MULTÍPLOS OLHARES EM CIÊNCIA DA INFORMAÇÃO
Vol. 9, Nº. 2 Ano 2019

MODELAGEM DE TÓPICOS

Resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina

Marcos de Souza
Universidade Federal de Minas Gerais
Marcosdesouza82@gmail.com

Renato Rocha Souza
Universidade Federal de Minas Gerais
rsouzaufmg@gmail.com

RESUMO

A pesquisa compara os resultados e desempenho dos modelos *Latent Semantic Indexing* (LSI) e *Latent Dirichlet Allocation* (LDA) de Machine Learning quando aplicado Modelagem de Tópicos em documentos dos canais formais da comunicação científica, constituído por 2006 artigos científicos e resumos expandidos do XIII ao XVII Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB). Constituem as etapas da pesquisa empírica a coleta dos dados para constituição, limpeza, manipulação, combinação, normalização, tratamento e transformação dos dados do corpus para conectar aos modelos de aprendizagem de máquina. Os modelos resumiram e organizaram o *corpus* de dados em tópicos que são constituídos por termos e pesos. O modelo LSI apresentou uma maior variedade entre os termos e pesos contidos em cada tópico, diferente do modelo LDA que apresentou uma maior similaridade nos resultados, facilitando, assim, para o especialista de domínio, criar a suposição para os nomes dos tópicos.

Palavras-Chave: Modelagem de Tópicos; Aprendizagem de Máquina; Alocação de Dirichlet Latente; Indexação Semântica Latente.

ABSTRACT

The research compares the results and performance of the Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) models of Machine Learning when applied Topic Modeling in documents of the formal channels of scientific communication, consisting of 2006 scientific articles and expanded abstracts from the XIII to the XVII Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB). The stages of empirical research are the collection of data for constitution, cleaning, manipulation, combination, normalization, treatment and transformation of corpus data to connect to machine learning models. The models summarized and organized the corpus of data into topics that consist of terms and weights. The LSI model showed a greater variety between the terms and weights contained in each topic, different from the LDA model which showed greater similarity in the results, thus making it easier for the domain specialist to create the assumption for the names of the topics.

Keywords: Topic Modeling; Machine Learning; Latent Dirichlet Allocation; Latent Semantic Indexing.

Universidade Federal de Minas Gerais

Correspondência/Contato
Av. Antônio Carlos, 6627
Pampulha: 31270-901
BELO HORIZONTE - MG

Escola de Ciência da Informação da UFMG

1. INTRODUÇÃO

O uso de ferramentas computacionais tem sido cada vez mais utilizado para organizar, recuperar e compreender os diversos tipos informações que são disponibilizadas diariamente no ciberespaço. As instituições de ensino superior, públicas ou privadas, por meio dos cursos de graduação, nas modalidades tecnológico, licenciatura ou bacharelado e pós-graduação lato ou stricto sensu tem contribuído diariamente para esse crescente volume de produção e disseminação de informações. Dentro desse cenário acadêmico, a comunicação científica produz, anualmente, uma quantidade incalculável de informação, seja por meio dos canais formais como teses, dissertações, artigos, resumos, resumos expandidos, livros ou informais como atas de reuniões, relatórios de pesquisa ou dados de pesquisa.

Resumir e organizar grandes coleções de informações pode se tornar uma tarefa humanamente impossível, suscetível de erro e exaustiva quando realizada manualmente. A Modelagem de Tópicos tem possibilitado realizar tal atividade por meio de algoritmos de *Machine Learning* que utilizam métodos estatísticos. Esses modelos buscam descobrir temas e suas relações quando aplicado em corpora de dados.

Partindo desse princípio, questiona-se: de que forma tem se apresentado os modelos de aprendizagem de máquina *Latent Semantic Indexing* e *Latent Dirichlet Allocation* quando aplicado a Modelagem de Tópicos em corpus de dados constituído por documentos dos canais formais da comunicação científica? A pesquisa possui como objeto geral resumir e organizar, por meio de modelos de aprendizagem de máquina, corpus de dados constituídos por documentos científicos. Além disso, busca analisar e comparar os resultados entre os modelos utilizados.

