Universidade Federal do ABC

CMCC - Centro de Matemática, Computação e Cognição

ANÁLISE EXPLORATÓRIA DA MODELAGEM DE TÓPICOS SOBRE OS ARTIGOS APRESENTADOS NO ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO (EATI) COM O MÉTODO LDA

Carlos Eduardo Ramos

Orientador: Prof. José Artur Quilici Gonzalez

Santo André

2021

UFABC

Introdução

SITUAÇÃO PROBLEMA

- Documentação da produção científica;
- Artigos científicos alocados em repositórios na internet;
  - EATI (Encontro Nacional de Tecnologia da Informação)
- Crescimento do volume dos documentos armazenados;
- Dificuldade em resumir e organizar grandes coleções de documentos manualmente.

Título	**Data mining : conceitos, técnicas, algoritmos, orientações e aplicações**
Edição	2. ed.
Imprenta	Rio de Janeiro, RJ : Elsevier, 2015.
Desc. física	xvii, 276 p. : il.
Assuntos	1. MINERAÇÃO DE DADOS 2. INTELIGÊNCIA ARTIFICIAL
Ent. sec.	I. Passos, Emmanuel (Coautor) II. Bezerra, Eduardo (Coautor)
Link do título	http://biblioteca.ufabc.edu.br/index.php?codigo_sophia=107305

Figura 1 - Detalhes de um livro consultado no portal da biblioteca da UFABC

UFABC

MODELAGEM DE TÓPICOS

- Tópicos
  - Distribuição de palavras relacionadas ao tema do tópico;
- Documentos
  - Representados como uma mistura randômica de tópicos;
- Modelagem
  - Descobrir padrões ocultos (latentes) em um corpus;
  - Temas abordados e suas relações;
  - Modelo capaz de produzir documentos similares ao corpus.

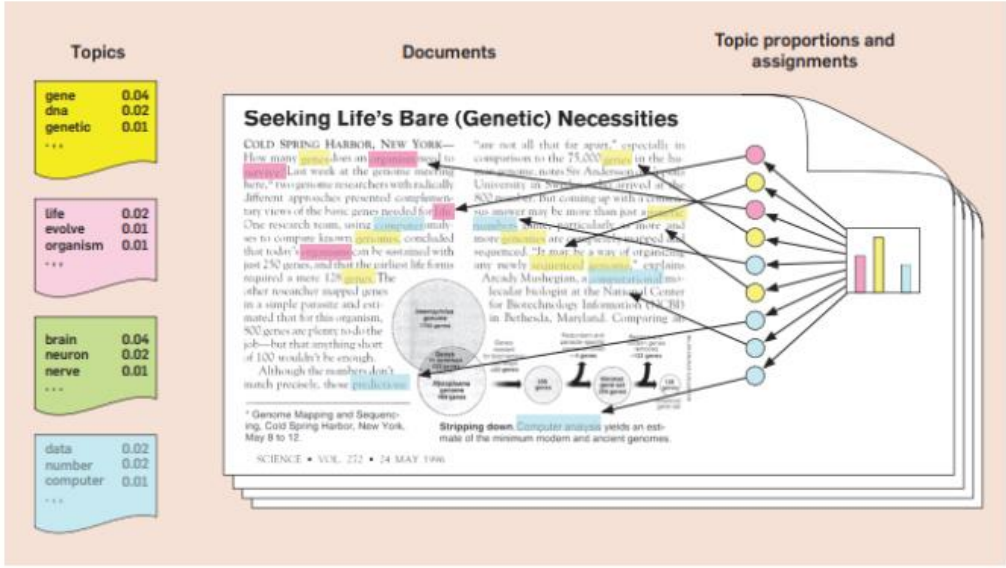

Figura 2 - Distribuição dos tópicos em um Documento. Adaptado de (BLEI, 2011)

LATENT DIRICHLET ALLOCATION (LDA)

- Algoritmo de modelagem probabilística dos tópicos;
  - Aplicado com **Machine Learning** por ***David Blei, Andrew Ng and Michael I. Jordan*** (2003);
  - Distribuição de Dirichlet.

- **Infere** os parâmetros que formam um modelo de tópicos;
  - Crescem na medida em que mais dados são observados;
  - Modelar distribuições sobre distribuições;

- Algoritmo de aprendizado não-supervisionado.

