que se atinjam níveis satisfatórios. O objetivo é explorar os dados e apresentar estruturas, mas o computador realizará determinados procedimentos de *machine learning* sem informar como (AYODELE, 2010; PUSTEJOVSKY; STUBBS, 2012).

A modelagem de tópicos utiliza métodos probabilísticos e conceitos da área da ciência da computação e da estatística para compreender o conteúdo e os documentos utilizados em grandes coleções de documentos (BLEI, 2012; KASZUBOWSKI, 2016). Um dos grandes desafios está em desenvolver interfaces inteligentes para obter uma melhor interação homem-máquina e para assim proporcionar aos usuários uma melhor recuperação da informação que procuram (HOFMANN, 1999).

Os modelos probabilísticos de extração de tópicos utilizam premissas nas quais os documentos são representados em um conjunto de tópicos misturados. Com isso, um tópico é constituído por uma distribuição probabilística de palavras. São destacados três processos: 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 (STEYVERS; GRIFFITHS, 2007).

A modelagem de tópicos é realizada após a definição do quantitativo de tópicos que serão abordados em um *corpus*. Esses tópicos são responsáveis por

determinar os termos que serão utilizados no documento. Por meio de um modelo generativo, pelo qual o documento emerge e os parâmetros utilizados para a construção dos documentos são desconhecidos, torna-se possível estimar o quantitativo de termos a partir dos documentos e das palavras por meio das variáveis observadas (SANTOS, 2015).

Ainda de acordo com Santos (2015), o processo de extração de tópicos está intricadamente associado à melhor maneira de estimar os parâmetros responsáveis por originar os documentos de um *corpus*, passando a assumir o modelo generativo. A partir dessa etapa, obtêm-se resultados de uma representação denominada documento-tópico, responsável por determinar o peso de cada tópico, de cada documento de e uma representação termo-tópico, que se relaciona ao modelo generativo escolhido (SANTOS, 2015).

O modelo generativo probabilístico LDA adota uma abordagem bayesiana e parte do princípio de que os documentos de um *corpus* sejam representados por misturas aleatórias de tópicos latentes. Dessa forma, cada tópico passa a ser caracterizado por uma distribuição de palavras que compreendem cada um dos documentos (BLEI, 2012). O modelo é constituído por três níveis: a) os itens de uma coleção são moldados por meio de uma mistura finita do conjunto subjacente de tópicos; b) cada tópico é moldado por uma mistura infinita do conjunto subjacente de probabilidade de tópicos; e c) as probabilidades dos tópicos resultam em uma representação de um documento (BLEI; NG; JORDAN, 2003).

Quando se aplica o modelo LDA em um *corpus*, os tópicos são interpretáveis como temas na referida coleção e as representações dos documentos remetem ao tema de cada documento. Para isso, as variáveis ocultas e aleatórias codificam a estrutura temática, os tópicos aprendidos resumem a coleção e a representação dos documentos, e os *corpora* são organizados pela representação do documento (CHANEY; BLEI, 2012).

Ao executar o modelo LDA, é determinado um número fixo de tópicos e uma variável aleatória é responsável por atribuir, a cada tópico, uma probabilidade de distribuição associada às palavras contidas no *corpus*. Dessa forma, é pensada a probabilidade de visualizar a palavra para determinado

tópico. Outra distribuição pode ser obtida de maneira aleatória para cada documento, atribuindo a probabilidade de distribuição do tópico e sendo considerado uma mistura de tópicos no documento. Nesse caso, as palavras são geradas, inicialmente, pela escolha aleatória do tópico, sendo uma distribuição do tópico no documento. Posteriormente, é gerada a palavra referente à distribuição das palavras dos tópicos (GRUS, 2016).

4 Metodologia

A fase empírica da pesquisa foi adaptada de McKinney (2018), cumprindo-se as seguintes etapas:

a) interação com o mundo externo - constituição do *corpus* de teses e dissertações do PPGCI da Universidade Federal de Minas Gerais. A amostragem totaliza 189 teses e dissertações, na íntegra, publicadas na Biblioteca Digital de Teses e Dissertações da própria instituição entre 2012 e abril de 2017, intervalo também utilizado para coleta de dados e formação de *corpora* para outras pesquisas envolvendo modelagem de tópicos. Os documentos que compõem o *corpus* foram coletados em abril de 2017 por meio da extensão de navegador *Copy All URLs* e do *software Download Accelerator Plus*, para gerenciar o *download* dos documentos;

b) pré-processamento e preparação - limpeza, manipulação, combinação, normalização, tratamento e transformação dos dados para a análise descritiva. Os documentos foram convertidos para um formato legível pelo computador e foi estabelecido um padrão de caixa-baixa para os caracteres. Foram eliminados caracteres problemáticos de Unicode gerados no processo de conversão de documentos como "\x0c6011", "\uf0fc". Posteriormente, ocorreu uma conversão de siglas em termos e vice-versa, por meio de expressões regulares, para que os pesos dos termos não fossem divididos com palavras com o mesmo significado. Em seguida, foram eliminadas as *stop words* – palavras de parada como "e", "como" e "são".