Pressupõe-se que resumir e organizar corpora de dados por meio de Modelagem de Tópicos pode contribuir para novas perspectivas e tomadas de decisões estratégicas com base num novo conjunto de dados resultantes de insights de *Machine Leraning*, tais como novos temas de interesse e lacunas a serem preenchidas dentro de uma determinada área de domínio. A metodologia de análise de *corpus* de dados realizada nesta pesquisa permite identificar qual modelo apresenta melhor resultado para a tomada de decisão do especialista de domínio.

2. REFERENCIAL TEÓRICO

Para compreender uma quantidade expressiva de informações, tem sido utilizado, cada vez mais, um aparato de ferramentas computacionais que possibilitam organizar, pesquisar e compreender melhor um determinado volume de informações disponibilizado na internet (HOFMANN, 1999b; BLEI, 2012). A Modelagem de Tópicos possibilita realizar a tarefa de resumir e organizar corpus de dados por meio de algoritmos de *Machine Learning* e métodos estatísticos de forma que seja possível descobrir temas e suas relações, como as mudanças dos termos ao longo dos anos (BLEI, 2012; KASZUBOWSKI, 2016).

Um dos primeiros algoritmos representativos utilizados para otimizar a extração de tópicos de corpus de dados foi descrito por Papadimitriouet et al. (1998), o modelo *Latent Semantic Indexing* (LSI). Desde então, outros modelos têm sido criados ao longo dos últimos anos. Ao se realizar a Modelagem de Tópicos em um corpus de dados, cada documento é representado com uma combinação de tópicos, sendo cada tópico representado por um conjunto de termos. Tanto os tópicos quantos os termos possuem probabilidades associadas. Dessa forma, cada tópico extraído do corpus possui um conjunto de termos relevantes e cada documento possui tópicos mais relevantes de acordo com as respectivas probabilidades.

O modelo LSI é uma variação do método de recuperação vetorial onde as dependências entre os termos dos documentos de um corpus possuem relevância em sua representação, sendo simultaneamente explorada na recuperação por meio das interrelações entre os termos dos documentos. Uma consulta, por exemplo, pode possuir similaridade nos documentos mesmo quando não compartilham palavras (DUMAIS, 1995). Dessa forma, assume-se que, nos documentos que contenham alguma estrutura subjacente ou latente no padrão de uso das palavras, o LSI utiliza técnica estatística para estimar a estrutura latente do conteúdo semântico contido no corpus coleção (DEERWESTER et al., 1990; DUMAIS, 1995).

Já o modelo *Latent Dirichlet Allocation* (LDA) é um dos modelos generativos mais utilizados para organizar grandes coleções de documentos. Trata-se de um modelo que utiliza uma abordagem bayesiana, no qual os documentos contidos em um corpus são representados como uma mistura aleatória de tópicos latentes que emergem por meio de uma estrutura não supervisionada. Cada tópico é caracterizado por uma distribuição de palavras e pesos que compreendem cada um dos documentos (BLEI, 2012).

Corpus são coleções de textos produzidos pelo homem em seu ambiente natural de comunicação e utilizados para fins específicos da pesquisa. Os documentos não são criados com o propósito de compor um corpus, porém, seu conteúdo é tratado como um fenômeno linguístico que representa uma variedade de linguagem que deve estar legível para ser interpretada por computadores (SARDINHA, 2000; PUSTEJOVSKY; STUBBS, 2012).

Machine Learning é uma das áreas da Inteligência Artificial que tem como objetivo realizar aprendizagem de sistemas computacionais por meio de algoritmos que aprendem de forma interativa em um processo repetitivo a partir de dados fornecidos. Os algoritmos de *Machine Learning* buscam encontrar insights ocultos nos dados de forma que seja possível encontrar informações específicas (VASCONCELOS; BARÃO, 2017; MACHADO, 2018).