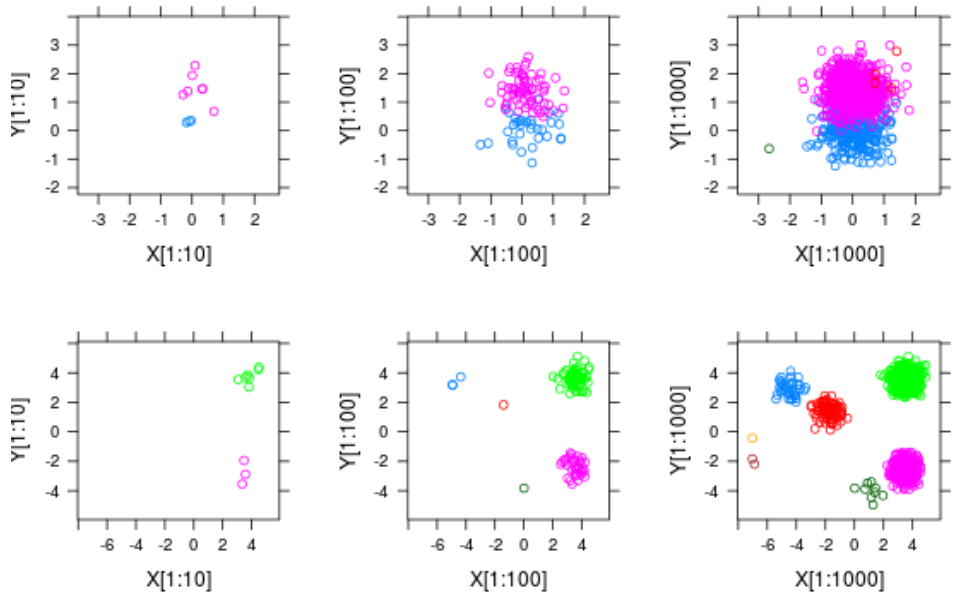

Figura 3 – Observação do modelo de mistura de Dirichlet.

UFABC

LATENT DIRICHLET ALLOCATION (LDA)