c) transformação - operações matemáticas e estatísticas aplicadas em conjuntos de dados, com o objetivo de obter novos conjuntos de dados;
d) modelagem e processamento - conexão dos dados já tratados ao modelo LDA;
e) apresentação - visualizações gráficas ou sínteses textuais;
f) documentação - análise e discussão dos resultados.

Para alcançar os resultados, foram utilizados o *framework Jupyter Notebook*, a linguagem de programação *Python* e as bibliotecas *Pdfminer*, *Gensim*, *NLTK*, *Numpy*, *Matplotlib* e *Plotly*. O processador utilizado foi um Intel Core i7 – 2630QM 2.00 GHz com memória de 8 GB.

A pesquisa se classifica, quanto à finalidade/natureza, como aplicada; quanto à abordagem do problema, como qualiquantitativa; quanto aos objetivos, como exploratória (GIL, 2010).

5 Resultados e discussões

O algoritmo de *machine learning*[2], por meio do modelo LDA utilizado, foi configurado para identificar 6, 8, 10, 12, 14, 16, 18 e 20 tópicos junto ao *corpus* de dados. Cada tópico tem um conjunto de 10 palavras que o representam melhor e cada palavra é acompanhada por um peso de representatividade. Além disso, configurou-se: (1) *chunksize* = 300, número de documentos a serem utilizados em cada bloco de treinamento. Optou-se por tal como margem de segurança para trabalhar com a totalidade dos documentos; (2) *passes* = 40, número de passagens de treinamento pelos documentos, e (3) *iterations* = 250, número máximo de iterações no *corpus*, ao inferir a distribuição de tópico de um *corpus*. O equipamento utilizado para processamento e as configurações utilizadas junto ao modelo LDA resultaram em um tempo total de *machine learning* de 8h14min02s.

Algumas características podem ser encontradas entre tópicos, termos e pesos, como, por exemplo, tópicos e termos generalistas, que indicam assuntos gerais junto ao domínio de linguagem, e tópicos e termos especialistas, que permitem ao especialista criar suposições de nomes para o tópico de maneira mais assertiva, através da análise de assunto e de termos-chave, que, por sua

vez, permite ao indexador localizar informações complementares no *corpus* através do termo e assim ter acesso a mais informações.

Também é possível identificar tópicos e termos com características fortes e fracas, nos quais são considerados os pesos e a qualidade dos termos. O Quadro 1 apresenta exemplos de resultados de tópicos com características fracas, contendo ruídos. A primeira coluna do quadro diz respeito ao modelo configurado para extrair 16, 18 e 20 tópicos. Já a segunda coluna apresenta a seleção de um determinado tópico do conjunto de tópicos extraídos entre os resultados da modelagem de tópicos.

Quadro 1 – Tópicos com ruídos

MODELO	TÓPICOS
16 tópicos	Tópico 4: 0.000*"porexemplo" + 0.000*"americanpsychologicalassociation" + 0.000*"veerkampeyoshikawa" + 0.000*"eportanto" + 0.000*"cutrell" + 0.000*"konstan" + 0.000*"czerwinski" + 0.000*"taylor_francis" + 0.000*"ouseja" + 0.000*"conaill"
18 tópicos	Tópico 12: 0.000*"newyork" + 0.000*"usa.proceedings" + 0.000*"newyork_usa" + 0.000*"spiteri,1998" + 0.000*"heidelberg_germany" + 0.000*"lima,2004a" + 0.000*"germany" + 0.000*"gaithersburg_usa.proceedings" + 0.000*"gaithersburg" + 0.000*"usa.proceedings_newyork"
20 tópicos	Tópico 11: 0.000*"mccallum" + 0.000*"ouseja" + 0.000*"peng_mccallum,2004" + 0.000*"granitzeretal.,2012b" + 0.000*"mccallum,2004" + 0.000*"peng" + 0.000*"zoologia" + 0.000*"porexemplo" + 0.000*"hanetal.,2003" + 0.000*"lafferty_mccallum_pereira,2001"

Fonte: Elaborado pelos autores.

Embora os resultados apresentados no tópico 4 do modelo 16 permitam inferir um possível tópico, como, por exemplo, literatura, associando as palavras **0.000*"czerwinski"** – autor da área de Psicologia –, **0.000*"konstan"** – autor da área da Ciência da Computação e da Engenharia – e **0.000*"veerkampeyoshikawa"** – autores que abordam processos de design de modelagem associados ao termo **0.000*"americanpsychologicalassociation"**, pode-se perceber a existência de palavras que não contribuem para essa inferência, como **0.000*"porexemplo"**, **0.000*"eportanto"** e **0.000*"ouseja"**, caracterizadas as palavras como ruídos por não apresentarem características de relevância ao tópico. Os termos tratados como ruídos podem ser excluídos através de uma lista adicional de *stop words*, além da já utilizada junto ao algoritmo.

Além disso, é possível perceber a existência de termos duplicados, dificultando ainda mais a suposição para o nome do tópico. Com isso, os resultados se tornam menos precisos. O mesmo ocorre para os modelos com 18 e 20 tópicos, tendendo, assim, a aumentar o quantitativo de ruídos para cada modelo superior a esses quantitativos de tópicos e tornando-os cada vez mais modelos com características fracas nesse *corpus*.

Já o modelo que melhor representou o *corpus* foi o de 14 tópicos, treinado durante 1h11min28s. Ele apresentou conjuntos de tópicos com quantitativo de ruídos menor do que o dos demais resultados. O Quadro 2 apresenta os tópicos e os termos que compõem cada tópico e seus referidos pesos, calculados pelo modelo LDA.

Quadro 2 – Tópicos, termos e pesos do modelo LDA aplicado ao *corpus*

TÓPICO	TERMOS E PESOS
1	0.002*"quadrinhos" + 0.001*"histórias_quadrinhos" + 0.001*"gyn" + 0.001*"super-heróis" + 0.000*"quadrinhos_super-heróis" + 0.000*"comics" + 0.000*"marvel" + 0.000*"histórias" + 0.000*"histórias_quadrinhos_super-heróis" + 0.000*"core_comics"
2	0.008*"informação" + 0.003*"pesquisa" + 0.003*"conhecimento" + 0.002*"dados" + 0.002*"trabalho" + 0.002*"forma" + 0.002*"informações" + 0.002*"processo" + 0.002*"relação" + 0.002*"documentos"
3	0.001*"alinhamento" + 0.001*"ontologias" + 0.001*"baixa" + 0.000*"multinucleate" + 0.000*"alta" + 0.000*"ontologia" + 0.000*"mpd" + 0.000*"uninucleate" + 0.000*"cobertura" + 0.000*"visualização"
4	0.000*"ontouml" + 0.000*"oled" + 0.000*"parte-todo" + 0.000*"grp" + 0.000*"modelo_uml" + 0.000*"teste_oled" + 0.000*"uml" + 0.000*"relações_parte-todo" + 0.000*"modelo_ontouml" + 0.000*"critérios_ontológicos"
5	0.000*"informação" + 0.000*"pesquisa" + 0.000*"trabalho" + 0.000*"conhecimento" + 0.000*"museu" + 0.000*"relação" + 0.000*"forma" + 0.000*"processo" + 0.000*"biblioteca" + 0.000*"informações"
6	0.001*"regimes" + 0.001*"vagas" + 0.001*"regime" + 0.001*"power" + 0.001*"reservadas" + 0.001*"vagas_reservadas" + 0.000*"candidatos" + 0.000*"governança" + 0.000*"cotas" + 0.000*"governance"
7	0.000*"religião" + 0.000*"religiões" + 0.000*"cristianismo" + 0.000*"teologia" + 0.000*"classificação_bibliográfica" + 0.000*"decimal" + 0.000*"classificação_decimal" + 0.000*"espírita" + 0.000*"sistemas_classificação_bibliográfica" + 0.000*"espiritismo"
8	0.000*"exames" + 0.000*"laboratoriais" + 0.000*"hepatites" + 0.000*"hepatites_virais" + 0.000*"regras_associação" + 0.000*"virais" + 0.000*"kdd" + 0.000*"exames_laboratoriais" + 0.000*"hepatite" + 0.000*"testes_laboratoriais"
9	0.003*"ontologia" + 0.003*"ontologias" + 0.002*"domínio" + 0.001*"multimídia" + 0.001*"conteúdo" + 0.001*"ontology" + 0.001*"escopo" + 0.001*"classes" + 0.001*"mpeg7" + 0.001*"metadados"
10	0.002*"cartas" + 0.001*"carta" + 0.001*"jornal" + 0.000*"monde" + 0.000*"publicada" + 0.000*"enviada" + 0.000*"enviada_publicada" + 0.000*"carta_enviada" + 0.000*"carta_enviada_publicada" + 0.000*"leitoras"

11	0.000*"informação" + 0.000*"pesquisa" + 0.000*"biblioteca" + 0.000*"conhecimento" + 0.000*"uso" + 0.000*"trabalho" + 0.000*"relação" + 0.000*"processo" + 0.000*"dados" + 0.000*"bibliotecas"
12	0.001*"neural" + 0.001*"artigos" + 0.001*"networks" + 0.001*"neural_networks" + 0.001*"network" + 0.001*"neural_network" + 0.000*"referências" + 0.000*"rna" + 0.000*"categorização" + 0.000*"recurrent"
13	0.001*"performance" + 0.001*"convenção" + 0.001*"docentes-pesquisadores" + 0.001*"cancerologia" + 0.001*"trabalho_infantil" + 0.001*"cancer" + 0.001*"campo_cancerologia" + 0.000*"journal" + 0.000*"musical" + 0.000*"barreiras"
14	0.005*"informação" + 0.005*"biblioteca" + 0.004*"bibliotecas" + 0.003*"digital" + 0.003*"usuários" + 0.002*"pesquisa" + 0.002*"uso" + 0.002*"usuário" + 0.001*"bibliotecário" + 0.001*"inclusão"

Fonte: Elaborado pelos autores.

A modelagem de tópicos não denomina um determinado nome para cada tópico. Sendo assim, os possíveis nomes ou assuntos tratados em cada tópico devem ser supostos por meio da análise dos resultados, com base nas palavras e pesos. Dessa forma, supõe-se que os tópicos abordem os seguintes assuntos dentro do contexto do domínio da linguagem do *corpus*:

a) **tópico 1** - Literatura. História em quadrinhos;
b) **tópico 2** - Gestão da informação e do conhecimento;
c) **tópico 3** - Ontologia. Organização e representação do conhecimento;
d) **tópico 4** - Ontologia. Organização e representação do conhecimento;
e) **tópico 5** - Biblioteca. Museu. Espaços informacionais;
f) **tópico 6** - Ações afirmativas;
g) **tópico 7** - Religião. Organização e representação do conhecimento;
h) **tópico 8** - Informação e saúde;
i) **tópico 9**- Ontologia. Organização e representação do conhecimento;
j) **tópico 10** - Leitura;
k) **tópico 11** - Biblioteca. Espaços informacionais;
l) **tópico 12** - Redes neurais. Organização e representação do conhecimento;
m) **tópico 13** - Comunicação científica;
n) **tópico 14** - Uso e usuário da informação.

Os tópicos 1 e 10 apresentam resultados que remetem à literatura, mas com abordagens diferentes. Enquanto o tópico 1 apresenta literatura sobre história em quadrinhos, com pesos representativos de **0.002*"quadrinhos"** +

0.001*"histórias_quadrinhos", o tópico 10 apresenta literatura sobre cartas e jornais, com os pesos **0.002*"cartas"**, **0.001*"carta"** e **0.001*"jornal"**.

O tópico 2 tem a maior representatividade de todo o *corpus*, mediante os valores dos pesos atribuídos às palavras. O conjunto peso/palavra com maior representatividade do *corpus* também está no tópico 2: **0.008*"informação"**.

Percebe-se que os tópicos 3, 4 e 9 apresentam resultados na mesma área, mas com termos e pesos variados. O tópico 3 apresenta três palavras com maior peso, dentre elas **0.001*"ontologias"**, que permite inferir o nome do tópico. Embora o tópico 4 não apresente a palavra "ontologia" explicitamente, percebe-se um conjunto de palavras/pesos que diz respeito ao assunto, tais como **0.000*"parte-todo"** e **0.000*"critérios_ontológicos"**. Já o tópico 9 apresenta indicativos claros do assunto nos resultados com as palavras e pesos **0.003*"ontologia"** + **0.003*"ontologias"** + **0.002*"domínio"**. Tal quantidade de tópicos resultantes sobre ontologia e organização e representação do conhecimento ocorre devido à quantidade de pesquisas realizadas na área. O tópico 12 apresenta resultados referentes à organização do conhecimento, mas aborda redes neurais.

Os tópicos 5 e 11 também apresentam similaridade no que diz respeito ao tratamento de espaços informacionais, representados pelos pesos/palavras **0.000*"biblioteca"** e **0.000*"bibliotecas"**. Além disso, há poucas variações entre as palavras em seus respectivos resultados. É necessário destacar que, em ambos os resultados, todas as palavras têm pesos iguais; destacam-se termos como **0.000*"informação"**, **0.000*"pesquisa"** e **0.000*"conhecimento",** que fazem parte do contexto de espaço informacional.

O tópico 13 apresenta uma variação de assuntos abordados no conjunto de palavras. Isto é, a variação observada entre as palavras dificulta uma suposição mais assertiva do nome ou do assunto a representar. Entretanto, o conjunto **0.001*"docentes-pesquisadores"** passa a ser palavra-chave para representar assuntos da comunicação científica e que constituem assuntos do *corpus* como **0.001*"cancerologia"**, **0.001*"trabalho_infantil"** e **0.000*"musical"**. Em contrapartida, o tópico 6 apresenta um conjunto de

pesos/palavras equilibrado e específico ao assunto de ações afirmativas, tornando possível, assim, supor o nome do tópico sem grandes dificuldades.

O tópico 14 tem, em seu conjunto de resultados, a segunda maior representação do *corpus*, devido ao conjunto dos resultados formados por pesos/palavras, como, por exemplo, **0.005*"informação"**. Os demais conjuntos indicam a suposição para o assunto tratado no tópico, por exemplo, **0.005*"biblioteca"** + **0.003*"usuários"** + **0.002*"uso"**.

Por meio da análise que foi feita do *corpus*, ainda na fase de pré-processamento, e que antecede a modelagem de tópicos, foi possível identificar os termos mais frequentes de todos os documentos, resultando em 5.034.530 unigramas, 5.034.341 bigramas e 5.034.152 trigramas. O Quadro 3 apresenta uma lista com os 25 termos mais frequentes de cada n-grama.

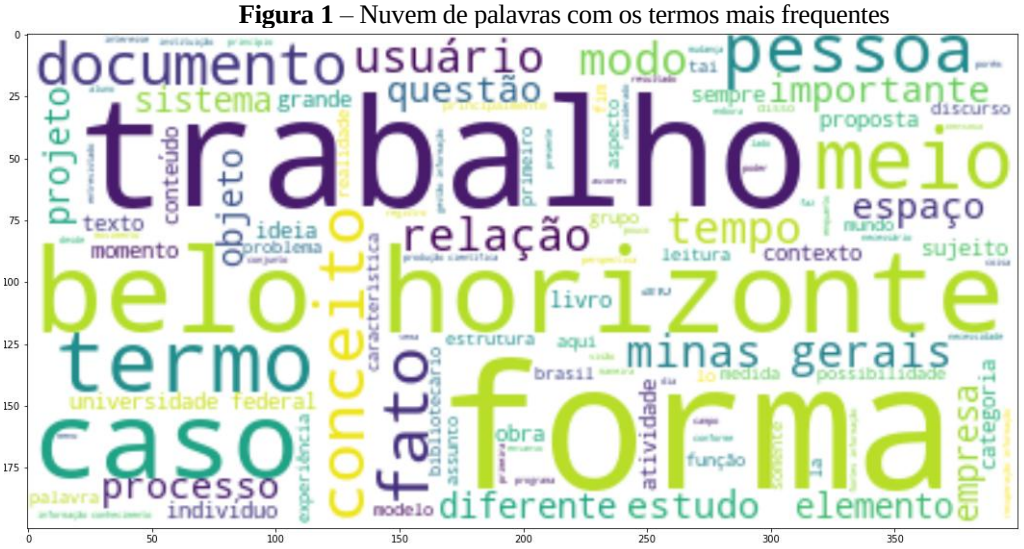
Quadro 3 – Lista de frequência de termos

Nº	Unigramas	Bigramas	Trigramas
1	informação,59877	belo_horizonte,4841	federal_minas_gerais,1234
2	pesquisa,24436	minas_gerais,3438	universidade_federal_minas,1126
3	conhecimento,21872	universidade_federal,2602	fonte_dados_pesquisa,1030
4	dados,15969	informação_conhecimento,1936	portal_periódicos_capes,928
5	trabalho,15344	produção_científica,1919	gestão_informação_conhecimento,722
6	forma,14995	recuperação_informação,1900	fonte_elaborado_autora,699
7	biblioteca,14776	fontes_informação,1742	tecnologias_informação_comunicação,581
9	informações,14353	gestão_informação,1681	informação_universidade_federal,574
10	processo,13488	gestão_conhecimento,1596	dissertação_mestrado_informação,520
11	relação,12539	escola_informação,1565	escola_informação_ufmg,481
12	uso,12181	uso_informação,1541	estado_minas_gerais,454
13	documentos,11629	fonte_elaborado,1523	international_organization_standardization,433
14	sistema,10710	ponto_vista,1468	minas_gerais_escola,412
15	social,10695	ensino_superior,1454	gerais_belo_horizonte,409
16	bibliotecas,10625	sistemas_informação,1378	acute_myeloid_leukemia,409
17	organização,10582	base_dados,1346	minas_gerais_belo,408
18	gestão,10317	coleta_dados,1326	domínio_escopo_mpeg7,406
19	usuários,9646	redes_sociais,1323	instituições_ensino_superior,396
20	desenvolvimento,9524	bases_dados,1314	gerais_escola_informação,383
21	brasil,9497	informação_tecnologia,1313	informação_belo_horizonte,379
22	busca,9396	dados_pesquisa,1277	instituição_ensino_superior,379
23	produção,9123	muitas_vezes,1276	política_nacional_arquivos,364
24	fonte,9056	tomada_decisão,1268	arquivologia_biblioteconomia_museologia,358
25	meio,8630	federal_minas,1261	web_ontology_language,346

Fonte: Elaborado pelos autores.

Na lista contendo os primeiros n-gramas mais frequentes de cada tipo, torna-se possível perceber que os unigramas têm maior frequência no que diz respeito aos termos do vocabulário da área da Ciência da Informação. Já os bigramas apresentam alguns termos que não condizem com o vocabulário da área e sim com palavras-chave de itens pré-textuais da comunicação científica. Entretanto, entre termos bigramas, como "informação_conhecimento"; "produção_científica"; "recuperação_informação"; "fontes_informação"; "uso_informação"; "tomada_decisão", "base_dados" e "redes_sociais", e entre termos trigramas, como "gestão_informação_conhecimento", "fontes_dados_pesquisa" e "arquivologia_biblioteconomia_museologia", há mais termos que fazem parte do vocabulário da Ciência da Informação do que entre termos unigramas. Além disso, há, entre termos bigramas e termos trigramas, termos que são mais representativos do que termos unigramas apresentados de forma isolada. Ainda assim, esses termos bigramas e trigramas aparecem com frequências menores. A Figura 1 apresenta uma nuvem de palavras contendo os 150 termos mais frequentes do *corpus*:

Figura 1 – Nuvem de palavras com os termos mais frequentes



Fonte: Elaborado pelos autores.

Entre os termos de cada tipo de n-grama mais bem ranqueados, estão os unigramas "informação", com frequência de 59.877; "pesquisa", com frequência de 24.436; "conhecimento", com frequência de 21.872; "dados", com frequência de 15.969; e "trabalho", com frequência de 15.344, respectivamente do primeiro ao quinto termo mais frequente extraído do *corpus*.

Já entre os bigramas mais bem ranqueados, estão "belo_horizonte", com frequência de 4.841, na posição 102; "minas_gerais", com frequência de 3.438 e posição 177; "universidade_federal", com frequência de 2.602, na posição 274; "informação_conhecimento", com frequência de 1.936, na posição 416, e "produção_científica", com frequência de 1.919 e na posição 420.

Já entre os trigramas, constam "federal_minas_gerais", com frequência de 1.234 e posição 751; "universidade_federal_minas", com frequência de 1.126 e posição 827, e "fonte_dados_pesquisa", com frequência de 1.030 e posição 911. Os próximos trigramas – "portal_periódicos_capes", com frequência de 928, e "gestão_informação_conhecimento", com frequência de 722 – apresentam frequência após o milésimo termo extraído do *corpus*.

A frequência dos termos é determinante para a construção da nuvem de palavras apresentada na Figura 1, não sendo descartados, por exemplo, termos de baixa aderência ou pouca representatividade junto ao domínio da linguagem.

6 Considerações finais

A CI, alinhada com as TICs, tem contribuído para uma melhor fluidez e para um melhor uso da informação por meio de ferramentas tecnológicas que buscam organizar, recuperar e compreender melhor um crescente volume de dados. Segundo autores da CI, são processos que envolvem a informação a geração, a coleta, a transformação, a interpretação, a organização, a disseminação, o armazenamento, a recuperação e o uso da informação.

A comunicação científica também tem contribuído, por meio de publicações formais e informais, para esse volume de dados. Entretanto, organizá-los de forma que se possam extrair novos valores de um *corpus* acaba por ser um processo longo e passível de erro, quando realizado manualmente. A partir da modelagem de tópicos, foi possível organizar e resumir uma coleção de teses e dissertações do PPGCI, da Universidade Federal de Minas Gerais.

O algoritmo de *machine learning* utilizado foi o modelo LDA, que resultou em uma melhor representação quando configurado para aprender com 14 tópicos em uma amostragem constituída por um *corpus* de 189 documentos. Os resultados são constituídos por um conjunto de palavras, às quais são

atribuídos pesos, que representam determinado tópico; entretanto, o algoritmo não gera um nome específico para cada tópico. Dessa forma, é necessário o conhecimento de um ou mais especialistas do domínio da linguagem estudada, para supor o nome ou assunto abordado em cada tópico por meio da análise dos resultados, constituídos por um conjunto de termos/pesos.

Os resultados alcançados puderam ser organizados junto às linhas de pesquisa do PPGCI da UFMG. À linha 1, *Memória social, patrimônio e produção do conhecimento*, foi associado o tópico 13 – Comunicação científica. À linha 2, *Políticas públicas e organização da informação*, foram associados o tópico 3 – Ontologia. Organização e representação do conhecimento; o tópico 4 – Ontologia. Organização e representação do conhecimento; o tópico 7 – Religião. Organização e representação do conhecimento; o tópico 9 – Ontologia. Organização e representação do conhecimento e o tópico 12 – Redes neurais. Organização e representação do conhecimento. À linha 3, *Usuários, gestão do conhecimento e práticas informacionais*, foram associados o tópico 1 – Literatura. História em quadrinhos; o tópico 2 – Gestão da informação e do conhecimento; o tópico 5 – Biblioteca. Museu. Espaços Informacionais; o tópico 6 – Ações afirmativas; o tópico 8 – Informação e saúde; o tópico 10 – Leitura; o tópico 11 – Biblioteca. Espaços informacionais, e o tópico 14 – Uso e usuário da informação.

Os temas resultantes da modelagem de tópicos realizada nas produções científicas de teses e dissertações do PPGCI da UFMG no intervalo de 2012 a abril de 2017 apresentaram uma organização por linhas de pesquisas. A primeira linha de pesquisa do PPGCI resultou em um (1) tópico. Já a segunda linha de pesquisa resultou em cinco tópicos. Por fim, a terceira linha de pesquisa resultou em oito tópicos. A partir dos resultados obtidos, podem-se prospectar diferentes estratégias de desenvolvimento para o PPGCI, como a atuação em novas atividades práticas, metodológicas ou científicas nas linhas de pesquisa com poucos tópicos, o aprofundamento de pesquisas em linhas que contemplam um maior número de tópicos ou mesmo a criação/reorganização de linhas de pesquisas.

Os objetivos da pesquisa foram atingidos por meio da modelagem de tópicos, que permitiram realizar o mapeamento científico da área por meio da classificação dos tópicos encontrados e da associação desses tópicos às linhas de pesquisa do PPGCI. O conjunto de 14 tópicos apresentou termos mais coesos e menos termos com ruídos do que os demais conjuntos. Tais ruídos podem ser reduzidos, por exemplo, ao serem adicionados a uma lista extra de *stop words*. Além disso, foram apresentados os termos mais frequentes do *corpus* estudado no formato de nuvem de palavras. Embora a frequência dos unigramas seja superior à dos demais n-gramas, os bigramas e trigramas acabam por apresentar semanticamente ao profissional indexador um maior entendimento dos termos junto do domínio da linguagem.

A modelagem de tópicos aplicada ao *corpus* apresentou resultados sólidos em um tempo de processamento de 1h11min28s para organizar e resumir 189 documentos científicos. A mesma tarefa, feita manualmente, pode levar semanas, meses ou anos até ser concluída, considerada também a subjetividade do profissional, que pode interferir nos resultados.

Sugerem-se como pesquisas futuras uma análise visual das relações dos tópicos encontrados através da biblioteca pyLDAvis e um mapeamento do comportamento anual dos principais termos do *corpus*, uma vez que a modelagem de tópicos e a frequências de termos apresentam diferentes resultados, mas são complementares uma à outra. Além disso, ressalta-se que a pesquisa pode ser aplicada em outros Programas de Pós-Graduação e que, consequentemente, podem comparar-se, por exemplo, os resultados dos tópicos, termos e frequências por escolas ou diferentes cursos, IESs, regiões do país, eventos nacionais e internacionais.

Financiamento

Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Referências

AYODELE, Taiwo Oladipupo. Types of Machine Learning Algorithms. **New Advances in Machine Learning,** *[S.l.]*: InTech, 2010. p. 19-48

BLEI, David M. Probabilistic topic models. **Communications of the ACM**, *[S.l.]*, v. 55, n. 4, p. 77–84, 1 abr. 2012.

BLEI, David M.; NG, Andrew Y; JORDAN, Michael I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, *[S.l.]*, v. 3, p. 993-1022, 2003.

BORKO, Harold. **Information science**: what is it? American Documentation, p. 5, 1968.

BRASIL. Lei n. 12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º... **Diário Oficial [da] União**, Brasília, 18 nev. 2011. Edição extra.

CAPURRO, Rafael; HJORLAND, Birger. O conceito de informação. **Perspectivas em Ciência da Informação**, *[S.l.]*, v. 12, n. 1, p. 148-207, 2007.

CHANEY, Allison J. B.; BLEI, David M. **Visualizing Topic Models**. ICWSM, 2012.

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

GRUS, Joel. **Data Science do zero:** primeiras regras com Pythhon. Rio de Janeiro - RJ: Alta Books, 2016.

HOFMANN, Thomas. **Probabilistic Latent Semantic Indexing**. 1999.

KASZUBOWSKI, Erikson. **Modelo de tópicos para associações livres**. 2016. 213 f. Universidade Federal de Santa Catarina, 2016.

LE COADIC, Yves-François. **A ciência da informação**. Tradução Maria Yêda Falcão Soares de Filgueiras Gomes. Brasília: Briquet de Lemos, 1996.

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

NHACUONGUE, Januário Albino; FERNEDA, Edberto. O campo da ciência da informação: contribuições, desafios e perspectivas. **Perspectivas em Ciência da Informação,** *[S.l.]*, v. 20, n. 2, p. 3-18, 2015.

PINHEIRO, Lena Vania Ribeiro. **A Ciência da Informação entre sombra e luz**: domínio epistemológico e campo interdisciplinar. 1997. 278 f. Tese (Doutorado em Comunicação) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1997.

PPGCI. **Programa de Pós-graduação em Ciência da Informação**: Apresentação. 201?a. Disponível em:

https://web.archive.org/web/20210312181856/http://ppgci.eci.ufmg.br/apresentacao/. Acesso em: 15 maio 2020.

PPGCI. **Programa de Pós-graduação em Ciência da Informação**: Histórico/cronologia. 201?b. Disponível em: https://web.archive.org/web/20210312182603/http://ppgci.eci.ufmg.br/historicocronologia/. Acesso em: 15 maio 2020.

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

RUSSO, Mariza. **Fundamentos de biblioteconomia e Ciência da Informação**. Editora E-papers, 2010.

SANTOS, Fabiano Fernandes dos. **Extração de tópicos baseado em agrupamento de regras de associação**. 2015. 157 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos, 2015.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspectiva em Ciência da Informação**, *[S.l.]*, v. 1, n. 1, p. 41-62, 1996.

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

STEYVERS, Mark; GRIFFITHS, Tom. **Probabilistic topic models**. Handbook of latent semantic analysis. *[S.l.]:* Lawrence Erlbaum Associates, Inc, 2007. p. 424–440.

SUKKARIEH, Jana Z.; PULMAN, Stephen G.; RAIKES, Nicholas. **Auto-marking**: using computational linguistics to score short, free text responses. Paper presented at the 29th annual conference. In: of the International Association for Educational Assessment (IAEA). 2003.

Mapping of scientific knowledge: modeling of the graduate program in Information Science of the Federal University of Minas Gerais

Abstract: The use of computational tools has been increasingly required to organize, retrieve and understand the growing volume of data. Scientific communication has contributed both formally and informally to this phenomenon. However, managing and organizing a large collection of documents may become humanly impossible, and refutable when done

manually. Topic modeling through machine learning algorithms has made it possible to organize and summarize data *corpora*. This study aims to identify the topics of the theses and dissertations by the graduate program in Information Science of the Federal University of Minas Gerais, southeastern Brazil (*Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais*). The main goal is to identify the most relevant topics of the *corpus* made up of documents such as theses and dissertations of that graduate program, such as the terms that constitute each topic as well as their respective weights. In the topic modeling we set a Latent Dirichlet Allocation model to identify 6, 8, 10, 12, 14, 16, 18 and 20 topics along with the data *corpus*. This allowed us to scientifically map the documents that we analyzed. The results obtained when the model was set to 14 topics were more cohesive and presented less noise and so allowed us to assume the names of the topics more assertively and to correlate the fields of research of the graduate program of the Federal University of Minas Gerais.

Keywords: Topic modeling. Latent Dirichlet Allocation. Machine Learning. Scientific Mapping. Information Science.

Recebido: 13/06/2020
Aceito: 20/08/2020

Declaração de autoria
Concepção e elaboração do estudo: Marcos de Souza
Coleta de dados: Marcos de Souza
Análise e discussão de dados: Marcos de Souza
Redação e revisão do manuscrito: Marcos de Souza e Renato Rocha Souza

Como citar
SOUZA, Marcos de; SOUZA, Renato Rocha. 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: http://dx.doi.org/10.19132/1808-5245273.228-250

[1] Um n-grama consiste em um pedaço de *n* caracteres, extraído de uma cadeia de caracteres maior, sendo uma palavra ou frase. Geralmente, *n* assume o valor 1, 2 ou 3, e o n-grama então é chamado respectivamente de unigrama, bigrama ou trigrama (SUKKARIEH et al., 2003).
[2] Algoritmo e resultados da Modelagem de Tópicos das teses de dissertações do PPGCI, da UFMG. Disponível em: https://web.archive.org/web/20201217115210/https://github.com/marcosdesouza82/topic-model-perspectiva.