3. METODOLOGIA

As etapas da pesquisa empírica foram adaptadas de McKinley (2018):

- coleta de dados e constituição do corpus – 2006 documentos formados por resumos expandidos e artigos completos publicados nos anais do ENANCIB (XIII ao XVII). A coleta dos documentos foi realizada por ferramentas *open source* de gerenciamento de *Uniform Resource Locator* (URL) e de downloads. Após a coleta dos documentos foi realizada a conversão de todos os arquivos do formato *Portable Document Format* (PDF) para um único documento no formato de arquivo de texto (.txt) com o tamanho de 81.020kb. A conversão foi realizada por meio de algoritmo[1] desenvolvido pelos autores;

- limpeza, manipulação, combinação, normalização, tratamento e transformação dos dados para realização da análise descritiva – destaca-se as etapas realizadas no corpus como a substituição de siglas contidas nos textos pelos respectivos nomes; adição de novas *stopwords* à lista padrão utilizada para retirada das palavras irrelevantes do corpus; utilização da função N-gramas para criação de listas de unigramas, bigramas e trigramas; utilização da função tokenização que transforma cada pa-

[1] Algoritmos para listar, ler, extrair e unir documentos no formato PDF para TXT. Disponível em: https://github.com/marcosdesouza82/topic-model/blob/master/01_listar_arquivos.ipynb e https://github.com/marcosdesouza82/topic-model/blob/master/02_leitura_arquivos.ipynb. Acesso em: 18 nov. 2019.

lavra em um token atribuindo uma identificação; utilização do modelo de *bag-of-words* usado para processamento de linguagem natural e recuperação de informações;

- conexão dos dados já tratados a modelos estatísticos e algoritmo de *Machine Learning* – nesta etapa foi utilizado o modelo *Latent Semantic Indexing* (LSI) e *Latent Dirichlet Allocation* (LDA). Ambos os modelos foram configurados para resumir e organizar o corpus de dados iniciando em 10 tópicos, sendo adicionado 4 tópicos até um total de 42. O modelo LSI utilizou a configuração padrão para realizar a Modelagem de Tópicos. Já o modelo LDA teve os parâmetros alterados: *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;

- apresentação, síntese textuais, análise e discussão dos resultados.

Para realização da Modelagem de Tópicos foi utilizado o framework *Jupyter* Notebook, a linguagem de programação *Python* e as bibliotecas *Pdfminer*, *Gensim*, NLTK, *Numpy*, *Matplotlib* e *Plotly*. A escolha do framework se deu por apresentar um ambiente de desenvolvimento iterativo, interface amigável, possibilidade para se conectar a mais de 40 tipos de linguagens de programação e por ser construída sobre algumas bibliotecas *open source*. Já a opção pela linguagem de programação *Python*, além de ser ideal para Ciência de Dados, tem se destacado na comunidade de programadores, tornando-se a terceira[2] linguagem de programação mais utilizada do mundo.

4. RESULTADOS E DISCUSSÕES

Para comparação de desempenho entre os modelos Latent Semantic Indexing (LSI) e Latent Dirichlet Allocation (LDA) após a realização da Modelagem de Tópicos[3] no corpus de dados, foi selecionada a opção contendo 10 tópicos para ambos os modelos, juntamente com suas respectivas palavras e pesos representativos.

A configuração do computador utilizado para processar os dados possui processador Intel® Core(TM) i7-8750H CPU @ 2.20GHz com 16GB de memória. O modelo LSI alcançou o resultado após 2 minutos e 51 segundos de processamento do corpus de dados conforme apresentado no quadro 01:

[2] TIOBE – the software quality company. Disponível em: https://www.tiobe.com/tiobe-index/. Acesso em: 18 nov. 2019.
[3] Modelagem de Tópicos – Algoritmos e resultados. Disponível em: https://github.com/marcosdesouza82/topic-model-forum-discente/blob/master/artigosresumos.ipynb. Acesso em: 19 nov. 2019.

Quadro 01: Tópicos do modelo LSI.