- Esquema do algoritmo

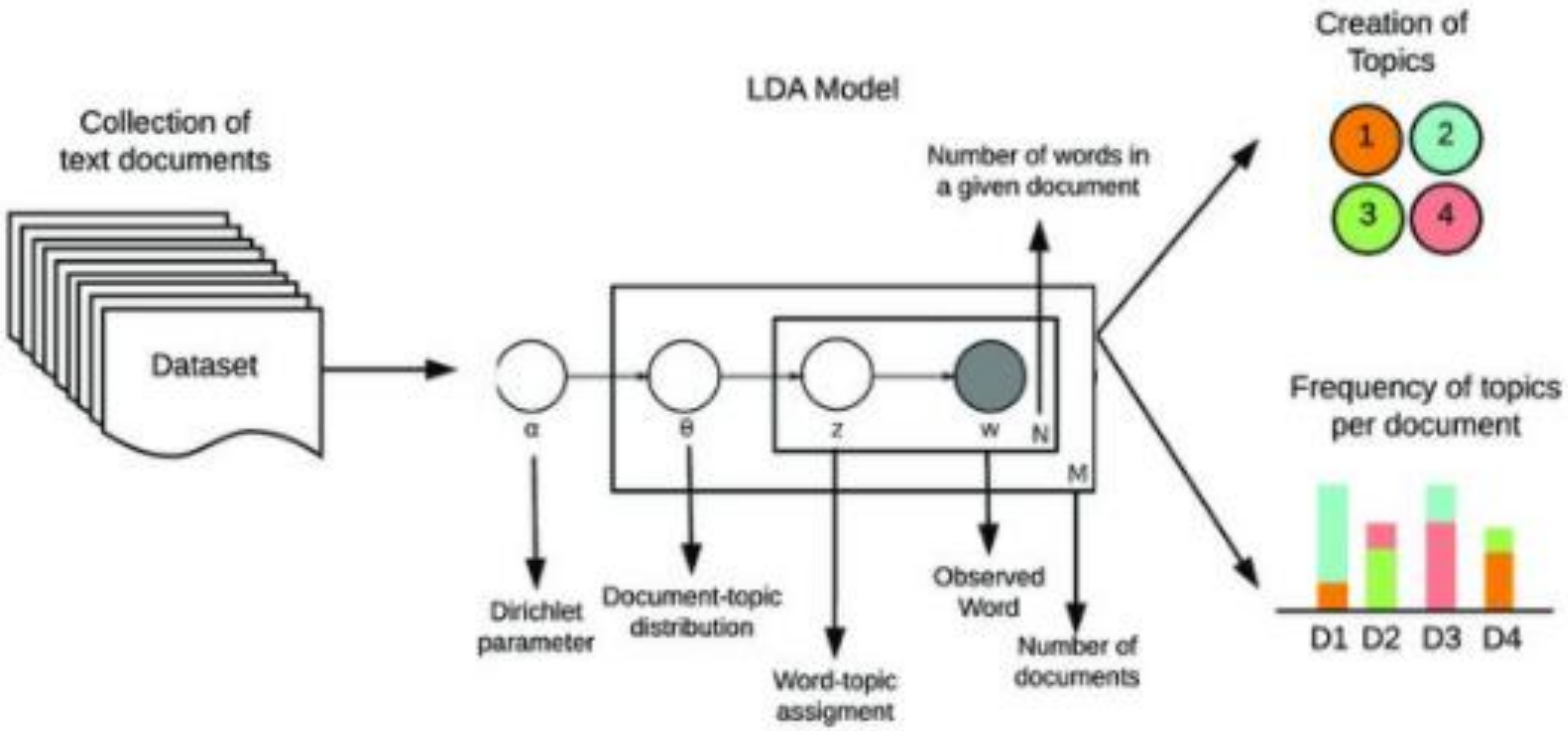

Figura 4 – Esquema do algoritmo LDA.

UFABC

OBJETIVOS

- Realizar a coleta e extração dos dados de modo automatizado;
- Encontrar os hiperparâmetros adequados para extração dos tópicos;
- Identificar os tópicos abordados nos documentos do EATI com o LDA;
- Analisar a variação e apresentação dos tópicos mais recorrentes ao longo de cada edição do EATI;
- Apresentar o comportamento dos termos mais frequentes nos artigos apresentados;
- Analisar, comparar e apresentar os resultados obtidos.

UFABC

Metodologia

METODOLOGIA

- Realizar pesquisa e levantamento bibliográfico;

- Automatizar a coleta de dados;
  - Extração dos Artigos da Base de Dados do simpósio;
  - Consolidação dos Artigos no Formato de Texto;

- Fazer o pré-processamento dos dados;
  - Limpeza e normalização dos dados

- Treinamento do modelo LDA
  - Métrica de coerência Cv
    - Cada palavra do tópico como um vetor;
    - Similaridade pelo **cosseno** entre o vetor desta palavra com o ***vetor soma*** das palavras dos tópicos;
    - Média aritmética destes valores de similaridade.

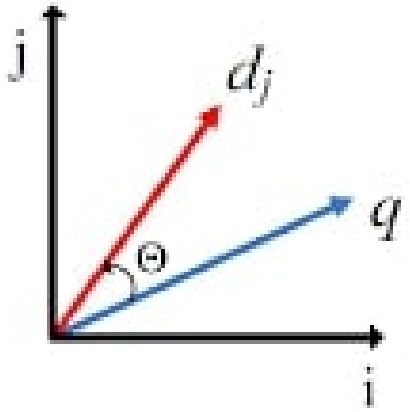

Figura 5 – Ângulo entre dois vetores.

METODOLOGIA

- Treinamento do modelo LDA
  - Obtenção dos Hiperparâmetros: **K = 50; α = "symetric" e β = 10^{-2}**

- Teste do modelo LDA
  - Corpus completo e por edições.

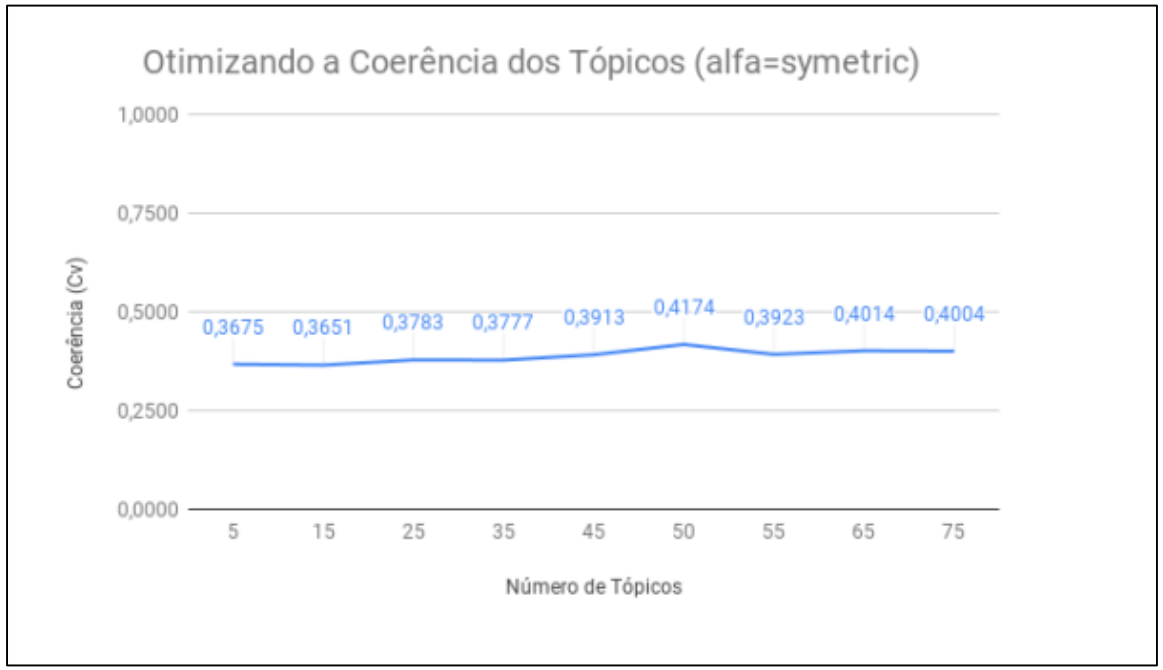

Figura 6 – Gráfico da coerência do modelo para hiperparâmetro *α* com valor *"symetric"*

.

Resultados

RESULTADOS

- **Sistemas de Informação**

Tópico	Termos e pesos do tópico
2	0.008*"sistema"; 0.006*"uso"; 0.005*"fuzzy"; 0.004*"software"; 0.004*"objetos"; 0.003*"desenvolvimento"; 0.003*"casos uso"; 0.003*"casos"; 0.003*"tabela"; 0.003*"processo"
42	0.031*"dados"; 0.008*"afirmações"; 0.008*"afirmações"; 0.007*"banco"; 0.006*"banco dados"; 0.005*"data"; 0.004*"trabalho"; 0.004*"forma"; 0.004*"analise"; 0.003*"teste"

- **Ensino, Aprendizagem e Interação Humano-Computador**

Tópico	Termos e pesos do tópico
7	0.010*"sistema"; 0.008*"desenvolvimento; 0.006*"forma"; 0.006*"usuário"; 0.005*"trabalho"; 0.005*"uso"; 0.005*"projeto"; 0.005*"ambiente"; 0.005*"aprendizagem"; 0.005*"web"
20	0.005*"docentes"; 0.004*"bloco"; 0.004*"lousa"; 0.004*"mão"; 0.003*"dedos"; 0.003*"algoritmo"; 0.003*"lousa digital"; 0.003*"distancia"; 0.003*"forma"; 0.003*"moodle"
43	0.008*"alunos"; 0.005*"atividades"; 0.005*"dados"; 0.004*"projeto" 0.004*"resultados"; 0.004*"programação"; 0.004*"desenvolvimento"; 0.003*"sistemas"; 0.003*"ferramenta"; 0.003*"ensino"

UFABC

RESULTADOS

- **Inteligência Artificial**

Tópico	Termos e pesos do tópico
5	0.006*"algoritmo"; 0.005*"dados"; 0.005*"treinamento" 0.005*"neural"; 0.005*"base"; 0.004*"aprendizagem" 0.004*"aprendizagem"; 0.004*"rede neural"; 0.004*"artificial; 0.004*"neural artificial"
34	0.012*"reforço"; 0.012*"reforço"; 0.007*"aprendizagem reforço"; 0.006*"agente"; 0.004*"ação"; 0.003*"conexionista"; 0.003*"neural"; 0.003*"neural"; 0.003*"q-learning"; 0.002*"ações"

- **Internet das coisas (IoT) e Arduíno**

Tópico	Termos e pesos do tópico
10	0.012*"dados"; 0.012*"dados"; 0.006*"aplicações"; 0.005*"cidade"; 0.005*"internet"; 0.004*"nota"; 0.004*"pagina"; 0.004*"energia"; 0.004*"iot"; 0.003*"social"
16	0.011*"energy"; 0.008*"system"; 0.007*"solar"; 0.007*"power"; 0.005*"arduino"; 0.004*"figure" 0.004*"gsm"; 0.004*"data"; 0.003*"generation"; 0.003*"grid"
31	0.007*"arduino"; 0.006*"jovens"; 0.004*"quedas"; 0.004*"pulseira"; 0.003*"eventos"; 0.003*"oficinas"; 0.003*"evento"; 0.003*"manifestações"; 0.002*"empreendedorismo"; 0.002*"cocos2d"

UFABC

RESULTADOS

- **Gestão de Projetos e Riscos**

Tópico	Termos e pesos do tópico
36	0.015*"dados"; 0.012*"informação"; 0.009*"tecnologia"; 0.009*"conhecimento"; 0.006*"tecnologia informação"; 0.004*"estratégico"; 0.004*"estratégico"; 0.004*"alinhamento"; 0.004*"data"; 0.004*"empresa"; 0.004*"projetos"
44	0.029*"informação"; 0.025*"tecnologia informação"; 0.023*"tecnologia" 0.023*"tecnologia"; 0.011*"riscos"; 0.011*"processos"; 0.010*"gestão"; 0.010*"gerenciamento"; 0.009*"serviços"; 0.006*"organização";

- **Redes, tráfego de informação e comunicação**

Tópico	Termos e pesos do tópico
6	0.010*"rede"; 0.010*"sistema"; 0.010*"sistema"; 0.005*"teste"; 0.005*"software"; 0.004*"servidor"; 0.004*"redes"; 0.004*"comunicação"; 0.004*"protocolo"; 0.004*"intrusão"
26	0.007*"flow"; 0.006*"network"; 0.005*"module"; 0.005*"router"; 0.005*"neutrality"; 0.004*"traffic"; 0.004*"metrics"; 0.004*"network neutrality"; 0.003*"figure"; 0.003*"breaking"
40	0.009*"sistema"; 0.006*"consumo"; 0.005*"dnssec"; 0.005*"dados"; 0.004*"informação"; 0.004*"estoque"; 0.004*"dns"; 0.004*"segurança"; 0.004*"web"; 0.004*"industria"

UFABC

RESULTADOS

- **Processadores, Processos e Threads**

Tópico	Termos e pesos do tópico
39	0.011*"thread"; 0.008*"documento"; 0.008*"xml"; 0.007*"threads"; 0.007*"work"; 0.006*"dados"; 0.006*"execução"; 0.006*"units"; 0.006*"work units"; 0.005*"thread vs"
41	0.012*"cluster"; 0.007*"desempenho"; 0.006*"tempo"; 0.004*"threads"; 0.004*"resultados"; 0.004*"computadores"; 0.004*"núcleos"; 0.003*"sistema"; 0.003*"beowulf"; 0.003*"número"

- **Gestão de Projetos e Riscos**

Tópico	Termos e pesos do tópico
36	0.015*"dados"; 0.012*"informação"; 0.009*"tecnologia"; 0.009*"conhecimento"; 0.006*"tecnologia informação"; 0.004*"estratégico"; 0.004*"estratégico"; 0.004*"alinhamento"; 0.004*"data"; 0.004*"empresa"; 0.004*"projetos"
44	0.029*"informação"; 0.025*"tecnologia informação"; 0.023*"tecnologia" 0.023*"tecnologia"; 0.011*"riscos"; 0.011*"processos"; 0.010*"gestão"; 0.010*"gerenciamento"; 0.009*"serviços"; 0.006*"organização";

UFABC

RESULTADOS

- **Termos mais frequentes do EATI (2011 – 2019)**

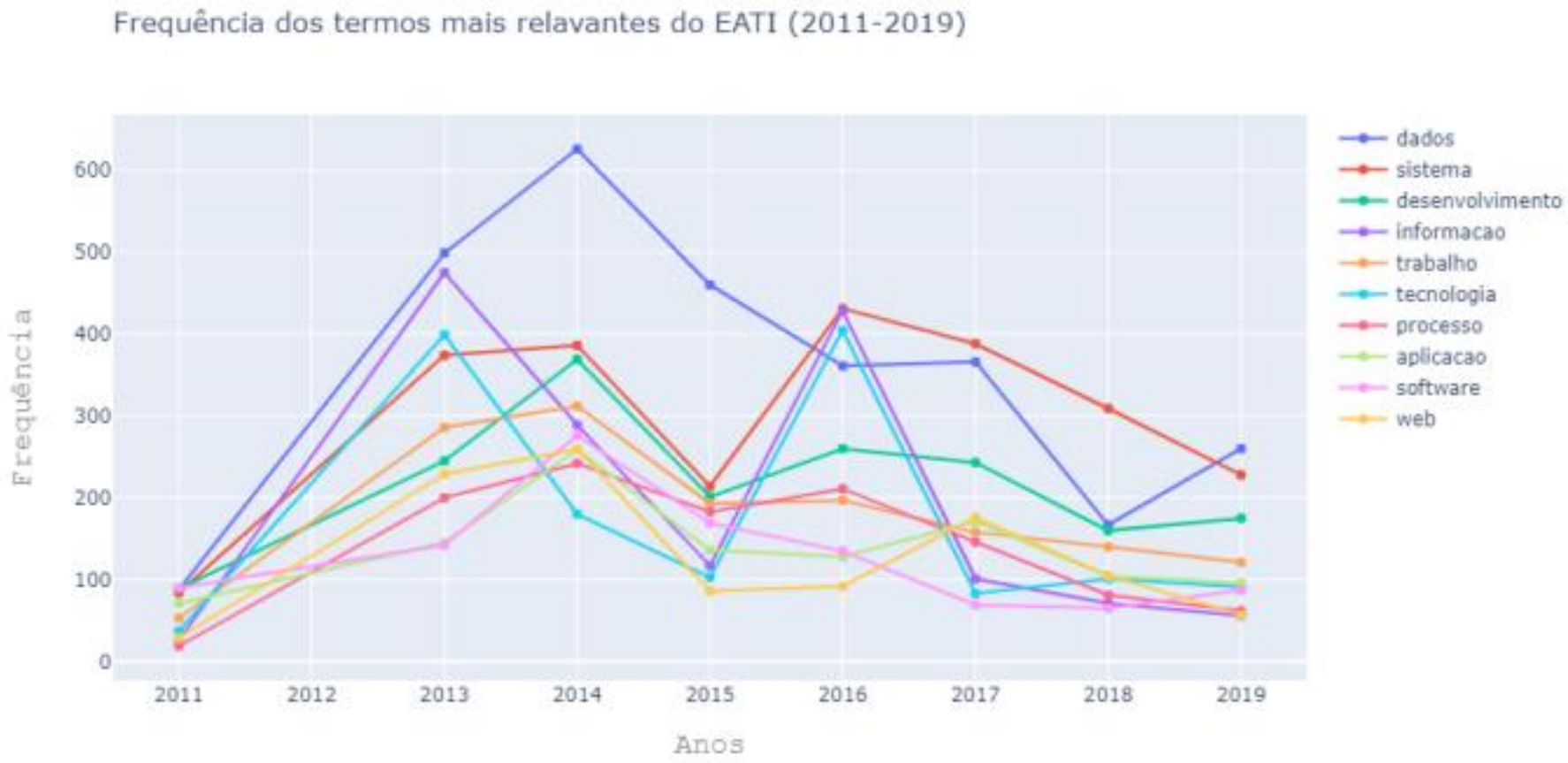

Figura 7 – Frequência dos termos mais relevantes do EATI (2011-2019)

.

UFABC

RESULTADOS

- **Edição de 2013 - A Educação na Área de TI e seus Desafios**

Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5	Tópico 6
• "web" • "informação" • "dados" • "documento" • "usuário" • "forma" • "segurança" • "xml" • "trabalho" • "conteúdo"	• "informação“, • "tecnologia“, • "tecnologia informação“, • "estratégico“, • "alinhamento“, • "dados“, • "sistema“, • "informações“, • "governança“, • "alinhamento estratégico“.	• "dados" • "riscos" • "sistema" • "projeto" • "informação" • "tecnologia" • "projetos“ • "identificação" • "analise" • "processo";	• "tempo" • "trabalho" • "desempenho" • "threads" • "speedup" • "dados" • "gpu" • "algoritmo" • "openmp" • "cpu";	• "sistema" • "jogos" • "desenvolvimento" • "forma" • "ensino" • "dados" • "processo" • "aprendizagem" • "pesquisa" • "tecnologia"	• "dados" • "alunos" • "aprendizagem" • "utilização" • "serviço" • "web" • "ambiente" • "ensino" • sistema • "informação"

UFABC

RESULTADOS

- **Edição de 2014 - As Tecnologias de Informação e Comunicação, a Academia e o Mercado de Trabalho**

Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5	Tópico 6
• "desenvolvimento" • "requisitos" • "sistema" • "modelo" • "forma" • "web" • "jogo" • "trabalho" • "dados" • "jogos"	• "dados" • "alunos" • "teste" • "aprendizagem" • "banco" • "sistema" • "banco dados" • "disciplinas" • "sql" • "atividades"	• "aplicações" • "dados" • "desenvolvimento" • "cidade" • "web" • "processos" • "processo" • "nota" • "aprendizagem" • "frameworks"	• "dados" • "trabalho" • "software" • "desenvolvimento" • "sistema" • "processo" • "informação" • "aplicação" • "tecnologia" • "forma"	• "sistema" • "desenvolvimento" • "web" • "usuário" • "aplicação" • "ambiente" • "aplicações" • "uso" • "caso" • "forma"	• "teste" • "informação" • "casos" • "casos teste" • "ambiente" • "mutantes" • "segurança" • "tecnologia" • "mutacão" • "rede"

UFABC

RESULTADOS

- **Edição de 2015 - Ensino, Pesquisa e Inovação para o Desenvolvimento Regional**

Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5	Tópico 6
• "software"	• "modelo"	• "aprendizagem"	• "dados"	• "sistema"	• "processo"
• "programação"	• "dados"	• "modelo"	• "informações"	• "rede"	• "seleção"
• "ferramenta"	• "população"	• "aluno"	• "conhecimento"	• "alunos"	• "recrutamento"
• "desenvolvimento"	• "numero"	• "usuário"	• "forma"	• "forma"	• "recrutamento seleção"
• "rede"	• "mineração"	• "alunos"	• "sistema"	• "filtro"	• "profissionais"
• "trabalho"	• "crescimento"	• "dados"	• "processo"	• "dados"	• "empresa"
• "jogo"	• "proposto"	• "forma"	• "ambiental"	• "dispositivos"	• "tecnologia"
• "engine"	• "contaminação"	• "tempo"	• "saúde"	• "desenvolvimento"	• "notas"
• "redes"	• "mineração dados"	• "reforço"	• "inteligência"	• "comunicação"	• "desenvolvimento"
• "forma"	• "taxa"	• "uso"	• "sistemas"	• "jogo"	• "usuários"

UFABC

Considerações Finais

CONSIDERAÇÕES FINAIS

- Conclusão

- Trabalhos Futuros
  - Utilização de outro algoritmo de modelagem de tópicos;
  - Rotulação de todos os tópicos com apoio de envolvidos com as áreas correlatas;
  - Identificação dos rótulos de forma automática.

UFABC

REFERÊNCIAS

BINKLEY, D.; HEINZ, D.; LAWRIE, D.; OVERFELT, J. Understanding lda in source code analysis. In: Proceedings of the 22nd international conference on program comprehension. [S.l.: s.n.], 2014. p. 26–36.

BLEI, D. Introduction to probabilistic topic models. Communications of the ACM, v. 55, 01 2011.

BLEI, D. M. Probabilistic topic models. Communications of the ACM, ACM New York, NY, USA, v. 55, n. 4, p. 77–84, 2012.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. Journal of machine Learning research, v. 3, n. Jan, p. 993–1022, 2003.

BUENAÑO-FERNANDEZ, Diego et al. Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. IEEE Access, v. 8, p. 35318-35330, 2020.

BRUNIALTI, L.; PERES, S.; FREIRE, V.; LIMA, C. Aprendizado de maquina em sistemas de recomendacao baseados em conteudo textual: Uma revisao sistematica. In: SBC. Anais do XI Simpósio Brasileiro de Sistemas de Informação. [S.l.], 2015. p. 203–210.

CAVNAR, W. B.; TRENKLE, J. M. et al. N-gram-based text categorization. In: CITESEER. Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. [S.l.], 1994. v. 161175.

EATI. Encontro Anual de Tecnologia da Informação | Sobre. 2020. Disponível em: <http://eati.info/sobre/>. Acesso em: 13 abr. 2021.

FALEIROS, T. d. P. Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais. Tese (Doutorado) — Universidade de São Paulo, 2016.

FONSECA, F. P. C. d. Inferência das áreas de atuação de pesquisadores. Tese (Doutorado) — Universidade de São Paulo, 2018.

KUCHLING, A. Regular expression howto. Regular Expression HOWTO—Python, v. 2, n. 10, 2014.

KUHLMAN, D. A python book: Beginning python, advanced python, and python exercises. [S.l.]: Dave Kuhlman Lutz, 2009.

REFERÊNCIAS

LI, S. Topic Modeling and Latent Dirichlet Allocation (LDA) in Python | Towards Data Science. 2019. Disponível em: <https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0>. Acesso em: 14 abr. 2021.

LOPER, E.; BIRD, S. Nltk: the natural language toolkit. arXiv preprint cs/0205028, 2002.

MABEY, B. Welcome to pyLDAvis's documentation. 2015. Disponível em: <https://pyldavis.readthedocs.io/en/latest/>. Acesso em: 14 abr. 2021.

MATTMANN, C. A.; ZITTING, J. L. Tika in action. Manning, 2012.

MELOTTI, G. Aplicação de autômatos celulares em sistemas complexos: um estudo de caso em espalhamento de epidemias. Universidade Federal de Minas Gerais, 2009.

NAVARRO, F. P.; CONEGLIAN, C. S.; SEGUNDO, J. E. S. Big data no contexto de dados acadêmicos: O uso de machine learning na construção de sistema de organização do conhecimento. XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XIX ENANCIB); XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XIX ENANCIB), v. 24, n. 2, 2018.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA, 2010. p. 45–50. <http://is.muni.cz/publication/884893/en>.

SELENIUM, D. Selenium webdriver. Selenium HQ, Feb, 2013. Disponível em: <https://www.selenium.dev/>.

SIEVERT, C.; SHIRLEY, K. Ldavis: A method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. [S.l.: s.n.], 2014. p. 63–70.

REFERÊNCIAS

SOUSA, D. N. F. Identificação automática de áreas de pesquisa em c&t. 2016.

SOUZA, M. d.; SOUZA, R. R. Modelagem de tópicos: Resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina. Múltiplos Olhares em Ciência da Informação-ISSN 2237-6658; Vol. 9 No. 2 (2019): PPGGOG-Discentes, v. 24, n. 2, 2018.

SOUZA, M. de; SOUZA, R. R. Modelagem de tópicos. Múltiplos Olhares em Ciência da Informação, v. 9, n. 2, 2019.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. Perspectivas em ciência da informação, SciELO Brasil, v. 11, n. 2, p. 161–173, 2006.

STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. Handbook of latent semantic analysis, v. 427, n. 7, p. 424–440, 2007.

VIEIRA, L. C. et al. Organização e disseminação da produção científica dos docentes do ccsh/ufsm em um repositório digital. Universidade Federal de Santa Maria, 2013.

WEITZEL, S. da R. O papel dos repositórios institucionais e temáticos na estrutura da produção científica. Em Questão, Universidade Federal do Rio Grande do Sul, v. 12, n. 1, p. 51–71, 2006.

ZIMAN, J. Conhecimento público. [S.l.]: Itatiaia Belo Horizonte, 1979.