Latent Semantic Indexing
Tópico 1: 0.608*"informação" + 0.231*"dados" + 0.179*"pesquisa" + 0.143*"web" + 0.131*"conhecimento" + 0.119*"informações" + 0.112*"uso" + 0.108*"forma" + 0.104*"information" + 0.100*"usuários"
Tópico 2: -0.463*"informação" + 0.408*"dados" + 0.271*"web" + -0.180*"conhecimento" + 0.163*"metadados" + 0.161*"data" + 0.105*"usuários" + 0.104*"sistemas" + 0.102*"digitais" + -0.100*"social"
Tópico 3: -0.517*"informação" + 0.212*"memória" + 0.170*"documentos" + 0.164*"pesquisa" + 0.161*"museu" + 0.145*"conhecimento" + 0.134*"cultural" + 0.132*"museus" + 0.130*"patrimônio" + 0.122*"produção"
Tópico 4: -0.691*"conhecimento" + -0.215*"organização" + -0.194*"gestão" + 0.183*"memória" + 0.135*"informação" + 0.117*"biblioteca" + -0.113*"knowledge" + 0.099*"cultural" + 0.099*"museu" + -0.099*"organizacional"
Tópico 5: -0.282*"pesquisa" + -0.279*"dados" + -0.268*"científica" + 0.253*"documentos" + -0.229*"artigos" + -0.194*"produção" + -0.191*"periódicos" + -0.141*"autores" + -0.138*"pesquisadores" + 0.133*"organização"
Tópico 6: 0.480*"documentos" + 0.223*"arquivos" + 0.202*"gestão" + 0.196*"arquivo" + -0.190*"conhecimento" + -0.167*"biblioteca" + -0.161*"social" + -0.147*"bibliotecas" + -0.129*"sociais" + 0.128*"documento"
Tópico 7: -0.389*"biblioteca" + -0.381*"bibliotecas" + -0.249*"gestão" + -0.168*"usuários" + -0.156*"serviços" + 0.141*"museu" + 0.141*"memória" + 0.125*"representação" + 0.114*"museus" + 0.112*"informação"
Tópico 8: 0.460*"dados" + 0.214*"gestão" + 0.189*"informações" + 0.162*"data" + -0.152*"indexação" + -0.144*"usuários" + -0.142*"bibliotecas" + -0.140*"biblioteca" + 0.131*"memória" + -0.127*"termos"
Tópico 9: -0.404*"museu" + 0.361*"memória" + -0.353*"museus" + 0.294*"dados" + -0.155*"digital" + -0.154*"gestão" + 0.142*"biblioteca" + -0.135*"patrimônio" + -0.123*"museologia" + -0.119*"objetos"
Tópico 10: -0.250*"dados" + 0.234*"memória" + -0.227*"museu" + 0.216*"rede" + 0.216*"digital" + 0.209*"sociais" + 0.206*"redes" + -0.202*"museus" + 0.193*"social" + -0.153*"biblioteca"

Fonte: Elaborado pelos autores

O tempo de processamento do modelo LDA foi de 38 minutos e 45 segundos para resumir e organizar o corpus de dados conforme apresentado no quadro 02:

Quadro 02: Tópicos do modelo LDA.

Latent Dirichlet Allocation
Tópico 1: 0.006*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"dados" + 0.001*"information" + 0.001*"informações" + 0.001*"gestão" + 0.001*"forma" + 0.001*"processo" + 0.001*"social"
Tópico 2: 0.006*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"dados" + 0.001*"information" + 0.001*"informações" + 0.001*"gestão" + 0.001*"forma" + 0.001*"processo" + 0.001*"comunicação"

Tópico 3: 0.006*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"dados" + 0.001*"informações" + 0.001*"information" + 0.001*"gestão" + 0.001*"forma" + 0.001*"uso" + 0.001*"processo"

Tópico 4: 0.007*"informação" + 0.002*"pesquisa" + 0.001*"conhecimento" + 0.001*"dados" + 0.001*"informações" + 0.001*"information" + 0.001*"social" + 0.001*"comunicação" + 0.001*"gestão" + 0.001*"uso"

Tópico 5: 0.011*"informação" + 0.002*"pesquisa" + 0.002*"information" + 0.002*"conhecimento" + 0.002*"gestão" + 0.002*"dados" + 0.001*"profissionais" + 0.001*"estudo" + 0.001*"informações" + 0.001*"pesquisas"

Tópico 6: 0.006*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"dados" + 0.001*"information" + 0.001*"informações" + 0.001*"gestão" + 0.001*"forma" + 0.001*"processo" + 0.001*"uso"

Tópico 7: 0.005*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"dados" + 0.001*"produção" + 0.001*"forma" + 0.001*"processo" + 0.001*"informações" + 0.001*"social" + 0.001*"information"

Tópico 8: 0.008*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.002*"bibliotecário" + 0.001*"prontuário" + 0.001*"information" + 0.001*"informações" + 0.001*"dados" + 0.001*"profissionais" + 0.001*"competência"

Tópico 9: 0.006*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"dados" + 0.001*"informações" + 0.001*"information" + 0.001*"forma" + 0.001*"processo" + 0.001*"gestão" + 0.001*"organização"

Tópico 10: 0.005*"informação" + 0.002*"dados" + 0.002*"pesquisa" + 0.001*"conhecimento" + 0.001*"informações" + 0.001*"forma" + 0.001*"information" + 0.001*"documentos" + 0.001*"uso" + 0.001*"organização"

Fonte: Elaborado pelos autores.

Torna-se importante ressaltar que a configuração do computador utilizado para realização da Modelagem de Tópicos, tanto para o modelo LSI quanto do modelo LDA, refere-se unicamente como ponto norteador referente ao fator de tempo de processamento, uma vez que dependendo da configuração do computador e ou do tamanho dos corpora de dados, os modelos poderão variar entre horas, dias e até meses para que o processamento seja finalizado.

O modelo LSI apresentou uma maior variedade entre os termos contidos em cada tópico, porém, em alguns casos, seus respectivos pesos possuem características distintas, apresentando assim, resultados positivos e negativos entre os termos dos tópicos 2 a 10. Mesmo com essas características e tratando-se de resultados totalmente condizentes com o domínio do assunto, torna-se necessário observar que os tópicos não seguem uma padronização se comparado ao modelo LDA, o que dificulta a suposição dos nomes dos tópicos.

O modelo LDA apresentou uma maior similaridade entre os termos contidos em cada tópico, por exemplo, nos tópicos 1 e 2, que apresentaram de forma idêntica os

nove primeiros termos e pesos, sendo eles: 0.006*"informação" + 0.002*"pesquisa" + 0.002*"conhecimento" + 0.001*"dados" + 0.001*"information" + 0.001*"informações" + 0.001*"gestão" + 0.001*"forma" + 0.001*"processo" + 0.001*"social", diferenciando assim somente o último termo de cada tópico, respectivamente 0.001*"social" e 0.001*"comunicação". Dessa forma, além de serem tópicos correlacionados e que se comunicam semanticamente de alguma forma, faz-se necessária uma análise aprofundada sobre os termos e pesos para que seja realizada a suposição dos nomes dos tópicos.

5. CONSIDERAÇÕES FINAIS

Por meio da Modelagem de Tópicos, utilizando os modelos *Latent Semantic Indexing* (LSI) e *Latent Dirichlet Allocation* (LDA), foi possível resumir e organizar, em sínteses textuais constituídas por tópicos, termos e pesos, o corpus de dados constituídos por 2006 documentos, sendo artigos científicos e resumos expandidos publicados nos anais do XIII ao XVII Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB).

Ambos os modelos foram configurados para apresentarem resultados com 10, 14, 18, 22, 26, 30, 34, 38 e 48 tópicos, constituídos por 10 termos cada e seus respectivos pesos. Para comparação de desempenho entre os modelos foram selecionados os resultados contendo 10 tópicos. O modelo LDA possui uma maior flexibilidade de configuração de parâmetros quando comparado ao modelo LSI, sendo possível configurar, por exemplo, o número de documentos e de passagens de treinamento pelos documentos do corpus durante o processo de *Machine Learning*.

Dessa forma, o modelo LDA apresentou resultados de maior proximidade entre os termos e pesos de cada tópico quando comparado ao modelo LSI que, por sua vez, resultou em um distanciamento e uma variação entre os termos e pesos. Essa diferença entre os resultados pode dificultar os processos subsequentes como a suposição dos nomes dos tópicos resultantes da Modelagem.

Trata-se de uma etapa de suma importância para uma melhor organização da informação e posterior tomada de decisões estratégicas dependendo do contexto.

Esse processo de definição da suposição dos nomes dos tópicos pode ser realizado através de análise de resultados por meio de um profissional especialista da linguagem de domínio. Outra técnica empregada é a de rotulagem, que exibe os tópicos

semanticamente mais coerentes e permite realizar o agrupamento de áreas para que seja gerada uma representação de tópicos da coleção. Nesse caso, acaba por reduzir a dependência de conhecimento especializado de um profissional na linguagem de domínio.

Apontam-se duas possibilidades para a pesquisa, relacionadas à limpeza e tratamento dos dados, de forma que seja possível contribuir para um melhor resultado na Modelagem de Tópicos, independente dos modelos a serem utilizados. A primeira trata-se da Lematização, que busca encontrar a palavra no seu lema, por exemplo, "*speaking*" e "*spoke*" que se converteria no lema "*speak*". Entretanto, ainda não existem bibliotecas que possam ser utilizadas na linguagem natural dos artigos analisados, neste caso, em português. Isso acabou por resultar tópicos contendo termos como "informação", "informações" e "*information*". O não uso, neste trabalho, da Lematização pode ser considerado uma limitação para um melhor resultado alcançado pelos modelos.

A segunda técnica diz respeito ao *Stemming*, que busca fatiar o início ou o fim das palavras com a intenção de remover os afixos, porém, o uso dessa técnica deve ser monitorado para que palavras em diferentes contextos não sejam unificadas. Deve-se ficar atento em relação à técnica de redução de palavras para que não sejam alterados o contexto da Modelagem de Tópicos e a origem das palavras, tais como os termos "Livro" e "Livraria" que possuem conceitos diferentes, mas seriam reduzidos para "Livr" se tornando um único termo.

Sugere-se, para práticas futuras de Modelagem de Tópicos que utilizam mais de uma linguagem natural como esta pesquisa, que possui textos em português e abstracts em inglês, que se realize uma unificação dos principais termos da linguagem de domínio, por exemplo:

- "informação" e "*information*";
- "conhecimento" e "*knowledge*".

A unificação abre possibilidades para o surgimento de novos termos que oportunizaria ao especialista de domínio mais opções para realizar a análise e posterior suposição dos nomes dos tópicos. Outra sugestão para minimizar tal duplicidade de informações pode ser realizada durante a etapa de limpeza dos dados excluindo os abstracts, por exemplo.

AGRADECIMENTOS

Agência de fomento Fundação de Amparo à Pesquisa do Estado de Minas Gerais

REFERÊNCIAS

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.

DEERWESTER, Scott et al. Indexing by latent semantic analysis. JASIS, v. 41, n. 6, p. 391–407, 1990.

DUMAIS, Susan T. Latent Semantic Indexing (LSI): TREC-3 Report. 1995, [S.l: s.n.], 1995. p. 2019–230. Disponível em: https://pdfs.semanticscholar.org/e410/6fb9539e7fc3bf30c730a3c1d2df45d4eff6.pdf. Acesso em: 6 ago. 2019.

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.

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.

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

PAPADIMITRIOU, Christos Harilaos et al. Latent semantic indexing: a probabilistic analysis. In: ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 17., 1998, Seattle. Proceedings… New York: ACM, 1998. p. 159-168. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.193.4072&rep=rep1&type=pdf. Acesso em: 10 nov. 2019.

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

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

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.

Marcos de Souza

Pós-graduação stricto sensu (Doutorando) em Gestão e Organização do Conhecimento pela Universidade Federal de Minas Gerais - UFMG na linha de pesquisa em "Gestão & Tecnologia da Informação e Comunicação" com previsão para defesa da tese em julho de 2019; Pós-graduação *stricto sensu* (Mestrado) em Cognição e Linguagem pela Universidade Estadual do Norte Fluminense Darcy Ribeiro - UENF na

linha de pesquisa em Pesquisas Interdisciplinares em Comunicação, Educação e Novas Tecnologias da Informação; Pós-graduação lato sensu em: Informática na Educação pelo Instituto Federal do Espírito Santo - IFES; Docência do Ensino Superior pelo Centro Universitário São Camilo - Espírito Santo - CeUSC; Desenvolvimento de Aplicação para WEB pelo Centro de Ensino Superior de Juiz de Fora – CESJF e; Graduado em Sistemas de Informação pelo Centro Universitário São Camilo - Espírito Santo – CeUSC.

Renato Rocha de Souza

Pós-Doutorado pela Österreichischen Akademie der Wissenschaften, OeAW, Austria. Pós-Doutorado pela Columbia University, COLUMBIA, Estados Unidos. Pós-Doutorado pela University of South Wales, SOUTHWALES, Gales. Pós-graduação stricto sensu (Doutorado) em Ciências da Informação pela Universidade Federal de Minas Gerais – UFMG; Pós-graduação *stricto sensu* (Mestrado) em Engenharia de Produção pela Universidade Federal de Santa Catarina – UFSC; Pós-graduação *lato sensu* em Informática na Educação pela Pontifícia Universidade Católica de Minas Gerais – PUC-Minas e; Graduação em Engenharia Elétrica pela Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio.