

UNIVERSIDADE FEDERAL DO ABC
CMCC - CENTRO DE MATEMÁTICA, COMPUTAÇÃO E COGNIÇÃO

CARLOS EDUARDO RAMOS
Orientador: Prof. José Artur Quilici Gonzalez

**ANÁLISE EXPLORATÓRIA DA MODELAGEM DE TÓPICOS SOBRE
OS ARTIGOS APRESENTADOS NO EATI COM O MÉTODO LDA**

Santo André, SP
2021

UNIVERSIDADE FEDERAL DO ABC
CMCC - CENTRO DE MATEMÁTICA, COMPUTAÇÃO E COGNIÇÃO

CARLOS EDUARDO RAMOS

**ANÁLISE EXPLORATÓRIA DA MODELAGEM DE TÓPICOS SOBRE OS ARTIGOS
APRESENTADOS NO EATI COM O MÉTODO LDA**

Monografia apresentada ao Curso de Ciências da Computação da Universidade Federal do ABC como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. José Artur Quilici Gonzalez

Santo André, SP
2021

Ramos, Carlos Eduardo.

Análise exploratória da Modelagem de tópicos sobre os artigos apresentados no EATI com o método LDA/ Carlos Eduardo Ramos. –, 2021-
47 p. 1 :il. (colors; grafs; tabs).

Orientador: Prof. José Artur Quilici Gonzalez

Monografia – Universidade Federal do ABC,
, CMCC - Centro de Matemática, Computação e Cognição, 2021.

1. Modelagem de Tópicos. 2. LDA. 3. Artigos Científicos. 4. Mapeamento Científico. I. Prof. José Artur Quilici Gonzalez. II. Universidade Federal do ABC. III. Análise exploratória da Modelagem de tópicos sobre os artigos apresentados no EATI com o método LDA

Resumo

Devido ao avanço da tecnologia nos meios de comunicação e informação, os artefatos produzidos nestes espaços passaram a ser virtualizados. Dessa forma, temos visto uma migração destas produções para ambientes online, mais acessíveis, abertos e informatizados. A comunicação científica produz, anualmente, uma enorme quantidade de informação, principalmente na forma de artigos científicos alocados em repositórios na internet. Estes repositórios formam enormes coleções de documentos em formato digital, ao passo que resumir e organizar grandes coleções de informações pode se tornar uma tarefa humanamente impossível, suscetível ao erro e exaustiva quando realizada manualmente. A modelagem de tópicos apresenta métodos que visam resolver problemas de mineração de textos e identificação dos principais assuntos em uma coleção de documentos. Utilizada para estruturar e/ou sumarizar grandes quantidades de dados, tal metodologia fornece uma solução para o problema de gerenciar grandes arquivos de documentos. Tendo isso em vista, este trabalho visou desenvolver uma análise exploratória, a partir da extração dos principais tópicos contidos nos artigos apresentados durante as edições do Encontro Anual de Tecnologia da Informação (EATI), utilizando a técnica de extração por *Latent Dirichlet Allocation* (LDA). Além disso, buscou-se realizar a coleta dos dados de modo automatizado; suceder o treinamento da modelagem de tópicos de maneira não supervisionada; apresentar o comportamento dos termos mais frequentes nos documentos do EATI e analisar a variação dos tópicos mais recorrentes de modo cronológico, ao longo de cada edição do evento. Como resultados, foram extraídos 50 tópicos referentes a base de 310 artigos, e analisamos os tópicos com seus possíveis rótulos. Para complementar a análise dos tópicos, a distribuição dos termos mais frequentes da base de dados foi apresentada. Além disso, foi executada a extração dos principais tópicos nas edições do EATI que possuíam um tema central definido, como em 2013, 2014 e 2015.

Palavras-chave: Modelagem de Tópicos. LDA. Artigos Científicos. Mapeamento Científico.

Abstract

Because of the advancement of technology in the means of communication and information, the artifacts produced in these spaces started to be virtualized, thus, we have seen a migration of these productions to the internet. Scientific communication produces large amounts of information, mainly in scientific articles form, allocated in repositories on the network. These repositories form huge collections of documents in digital format. Summarizing and organizing large collections of information can become a humanly impossible task, susceptible to error and exhaustive when performed manually. Topic modeling presents methods that aim to solve text mining problems and to identify the main subjects in a collection of documents. This methodology serves to structure large amounts of data, and provides a solution to the problem of managing large document files. By these means, this work will develop an exploratory analysis, from the inference of the main topics contained in the articles presented during the editions of the Annual Meeting of Information Technology (EATI), using the extraction technique by Latent Dirichlet Allocation (LDA). In addition, perform data mining in an automated way; succeeding in unsupervised topic modeling training; present the behavior of the most frequent terms in the EATI documents and analyze the variation of the most recurrent topics in a chronological way, throughout each edition of the event.

Keywords: Topic Modeling. LDA. Scientific Articles. Scientific Mapping.

Lista de Ilustrações

Figura 2.1 – Produção científica dentre as edições do EATI.	6
Figura 2.2 – Distribuição dos tópicos em um documento. Adaptado de (BLEI, 2012). . .	8
Figura 2.3 – Exemplo de tópicos gerados. Adaptado de (BLEI, 2012).	9
Figura 2.4 – Pseudocódigo do algoritmo LDA.	10
Figura 2.5 – Esquema ilustrado do algoritmo LDA.	11
Figura 4.1 – Página dos anais do evento da edição de 2018 do EATI (EATI, 2020).	16
Figura 4.2 – Tabela de mapeamento das informações dos artigos da base de dados.	17
Figura 4.3 – Arquivos dos artigos extraídos em formato PDF com nomenclatura padrão. .	17
Figura 4.4 – Exemplo de cabeçalhos e rodapés presentes nos artigos da base de dados. . .	19
Figura 4.5 – Gráfico da coerência do modelo para hiperparâmetro α com valor "symetric".	22
Figura 5.1 – Mapa da distância entre os tópicos.	29
Figura 5.2 – Frequência dos termos mais relevantes do EATI (2011-2019).	30
Figura 5.3 – Palavras mais frequentes nos artigos da edição IV do EATI em 2013.	30
Figura 5.4 – Tópicos da edição IV do EATI em 2013.	31
Figura 5.5 – Palavras mais frequentes nos artigos da edição V do EATI em 2014.	31
Figura 5.6 – Tópicos da edição V do EATI em 2014.	32
Figura 5.7 – Palavras mais frequentes nos artigos da edição VI do EATI em 2015.	32
Figura 5.8 – Tópicos da edição VI do EATI em 2015.	32

Lista de Tabelas

Tabela 4.1 – Exemplo de representação matricial documento-termo.	21
Tabela 5.1 – Quadro com os tópicos 2 e 42.	23
Tabela 5.2 – Quadro com os tópicos 7, 20 e 43.	24
Tabela 5.3 – Quadro com os tópicos 5 e 34.	24
Tabela 5.4 – Quadro com os tópicos 10, 16 e 31.	25
Tabela 5.5 – Quadro com os tópicos 39 e 41.	25
Tabela 5.6 – Quadro com os tópicos 6, 26 e 40.	25
Tabela 5.7 – Quadro com o tópico 29.	26
Tabela 5.8 – Quadro com os tópicos 36 e 44.	26
Tabela 5.9 – Quadro com o tópico 24.	26
Tabela 5.10–Quadro com o tópico 8.	27
Tabela 5.11–Quadro com o tópico 15.	27
Tabela 7.1 – Cronograma de atividades 2020.	37
Tabela 7.2 – Cronograma de atividades 2021.	37
Tabela A.1 – Tópicos 0-3 extraídos da base de artigos.	43
Tabela A.2 – Tópicos 4-7 extraídos da base de artigos.	43
Tabela A.3 – Tópicos 8-11 extraídos da base de artigos.	43
Tabela A.4 – Tópicos 12-15 extraídos da base de artigos.	44
Tabela A.5 – Tópicos 16-19 extraídos da base de artigos.	44
Tabela A.6 – Tópicos 20-23 extraídos da base de artigos.	44
Tabela A.7 – Tópicos 24-27 extraídos da base de artigos.	45
Tabela A.8 – Tópicos 28-31 extraídos da base de artigos.	45
Tabela A.9 – Tópicos 32-35 extraídos da base de artigos.	45
Tabela A.10–Tópicos 36-39 extraídos da base de artigos.	46
Tabela A.11–Tópicos 40-43 extraídos da base de artigos.	46
Tabela A.12–Tópicos 44-47 extraídos da base de artigos.	46
Tabela A.13–Tópicos 48-49 extraídos da base de artigos.	47

Lista de Abreviaturas e Siglas

EATI	Encontro Anual de Tecnologia da Informação
LDA	Latent Dirichlet Allocation
CPU	Unidade Central de Processamento
IDE	Integrated Development Environment
NLTK	Natural Language Toolkit
URL	Uniform Resource Locator
PDF	Portable Document Format
CSV	Comma-Separated Values
TXT	Text File

Sumário

1	Introdução	1
1.1	Justificativa	1
1.2	Objetivos	3
1.2.1	Principal	3
1.2.2	Específicos	3
1.3	Organização do Trabalho	4
2	Fundamentação Teórica	5
2.1	Comunicação Científica	5
2.2	Encontro Nacional de Tecnologia da Informação	5
2.3	Aprendizado de Máquina	7
2.4	Mineração de Textos	7
2.5	Modelagem de Tópicos	7
2.5.1	Tópicos	8
2.5.2	LDA	9
2.5.3	Gensim	11
2.5.4	LDAVis	11
2.5.5	Coerência C_v	11
3	Trabalhos Relacionados	13
3.1	Identificação automática de áreas de pesquisa em Ciência e Tecnologia	13
3.2	Modelagem de Tópicos: Resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina	13
3.3	<i>Big Data</i> no contexto de dados acadêmicos: O uso de <i>machine learning</i> na construção de sistema de organização do conhecimento	14
4	Detalhamento da Metodologia	15
4.1	Configuração do Ambiente	15
4.2	Coleta de Dados	15
4.3	Pré-Processamento	18
4.4	Modelagem de Tópicos	20
4.4.1	Corpus de Dados Como Representação Matricial	20
4.4.2	Treinamento do Modelo	21
4.4.3	Extração dos Tópicos	22
5	Resultados	23
5.1	Tópicos Extraídos	23
5.2	Sobreposição e Disjunção dos Tópicos	28
5.3	Análise dos Resultados	29
6	Considerações Finais	34

6.1	Conclusão	34
6.2	Trabalhos Futuros	34
7	Cronograma	36
 Referências		38
 Apêndices		41
APÊNDICE A Tópicos Extraídos		42

1 Introdução

A modelagem de tópicos apresenta métodos que visam resolver problemas de mineração de dados e identificação dos principais assuntos em uma coleção de documentos. Utilizada para estruturar e/ou sumarizar grandes quantidades de dados, tal metodologia fornece uma solução para o problema de gerenciar grandes arquivos de documentos (BLEI, 2012).

Estes modelos podem ser considerados através de algoritmos de *Machine Learning* não supervisionados, em que são capazes de inferir automaticamente padrões em conjuntos de dados, a partir de suas estruturas semânticas latentes e de um conjunto de pressupostos.

O arcabouço teórico proposto por (STEYVERS; GRIFFITHS, 2007) exprime a ideia de que nos modelos de tópicos, os documentos se formam através de uma combinação de tópicos, ou ainda, assuntos, os quais são uma distribuição de probabilidade de palavras sobre tais tópicos. Com isso, podemos descobrir os temas dos documentos e suas relações diante da aplicação em uma coleção de documentos.

Atualmente, os modelos de extração de tópicos com abordagem probabilística que utilizam o LDA (*Latent Dirichlet Allocation*), apresentados por (BLEI, 2012), representam uma metodologia amplamente utilizada diante da aplicação em documentos grandes, ou com grandes quantidades de palavras.

Sendo assim, a modelagem de tópicos representa uma área do conhecimento ampla e interdisciplinar para a extração de tópicos. O presente trabalho tem como propósito o LDA (*Latent Dirichlet Allocation*), para realizar uma análise exploratória sobre os tópicos descobertos nos artigos científicos apresentados nas edições do Encontro Anual de Tecnologia da Informação (EATI).

1.1 Justificativa

Diante do aumento da presença dos sistemas de informação e comunicação no âmbito da atividade humana, vemos que também é crescente o volume de informações disponibilizadas e armazenadas no ciberespaço (SOUZA, 2006).

A concepção de uma sociedade sem esta acentuada relação das tecnologias de informação que nela surgem e que a modifica, se tornou algo pouco natural nos dias atuais. Esta virtualização trouxe a tona a migração dos artefatos produzidos, principalmente na comunicação, para um ambiente online, mais acessível, aberto e informatizado (WEITZEL, 2006).

Dentre os sistemas de comunicação, 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 (VIEIRA et al., 2013).

A comunicação científica teve como principal marco de sua constituição estrutural a revista científica escrita, composta por artigos científicos e outros elementos (WEITZEL, 2006). No entanto, devido ao pleno estágio de reorganização dos processos e produtos da comunicação e informação, os produtos da comunicação sucederam sua consolidação no ambiente virtual, em formatos digitais.

Dessa forma, as revistas científicas passaram a atuar em conjunto com outros itens de comunicação científica, especialmente os repositórios institucionais online, que também estão presentes no processo que move o ciclo produtivo do desenvolvimento da ciência e da tecnologia (WEITZEL, 2006).

As instituições de ensino superior, públicas ou privadas, através dos cursos de graduação, nas modalidades tecnológico, licenciatura ou bacharelado e pós-graduação lato ou stricto sensu têm contribuído diretamente para o crescente volume de produção e disseminação de informações, de modo que os artefatos gerados, abstraídos em arquivos, são alocados nos repositórios online, formando enormes coleções de documentos eletrônicos, mantidos pelas referidas entidades.

O Encontro Anual de Tecnologia da Informação (EATI) é um evento promovido conjuntamente pela Universidade Federal de Santa Maria e o Instituto Federal Farroupilha que ocorre desde o ano de 2010. A organização do evento se constitui de palestras e minicursos ministrados por profissionais de renome regional e nacional, com conteúdos teóricos e práticos relativos à área de Tecnologia da Informação e espaços para o compartilhamento de estudos em andamento, resultados de pesquisas científicas ou mesmo experiências vivenciadas por estudantes e profissionais (EATI, 2020).

Ao longo das edições do evento, foram apresentados 310 documentos no formato de artigos longos e curtos. Esse acumulado de trabalhos constitui um leque amplo de informações, ao passo que, a Tecnologia da Informação é considerada uma área interdisciplinar.

Resumir e organizar grandes coleções de informações pode se tornar uma tarefa humanamente impossível, suscetível ao erro e exaustiva quando realizada manualmente (BLEI, 2012). Diante deste cenário, com um volume de documentos científicos cada vez maior, temos a necessidade de estudar o uso de novas técnicas ou ferramentas que lidam com a organização, pesquisa e indexação em grandes coleções de documentos de modo automático.

Tendo isso em vista, este trabalho tem como objetivo principal desenvolver uma análise exploratória, a partir da extração dos principais tópicos contidos nos artigos apresentados durante as edições do EATI, utilizando a técnica de extração *LDA*. Além disso, busca analisar a variação dos tópicos mais recorrentes de modo cronológico, ao longo de cada edição do evento; e apresentar o comportamento dos termos mais frequentes.

Para realizar a análise, pretende-se extrair e preparar os dados dos anais das edições do

EATI, removendo fragmentos e caracteres que venham a ser inúteis no problema em questão, para então realizar a modelagem com o LDA e a extração dos tópicos.

1.2 Objetivos

1.2.1 Principal

Desenvolver uma análise exploratória, a partir da extração dos principais tópicos contidos nos artigos apresentados durante as edições do EATI utilizando a técnica de extração LDA.

1.2.2 Específicos

- 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.

1.3 Organização do Trabalho

O trabalho está organizado da seguinte maneira: o capítulo 2 apresenta uma fundamentação teórica para contextualizar as abordagens deste projeto. O capítulo 3 aborda os trabalhos relacionados ao tema. No capítulo 4 temos o detalhamento da metodologia. Em seguida, o capítulo 5 expõe os resultados obtidos neste projeto. O capítulo 6 contém as considerações finais e na sequência o capítulo 7 apresenta o cronograma. Ao final, estão organizadas as referências bibliográficas.

Abaixo temos a estrutura da monografia:

Capítulo 1: Introdução.

Capítulo 2: Fundamentação Teórica.

Capítulo 3: Trabalhos Relacionados.

Capítulo 4: Detalhamento da Metodologia.

Capítulo 5: Resultados.

Capítulo 6: Considerações Finais.

Capítulo 7: Cronograma.

2 Fundamentação Teórica

Neste capítulo, uma fundamentação dos principais conceitos relacionados à modelagem de tópicos e mineração de documentos serão apresentados, para contextualizar algumas técnicas e abordagens adotadas neste projeto.

2.1 Comunicação Científica

No âmbito da pesquisa científica, a comunicação científica pode ocorrer a partir da disseminação de uma série de publicações, produzidas em diversos formatos. Isso pode acontecer durante o período de construção da pesquisa e pode se estender para além do encerramento da mesma. Dentre os exemplos de produção científica, temos: palestras, relatórios, congressos, artigos de periódicos, livros impressos e documentos no formato digital (ZIMAN, 1971).

O processo de publicação da pesquisa ocorre diante da exposição do estudo realizado pelo autor ao julgamento realizado por pares, de forma que seja possível alcançar o consenso, confiabilidade e validação da pesquisa científica. A divulgação da informação científica possibilita tornar pública a produção do conhecimento (ZIMAN, 1971).

Os artefatos produzidos nos canais formais de comunicação (testes, dissertações, artigos, resumos, livros entre outros) podem ser agrupados e armazenados no formato digital (ZIMAN, 1971). Um dos maiores problemas nesse cenário é o fato de que estas coleções de documentos tendem a ficar cada vez maiores. Com isso, a dificuldade de recuperar e organizar estes documentos se torna muito alta.

Os eventos científicos acabam por agradar os pesquisadores e a comunidade científica como um todo, pois eles permitem a exposição e a discussão de suas pesquisas de forma que possam ser avaliadas por outros pesquisadores. A realização de encontros científicos pode ocorrer de diversas formas, como, colóquios, encontros, fóruns, reuniões, seminários e simpósios (ZIMAN, 1971). Diante da efetivação de um evento científico, os documentos são publicados, na maioria das vezes, nos repositórios sob a forma de anais do evento.

2.2 Encontro Nacional de Tecnologia da Informação

O Encontro Anual de Tecnologia da Informação - EATI é um evento promovido conjuntamente pela Universidade Federal de Santa Maria (Campus de Frederico Westphalen) e o Instituto Federal Farroupilha (Campus Frederico Westphalen), ambos localizados na região sul do Brasil. Realizado anualmente desde 2010, caracteriza-se por proporcionar um momento de encontro de estudantes, profissionais e pesquisadores da área de Tecnologia da Informação, constituindo-se

como um espaço de integração, interlocução e interdisciplinaridade.

O evento foi pensado para reunir o público entusiasta da Tecnologia da Informação, proporcionando um conjunto de atividades técnico/científicas que visam, não apenas o debate sobre os temas atuais, mas também treinamentos específicos. Dessa forma, há um âmbito por parte dos organizadores em promover a integração dos alunos de diferentes instituições, o acesso ao conhecimento científico e inovações tecnológicas.

Na realização do EATI, ocorrem palestras e minicursos ministrados por profissionais de renome regional e nacional, com conteúdos relativos à área de Tecnologia da Informação e espaços para o compartilhamento de estudos em andamento e resultados de pesquisas científicas.

O evento tem como objetivo principal levar conhecimento, informar e debater temas relevantes da área de informática que estejam em evidência no país e no exterior para o enriquecimento acadêmico e profissional dos participantes envolvidos.

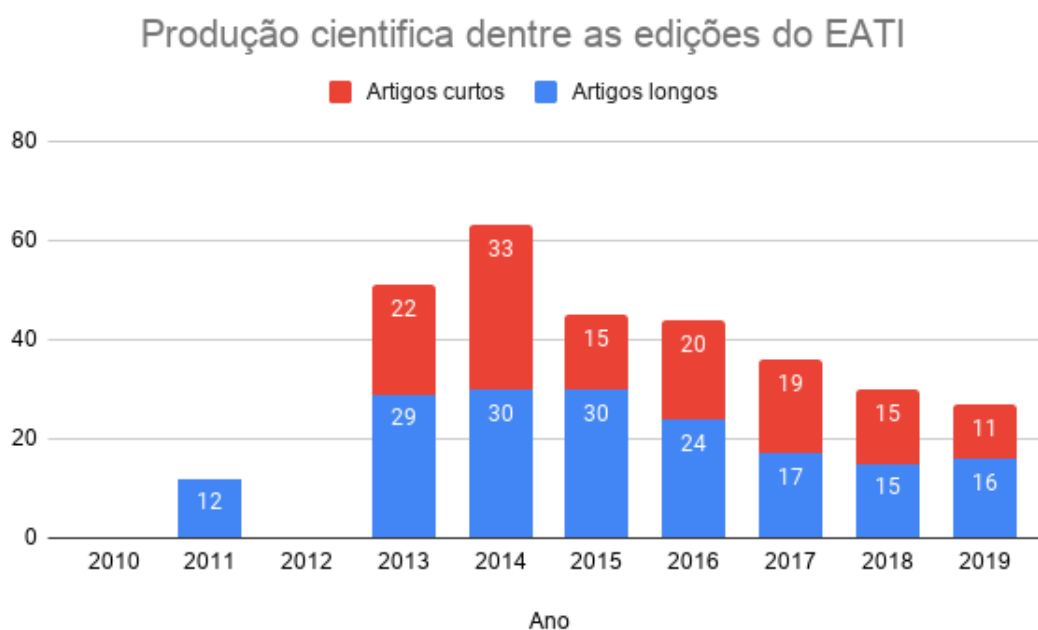


Figura 2.1 – Produção científica dentre as edições do EATI.

Entre a 1ª e a 10ª edição do EATI, foram apresentados 310 produções científicas nos anais dos eventos, sendo 135 artigos longos e 175 artigos curtos. A Figura 2.1 apresenta a quantificação das publicações, separadas por tipo de artigo e pelo ano em que o evento foi realizado.

Nas edições de 2010 e 2012, os artigos apresentados não foram disponibilizados pela organização do evento. O ano de 2012 teve o menor número de trabalhos apresentados, ao passo que na 5ª edição em 2014 a quantidade de produções alcançou seu valor máximo.

2.3 Aprendizado de Máquina

A teoria de Aprendizado de Máquina ou *Machine Learning* é baseada nos princípios de Inteligência Artificial, e possui como objetivo realizar a aprendizagem de sistemas computacionais a partir de algoritmos que aprendem interativamente, por meio de processo repetitivo diante dos dados fornecidos (BRUNIALTI et al., 2015).

Esse processamento realizado a partir dos algoritmos permite a extração de modelos capazes de explicar ou representar os dados sob algum aspecto, descobrir informações ocultas nos dados etc.

Na modalidade não supervisionada de *Machine Learning*, os parâmetros de um modelo são ajustados com base na maximização de medidas de qualidade das respostas obtidas (BRUNIALTI et al., 2015). Além disso, nesta categoria, os dados da base não possuem rótulos ou resultados conhecidos previamente.

2.4 Mineração de Textos

A mineração de texto pode ser definida como uma forma de ajudar usuários a analisar e obter informações de grandes conjuntos de textos, para facilitar o processo de decisão. Seu principal objetivo é inferir padrões interessantes, incluindo tendências em dados textuais (FONSECA, 2018).

Um dos desafios existentes na mineração de dados está atrelado na tratativa de textos esparsos e de alta dimensionalidade. Essas características dos textos são abstraídas através do *corpus* relativo aos documentos. O *corpus* é constituído por uma coleção de textos ou documentos produzidos pelo homem em um ambiente natural de comunicação. Assim, o *corpus* pode ser representado como uma matriz termo-documento de tamanho $n \times d$, na qual n é o número de documentos e d é o tamanho do dicionário. A entrada (i,j) da matriz é a frequência normalizada da j -ésima palavra no documento i .

A mineração de texto utiliza várias técnicas avançadas de mineração de dados, aprendizado de máquina, recuperação de informação, extração de informação e processamento de linguagem natural (FALEIROS, 2016).

2.5 Modelagem de Tópicos

Com a explosão no volume de arquivos de documentos eletrônicos atualmente, organizar, resumir e recuperar as informações contidas nestes documentos se tornou uma tarefa extremamente onerosa. Dessa forma, pesquisadores de aprendizado de máquina desenvolveram a modelagem probabilística de tópicos.

Consideraremos um documento como toda a informação registrada em um suporte material e fonte de informação. Pela teoria proposta por (BLEI, 2012), os modelos de tópicos visam descobrir padrões latentes (ocultos) em documentos, que além de sua aplicação em textos, podem ser utilizados em outros tipos de dados.

Dado um texto sobre um determinado tópico, aparecerão palavras relacionadas ao tema com certa frequência em que serão rotulados como sendo os tópicos descobertos pelo algoritmo de modelagem de tópicos. A Figura 2.2 ilustra a mistura e distribuição dos tópicos em um documento, onde os termos relacionados a um tópico estão destacados em cores iguais.

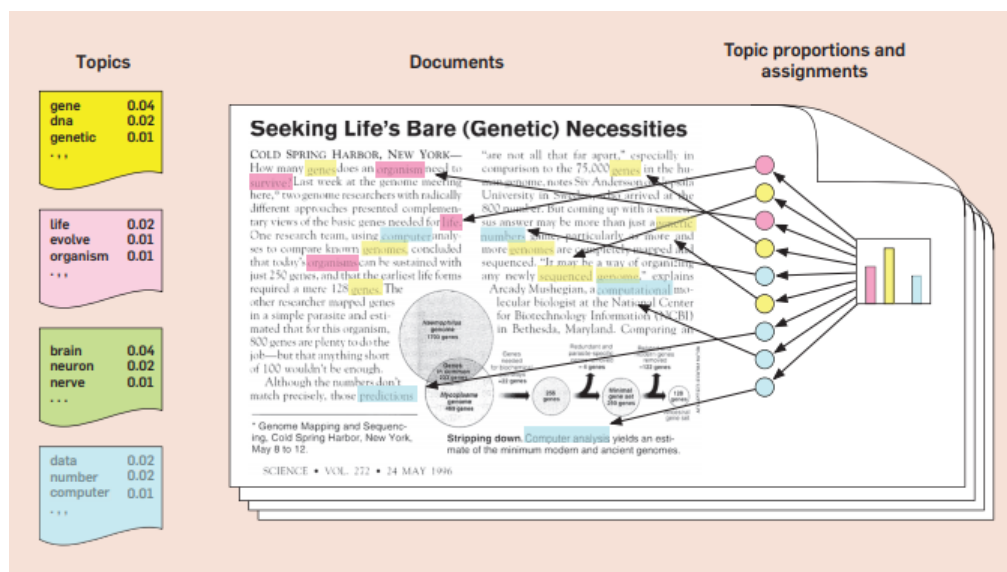


Figura 2.2 – Distribuição dos tópicos em um documento. Adaptado de (BLEI, 2012).

Isso posto, temos que a modelagem de tópicos é um conjunto de algoritmos que foram desenvolvidos com o objetivo de obter informações em grandes arquivos de texto. Tais informações permeiam as palavras dos textos originais, visando descobrir os temas abordados e suas relações.

2.5.1 Tópicos

Pelo princípio apresentado por (BLEI, 2012), pode-se definir tópico como uma distribuição sobre um vocabulário fixado. Esta estrutura é formada por distribuições probabilísticas de palavras a respeito de um determinado documento, ao passo que cada tópico é um padrão recorrente de co-ocorrência de palavras.

Exemplificando, um tópico sobre esportes possui uma maior probabilidade de apresentar palavras relativas à esportes, da mesma forma que em um tópico sobre cinema, há maior chance de aparecerem palavras ligadas com cinema.

Na Figura 2.3 temos um exemplo de tópicos extraídos a partir de uma base de dados. Nota-se na disposição das palavras de cada tópico que existe uma relação entre elas diante daquele assunto.

"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Figura 2.3 – Exemplo de tópicos gerados. Adaptado de (BLEI, 2012).

Nos dias de hoje, diversas abordagens são utilizadas na realização da modelagem de tópicos. Dentre as tratativas, destaca-se o *Latent Dirichlet Allocation (LDA)* devido a sua capacidade de produzir tópicos interpretáveis e semanticamente coerentes.

2.5.2 LDA

O *Latent Dirichlet Allocation (LDA)* é um modelo probabilístico generativo de um corpus. Nesta metodologia, a ideia central é que documentos são representados como misturas randômicas sobre tópicos e cada tópico é caracterizado por uma distribuição sobre palavras. Além disso, é escolhido um vocabulário básico de “palavras” ou “termos” e, para cada documento do *corpus*, uma contagem do número de ocorrências de cada palavra é realizada.

Diante disso, na abordagem do LDA, tem-se que existe um certo número de tópicos, que são distribuições de palavras, para toda a coleção. E também, presume-se que inicialmente cada documento é gerado a partir da escolha de uma distribuição sobre tópicos. Na sequência, para cada palavra, é escolhida uma distribuição de tópicos e definida a palavra para o tópico correspondente (BLEI; NG; JORDAN, 2003).

Neste modelo, consideramos os documentos como *bag of words*, onde cada documento é representado como um vetor das palavras que ocorrem no documento, ou em representações mais sofisticadas como frases ou sentenças. Dito isso, assume-se que o número de tópicos é conhecido e será constante.

Supondo a estrutura Tópicos como uma coleção de tópicos, Documentos como uma coleção dos documentos da base de dados e que os documentos são coleções de palavras, e conforme (BLEI; NG; JORDAN, 2003), a modelagem pelo LDA pode ser descrita conforme o pseudocódigo da Figura 2.4.

Algorithm 1 Pseudocódigo do modelo generativo para LDA

```

1: for tópico  $k = 1, 2, \dots, K$  do
2:   aplica a priori da distribuição de Dirichlet  $\text{Dir}(\beta)$  no tópico  $k$ 
3: end for
4: for documento  $m = 1, 2, \dots, M$  do
5:   aplica a priori da distribuição de Dirichlet  $\text{Dir}(\alpha)$  no documento  $m$ 
6:   mistura os documentos com a distribuição  $\text{Poiss}(\xi)$  e define  $N_m$ 
7:   for palavra  $n = 1, 2, \dots, N_m$  do
8:     mistura os índices  $z_{m,n}$  dos tópicos com distribuição  $\text{Mult}(\theta_m)$ 
9:     aplica a distribuição  $\text{Mult}(\phi_k)$  dos tópicos internamente para as
       palavras  $w_{m,n}$  do documento
10:   end for
11: end for

```

Figura 2.4 – Pseudocódigo do algoritmo LDA.

Presume-se que um conjunto de K tópicos já seja conhecido e corrigido. Um documento W_m é gerado escolhendo primeiro uma distribuição sobre tópicos θ_m de uma distribuição *Dirichlet* $\text{Dir}(\alpha)$, que determina a atribuição de tópico para palavras naquele documento. Então, a atribuição de tópico para cada índice de palavra $[m, n]$ é realizada amostrando um tópico particular $z_{m,n}$ de uma distribuição multinomial $\text{Mult}(\theta_m)$. E, finalmente, uma palavra particular $w_{m,n}$ é gerada para o índice $[m, n]$ por amostragem da distribuição multinomial $\text{Mult}(\phi_k)$.

No algoritmo apresentado na Figura 2.4, o primeiro laço se encarrega de realizar a distribuição de tópicos em todos os documentos, o segundo laço distribui os tópicos para cada documento, ao passo que o terceiro laço, interno ao segundo, repete a distribuição de tópicos para as palavras de um documento, sendo assim o responsável por realizar a mistura dos tópicos e modelar as probabilidades de que um conjunto de palavras pertença ao mesmo tópico.

A Figura 2.5 possui a ilustração deste processo, mostrando os níveis de tópicos, documentos e palavras por documento.

Podemos dizer que a complexidade computacional do algoritmo apresentado na Figura 2.4 é $O(mn)$, para m como o número total de documentos e n sendo o número de palavras do corpus. Por fim, estes processos iterativos ocorrem até que um critério de convergência (ou número de iterações) seja atingido.

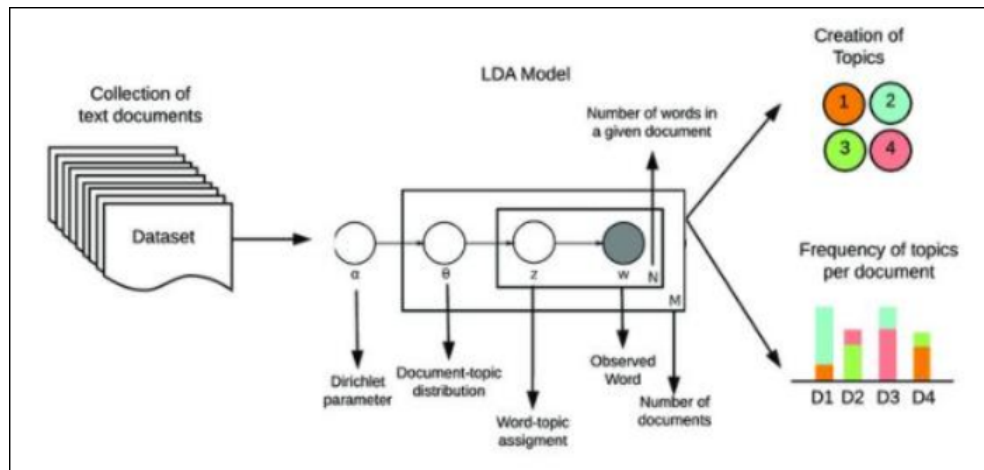


Figura 2.5 – Esquema ilustrado do algoritmo LDA.

2.5.3 Gensim

O Gensim é um conjunto de módulos *Open Source*, desenvolvido em *Python*. Ele serve para a modelagem de espaço vetorial e modelagem de tópicos, ao extrair automaticamente os tópicos com relação semântica (REHUREK; SOJKA, 2010). A biblioteca Gensim dispõe da modelagem de tópicos em duas abordagens, *LDAModel* e *LDAMulticore*, sendo que o *LDAMulticore* utiliza mais de um núcleo da CPU para paralelizar e acelerar o treinamento do modelo.

2.5.4 LDAVis

O LDAVis é uma biblioteca em *Python* destinada a visualização de modelos de tópicos de forma interativa em páginas Web. Através dos métodos desta biblioteca, é permitido aos usuários visualizar os tópicos de uma forma mais global, de modo que também é possível realizar uma inspeção mais profunda dos termos associados a cada tópico de maneira individual (SIEVERT; SHIRLEY, 2014).

Utilizando um modelo de tópicos LDA, o LDAVis pode construir um gráfico interativo que apresenta a distância entre os tópicos, colocando-os como círculos em um plano bidimensional, e em seguida, usando escala multidimensional para projetar as distâncias entre os tópicos em duas dimensões (SIEVERT; SHIRLEY, 2014).

2.5.5 Coerência C_v

Métricas são instrumentos destinados para medir e avaliar resultados. Na modelagem de tópicos existem dificuldades em relação a avaliar resultados pois os conjuntos de dados não possuem rótulos para aferir a coerência alcançada.

A métrica de Coerência C_v serve para calcular a co-ocorrência de cada palavra do tópico com todas as palavras deste tópico, avaliando a afinidade entre os tópicos. O resultado da aplicação

da métrica é um conjunto de vetores, um para cada palavra.

Nesta métrica, utiliza-se o *cosinus* como medida de similaridade, calculando o valor do cosseno entre o ângulo do vetor de cada palavra, com o vetor resultante da soma dos vetores de cada palavra do tópico. Para ângulos menores entre dois vetores, temos que o valor do cosseno se torna maior, isto é, se aproxima mais de 1. A média aritmética desses valores de similaridade resulta na coerência.

3 Trabalhos Relacionados

A extração de tópicos aplicada em coleções de documentos pode ser encontrada em diversos trabalhos na literatura. Este estudo visa a análise de artigos científicos, especialmente da área de Tecnologia da Informação e domínios correlatos, de modo que foram selecionados estudos similares, com a aplicação de modelagem de tópicos sobre textos científicos relacionados a Tecnologias da Informação.

3.1 Identificação automática de áreas de pesquisa em Ciência e Tecnologia

O trabalho de (NOLASCO, 2016) apresentou uma técnica integrada para a identificação automática de áreas de pesquisa presentes em uma coleção de documentos, e sua posterior representação através de rótulos para facilitar a compreensão dos respectivos conteúdos. Na abordagem utilizada pelos autores foi elaborada uma proposta composta de três partes: Agrupamento, Seleção de Grupos e Representação dos Grupos.

A base de dados aplicada para desenvolver este trabalho foi composta de coleções de documentos científicos apresentados em eventos e conferências como: *Knowledge Discovery and Data Mining* (KDD), entre os anos 2004 e 2014; *Scholar Data Challenge* (SDC); *Special Interest Group on Information Retrieval* (SIGIR); Simpósio Brasileiro de Banco de Dados (SBBD).

Na fase de agrupamento, os principais tópicos da coleção foram extraídos com a técnica LDA, ao passo que tais tópicos permitiram que os documentos fossem agrupados segundo seus respectivos conjuntos ao qual cada um pertencia.

Nos resultados obtidos, (NOLASCO, 2016) realizou a identificação do número de áreas presentes em uma coleção e rotulagem de grupos de documentos. Dentre cada uma destas áreas, foram selecionadas as melhores técnicas de maneira que elas foram adaptadas ao ambiente científico e às necessidades da proposta.

3.2 Modelagem de Tópicos: Resumir e organizar corpus de dados por meio de algoritmos de aprendizagem de máquina

No estudo realizado por (SOUZA; SOUZA, 2018) foi feita uma comparação dos resultados e do desempenho de duas técnicas de modelagem de tópicos baseadas na aprendizagem de máquina: *Latent Semantic Indexing* (LSI) e *Latent Dirichlet Allocation* (LDA).

A modelagem foi aplicada em uma base de dados constituída por 2006 artigos científicos e resumos expandidos do XIII ao XVII Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB). As etapas da metodologia do trabalho desenvolvido foram segmentadas na pesquisa empírica da coleta dos dados para constituição, limpeza, manipulação, combinação, normalização, tratamento e transformação dos dados do corpus para conectar aos modelos de aprendizagem de máquina abordados.

Como resultados, os modelos resumiram e organizaram o corpus de dados em tópicos 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.

3.3 *Big Data* no contexto de dados acadêmicos: O uso de *machine learning* na construção de sistema de organização do conhecimento

No trabalho desenvolvido por (NAVARRO; CONEGLIAN; SEGUNDO, 2018), foi proposto um modelo teórico, que vinculou os dados de repositórios digitais, com as técnicas de *machine learning*, para a construção de sistemas de organização do conhecimento.

Para validar esse modelo, realizou-se uma prova de conceito, que utilizou a técnica LDA para identificar tópicos de interesse dentro de um corpus baseado em um repositório digital contendo aproximadamente 255 documentos como base de dados, ao passo que a maioria dos documentos desse repositório consistia de trabalhos de conclusão de curso da Universidade Estadual Paulista (Unesp).

Por fim, os autores concluíram através da prova de conceito que a aplicação das técnicas de *machine learning* contribui para fornecer um panorama amplo sobre os dados contidos em repositórios digitais.

4 Detalhamento da Metodologia

O objetivo deste trabalho é, conforme 1.2, realizar a extração dos principais tópicos contidos nos artigos científicos apresentados entre 2012 e 2019 nas edições do EATI. Dessa forma, dividimos a metodologia de desenvolvimento em quatro etapas: Coleta da Base de Dados, Pré-Processamento, Modelagem dos Tópicos com LDA e Análise Exploratória dos Dados.

A linguagem escolhida para o desenvolvimento foi *Python*¹, uma linguagem de programação interpretada, de alto nível e que pode ser utilizada para diversos fins.

As bibliotecas *Python* mais utilizadas no desenvolvimento foram: *Pandas*², *Requests*³, *Selenium* e *Apache Tika*⁴ na coleta de dados; *NLTK*⁵ (*Natural Language Toolkit*) e *ReGex*⁶ no pré-processamento; *Gensim*⁷ na extração e modelagem dos tópicos; *Matplot*⁸ e *LDAVis*⁹ para apresentação dos resultados da análise.

4.1 Configuração do Ambiente

O ambiente para geração da coleção de artigos, pré-processamento dos dados, treinamento do modelo de tópicos, testes do modelo para extração dos tópicos e análise dos tópicos foi baseado em um *desktop* com processador Intel(R) Core(TM) i5-8400 CPU @ 2.80GHz, 6 Núcleos, 6 Processadores Lógicos, 8GB de memória RAM e disco de 1TB com sistema operacional *Windows 10 Pro* 64-bit.

Em relação ao desenvolvimento de modo mais detalhado, foram utilizados para estes procedimentos o ambiente computacional *Web Jupyter Notebook*¹⁰ e a IDE (*Integrated Development Environment*) *PyCharm*¹¹.

4.2 Coleta de Dados

Esta etapa elucida a criação da base de dados utilizada neste trabalho. A base de dados foi composta por 310 documentos formados por artigos longos e artigos curtos publicados entre

¹ <<https://www.python.org/>>

² <<https://pandas.pydata.org/docs/>>

³ <<https://pypi.org/project/requests/>>

⁴ <<https://pypi.org/project/tika/>>

⁵ <<https://www.nltk.org/>>

⁶ <<https://docs.python.org/pt-br/3.8/howto/regex.html>>

⁷ <https://radimrehurek.com/gensim/auto_examples/index.html>

⁸ <<https://matplotlib.org/>>

⁹ <<https://pyldavis.readthedocs.io/en/latest/index.html>>

¹⁰ <<https://jupyter.org/>>

¹¹ <<https://www.jetbrains.com/pt-br/pycharm/>>

os anos de 2012 e 2019 nos anais de cada edição do EATI.

Em todas as edições, a organização do evento mantém a seção dos anais de forma padronizada, sempre apresentando os artigos em uma tabela, com as colunas de “Título do artigo” e “Autor(es)”. Cada elemento da página web que representa um artigo possui uma URL atrelada ao *link* do artigo em formato *PDF* e o título do artigo relacionado ao texto do mesmo elemento.



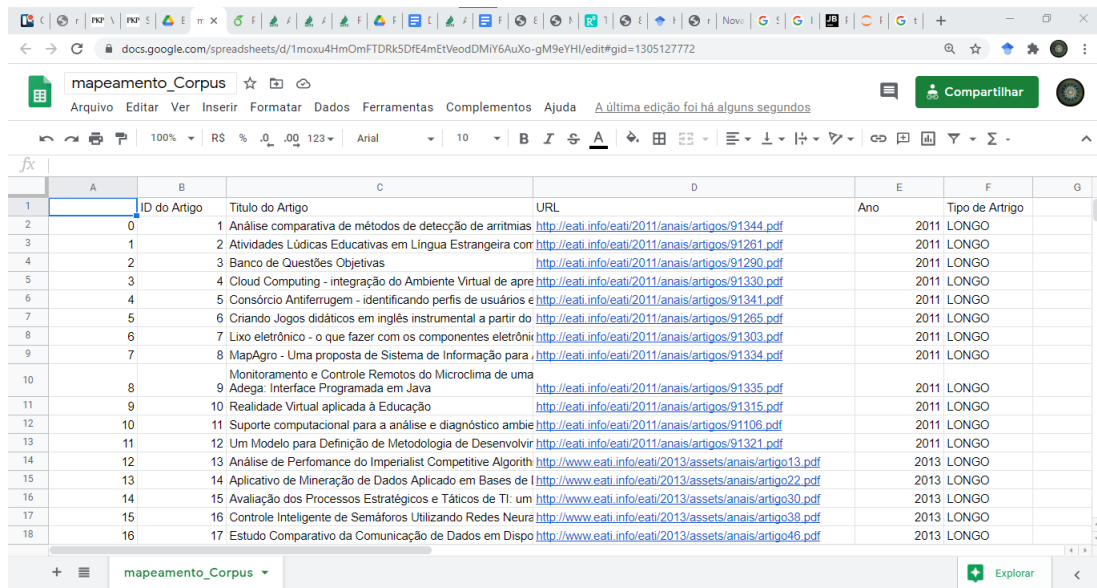
Título do Artigo	Autor(s)
ForceQoS: Ferramenta Open Source para Reconfiguração e Controle Eficiente de QoS	Cassiano Monego Mateus Victorio Zaganel Cristian Cleder Machado
Implementação de um Cluster Beowulf utilizando o framework Warewulf	Ludivan Bento Brandão João Paulo de Brito Gonçalves Paulo José Pereira de Oliveira
Elaboração de Grades Horárias Utilizando Algoritmos Genéticos	Lucas Bucior Fabio Asturian Zanin Marcos A. Lucas
Identificar o perfil dos estudantes do ensino médio para desenvolver pensamento computacional por meio do Scratch	Jonathan Pippi

Figura 4.1 – Página dos anais do evento da edição de 2018 do EATI (EATI, 2020).

A coleta dos documentos foi realizada inicialmente a partir da extração das *URLs* de cada artigo hospedado nas seções dos “Anais do Evento”, contidas no site de cada edição do EATI. Este processo foi concebido com o *Selenium*, uma ferramenta que automatiza o acesso e a interação a uma página Web e seus elementos, como botões, caixas de texto, links, e etc em um navegador específico (HUGGINS, 2013). Sendo assim, foi possível automatizar a realização de ações como obtenção do texto de um elemento e obtenção da URL de um link nos elementos da página.

O *script*¹² da automação da coleta de dados com *Selenium* foi desenvolvido para acessar a seção “Anais do Evento” no site do EATI, obter a URL dos elementos vinculados a cada artigo presente na página, e o Título do artigo. Feito isso, essas informações coletadas foram salvas em um arquivo do tipo CSV, onde adicionamos um número de identificação em sequencia para cada artigo mapeado e também o tipo do artigo, variando entre “Longo” e “Curto” conforme a Figura 4.2.

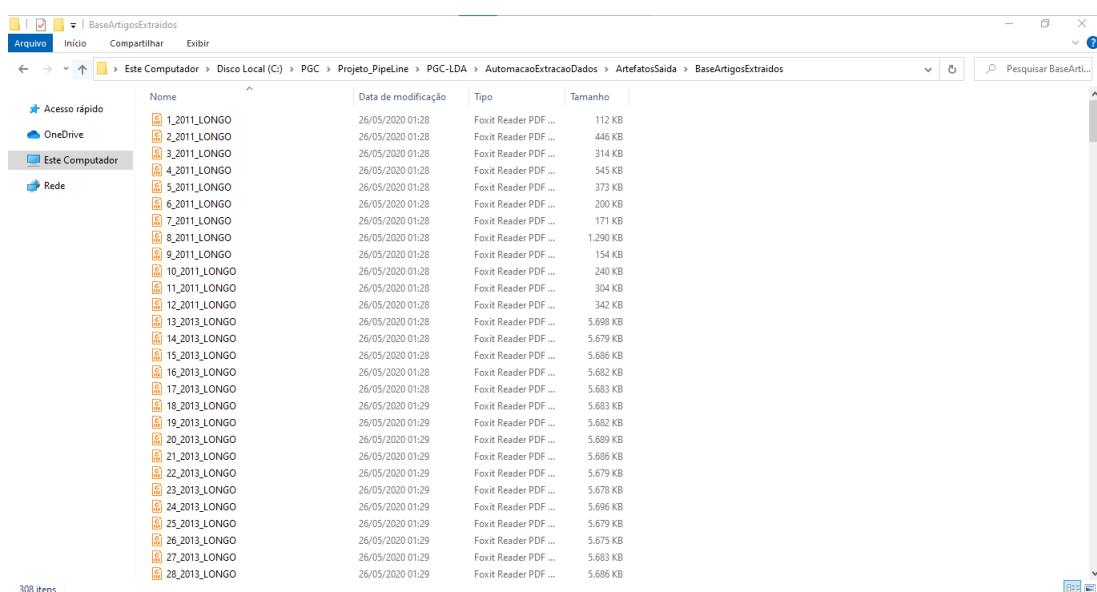
¹² Código para extrair as URLs dos artigos e mapear os dados dos artigos extraídos em uma planilha CSV. Disponível em <https://github.com/o-eduardo/PGC-LDA/blob/master/AutomacaoExtracaoDados/testes/automacao_extracao_dados.py>. Acesso em: 14 abr. 2021.



	A	B	C	D	E	F	G
1		ID do Artigo	Título do Artigo	URL	Ano	Tipo de Artigo	
2	0	1	Análise comparativa de métodos de detecção de arritmias	http://eati.info/eati/2011/anaais/artigos/91344.pdf	2011	LONGO	
3	1	2	Atividades Lúdicas Educativas em Língua Estrangeira com	http://eati.info/eati/2011/anaais/artigos/91261.pdf	2011	LONGO	
4	2	3	Banco de Questões Objetivas	http://eati.info/eati/2011/anaais/artigos/91290.pdf	2011	LONGO	
5	3	4	Cloud Computing - integração do Ambiente Virtual de apre	http://eati.info/eati/2011/anaais/artigos/91330.pdf	2011	LONGO	
6	4	5	Consórcio Antiferrugem - identificando perfis de usuários e	http://eati.info/eati/2011/anaais/artigos/91341.pdf	2011	LONGO	
7	5	6	Criando Jogos didáticos em inglês instrumental a partir do	http://eati.info/eati/2011/anaais/artigos/91265.pdf	2011	LONGO	
8	6	7	Lixo eletrônico - o que fazer com os componentes eletrôni	http://eati.info/eati/2011/anaais/artigos/91303.pdf	2011	LONGO	
9	7	8	MapAgro - Uma proposta de Sistema de Informação para	http://eati.info/eati/2011/anaais/artigos/91334.pdf	2011	LONGO	
10	8	9	Monitoramento e Controle Remotos do Microclima de uma	http://eati.info/eati/2011/anaais/artigos/91335.pdf	2011	LONGO	
11	9	10	Adega: Interface Programada em Java	http://eati.info/eati/2011/anaais/artigos/91315.pdf	2011	LONGO	
12	10	11	Realidade Virtual aplicada à Educação	http://eati.info/eati/2011/anaais/artigos/91315.pdf	2011	LONGO	
13	11	12	Um Modelo para Definição de Metodologia de Desenvolvir	http://eati.info/eati/2011/anaais/artigos/91321.pdf	2011	LONGO	
14	12	13	Análise de Performance do Imperialist Competitive Algorith	http://www.eati.info/eati/2013/assets/anaais/artigo13.pdf	2013	LONGO	
15	13	14	Aplicativo de Mineração de Dados Aplicado em Bases de	http://www.eati.info/eati/2013/assets/anaais/artigo22.pdf	2013	LONGO	
16	14	15	Avaliação dos Processos Estratégicos e Táticos de TI: um	http://www.eati.info/eati/2013/assets/anaais/artigo30.pdf	2013	LONGO	
17	15	16	Controle Inteligente de Semáforos Utilizando Redes Neura	http://www.eati.info/eati/2013/assets/anaais/artigo38.pdf	2013	LONGO	
18	16	17	Estudo Comparativo da Comunicação de Dados em Dispo	http://www.eati.info/eati/2013/assets/anaais/artigo46.pdf	2013	LONGO	

Figura 4.2 – Tabela de mapeamento das informações dos artigos da base de dados.

Seguindo o fluxo da coleta de dados, a partir das *URLs* extraídas na tarefa anterior, foi realizado o *download* de cada artigo no formato *PDF*, de modo que os 308 arquivos baixados foram alocados em um repositório e receberam uma nomenclatura com o padrão contendo: o número de identificação definido no mapeamento conforme a Figura 4.2, o ano da edição do evento em que o artigo foi apresentado e o tipo do artigo, de acordo com a Figura 4.3. A utilização deste padrão de nome facilitou a busca e recuperação dos artigos nas etapas de consolidação e pré-processamento dos dados.



Nome	Data de modificação	Tipo	Tamanho
1_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	112 KB
2_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	446 KB
3_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	314 KB
4_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	545 KB
5_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	373 KB
6_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	200 KB
7_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	171 KB
8_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	1.290 KB
9_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	154 KB
10_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	240 KB
11_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	304 KB
12_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	342 KB
13_2011_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	5.698 KB
14_2013_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	5.679 KB
15_2013_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	5.686 KB
16_2013_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	5.682 KB
17_2013_LONGO	26/05/2020 01:28	Foxit Reader PDF ...	5.683 KB
18_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.683 KB
19_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.682 KB
20_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.689 KB
21_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.686 KB
22_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.679 KB
23_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.678 KB
24_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.696 KB
25_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.679 KB
26_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.675 KB
27_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.683 KB
28_2013_LONGO	26/05/2020 01:29	Foxit Reader PDF ...	5.686 KB

Figura 4.3 – Arquivos dos artigos extraídos em formato PDF com nomenclatura padrão.

Após esta etapa, utilizamos um algoritmo¹³ para realizar a conversão destes arquivos para o formato de texto *TXT* com a biblioteca de extração de textos de arquivos *Apache Tika* (MATT-MANN; ZITTING, 2012). Por fim, realizamos a incorporação dos artigos em um documento no formato de texto para cada ano de edição do evento, e um documento também no formato de texto incorporando o conteúdo de todos os artigos extraídos. Ao total, obtivemos 8 arquivos de texto referentes às edições dos eventos e 1 arquivo para a base completa.

Este arquivo referente a base de artigos de todos os eventos do EATI foi composto por 308 documentos, 5215896 palavras no total, 28594 palavras distintas com média de 16935 palavras por documento e tamanho de 5.11MB.

4.3 Pré-Processamento

Para a etapa de preparação dos dados, foi realizada uma limpeza e normalização dos dados com o objetivo de evitar ruídos nos resultados e para melhor desempenho e qualidade nos tópicos.

O pré-processamento foi realizado através de um algoritmo¹⁴, onde as etapas de limpeza e normalização dos dados foram:

- Remoção de cabeçalhos e rodapés presentes por padrão nos artigos;
- Remoção de endereços de e-mail dos autores dos artigos;
- Remoção de URLs;
- Remoção de termos consistidos de números;
- Remoção de pronomes oblíquos, como por exemplo "lo", "la", "lhe" entre outros;
- Remoção de pontuação e símbolos especiais;
- Substituição de siglas por seus respectivos nomes;
- Remoção de *stopwords*, i.e palavras que adicionam pouco valor à análise;
- Remoção de termos com tamanho unitário;
- Tokenização dos textos em unidades;
- Padronização das palavras em formato minúsculo.

¹³ Código para converter os arquivos dos artigos para o formato de texto. Disponível em <https://github.com/o-eduardo/PGC-LDA/blob/master/AutomacaoExtracaoDados/testes/automacao_construcao_corpus_dados.py>. Acesso em: 14 abr. 2021.

¹⁴ Código para carregar o corpus com os textos dos artigos extraídos e realizar o pré-processamento dos textos. Disponível em <https://github.com/o-eduardo/PGC-LDA/blob/master/AutomacaoExtracaoDados/testes/automacao_pre_processamento.py>. Acesso em: 14 abr. 2021.

Anais do EATI	Frederico Westphalen - RS	Ano 5 n. 1	p. 204-211	Nov/2015
Anais do EATI - Encontro Anual de Tecnologia da Informação e Semana Acadêmica de Tecnologia da Informação				205
entre os indivíduos da grade, simulando a dinâmica da epidemia, e o modelo de Malthus referente ao crescimento populacional. Utiliza-se, portanto, uma grade bidimensional				

Figura 4.4 – Exemplo de cabeçalhos e rodapés presentes nos artigos da base de dados.

Devido à natureza da base de dados e à presença da seção *Abstract* nos artigos, a remoção das *stopword* foi realizada tanto para *stopwords* da língua portuguesa quanto da inglesa. Para isso foi utilizada a biblioteca *NLTK* (LOPER; BIRD, 2002). Nos baseamos nas práticas apresentadas em (KUCHLING, 2014) para realizar a substituição das siglas e remoção de pontuação e caracteres especiais, através de expressões regulares com suporte da biblioteca *ReGex*.

O processo de remoção de cabeçalhos e rodapés presentes por padrão nos artigos foi realizado após uma análise superficial nos artigos de cada edição. Notamos que os artigos apresentavam por padrão o nome do evento, nome das instituições envolvidas ao encontro, e outros padrões textuais em cabeçalhos e rodapés, conforme a Figura 4.4. Dessa forma, após mapear estes textos realizamos a remoção e limpeza dos mesmos com base nos exemplos de utilização da biblioteca *Regex* apresentados em (KUCHLING, 2014).

Após estas etapas de limpeza e normalização, realizou-se uma incorporação dos termos ditos como *N-grams* junto à base de dados. Um *N-gram* é uma sequência de itens de uma determinada cadeia de texto, onde esta amostra textual apresenta os *N* itens de forma contígua entre si (CAVNAR; TRENKLE et al., 1994). A expressão “Inteligência Artificial” exemplifica um bigrama, ao passo que “Rio de Janeiro” representa um trigramma, e assim por diante.

Apenas os *N-grams* mais frequentes foram incorporados em cada documento dos artigos separadamente, ou seja, para cada linha do documento com todos os artigos, obtivemos tais *N-grams* e concatenamos ao final da linha. Destacamos que devido à distribuição de probabilidade aplicada na modelagem dos tópicos, a ordem das palavras em cada documento não importa.

Os principais *N-grams* extraídos foram: “lixo eletrônico”, “caso de uso”, “design participativo”, “rede neural artificial”, “árvore geradora mínima”, “ambiente virtual”, “língua portuguesa”, “aprendizagem reforço”, “tecnologia informação”, “banco dados” e “web services”.

4.4 Modelagem de Tópicos

Com o pré-processamento realizado, a próxima etapa foi a modelagem dos tópicos com o *LDA*. A modelagem foi realizada com um algoritmo¹⁵, adaptado do *script* disponibilizado por (REHUREK; SOJKA, 2010 apud SOUZA; SOUZA, 2018), destinado a: preparar os dados de entrada da maneira correta para a execução do método de obtenção do modelo LDA existente na biblioteca Gensim e realizar o treinamento do modelo,

O treinamento foi realizado aplicando o método de obtenção do modelo para diferentes hiperparâmetros de entrada e avaliando a coerência de cada modelo.

4.4.1 Corpus de Dados Como Representação Matricial

Nesta etapa, fez-se necessária a realização de condições adequadas para o treinamento do algoritmo de modelagem de tópicos. A primeira delas foi a criação de um dicionário, composto a partir das palavras tokenizadas que foram obtidas no pré-processamento.

A linguagem *Python* possui uma estrutura de dados que modela um dicionário, a estrutura de Dicionário, ou *Dictionary*. Os dicionários são indexados por **chaves**, que podem ser de qualquer tipo imutável como strings e números. Esta estrutura é composta por uma coleção de pares {**chave: valor**}, com o requisito de que as **chaves** sejam únicas (dentro de um dicionário) (KUHLMAN, 2009).

A geração deste dicionário, efetuada por um módulo da biblioteca Gensim, realizou o mapeamento entre os *tokens*, atribuindo identificadores únicos a cada diferente *token*. Por exemplo, para um documento com os tokens iguais a: {“tecnologia”, “informação”, “aplicada”}, um dicionário resultante seria neste caso: {“tecnologia”:1, “informação”:2, “aplicada”:3}.

Com o dicionário armazenado, a partir do mesmo, aplicou-se o conceito de *Bag of Words* para mapear a frequência de cada *token* nos documentos, ou seja o número de ocorrências de cada palavra em cada documento, acarretando na criação do *corpus*.

O corpus é a representação matricial que relaciona os documentos e os termos que compõem cada documento da base de dados, resultando numa visão de frequência de termos por documento. A Figura 4.1 é um exemplo da representação da matriz documento-termo para uma coleção de quatro documentos com três palavras com frequências diferentes para cada documento.

Neste caso, ele é composto por uma coleção de pares **palavra:frequência**, e também podem ser abstraídos como coleção de pares **id:frequência**, onde *id* é o identificador único atrelado a cada termo dos documentos. Um exemplo extraído na construção do dicionário, com

¹⁵ Código para preparar os dados de entrada para obter um modelo LDA com a biblioteca Gensim e realizar o treinamento do modelo. Disponível em <https://github.com/o-eduardo/PGC-LDA/blob/master/AutomacaoExtracaoDados/testes/automacao_treinamento_modelo.py>. Acesso em: 14 abr. 2021

Matriz de Termo-Documento			
	segurança	acessibilidade	jogos
Documento 1	2	1	0
Documento 2	1	1	0
Documento 3	0	1	3
Documento 4	2	1	0

Tabela 4.1 – Exemplo de representação matricial documento-termo.

uma lista de listas contendo os pares palavra-frequência: [(‘abstract’, 1), (‘algum’, 2), (‘ambiente’, 1), (‘aplicacao’, 2), (‘aprendizado’, 2), (‘baseado’, 2)...]

4.4.2 Treinamento do Modelo

A aplicação do modelo LDA necessita além do dicionário e do *corpus*, da informação de valores para hiperparâmetros como: número de iterações, passos, número de tópicos (K), α e β . Ao definir os valores destes hiperparâmetros, realizamos a etapa de treino do modelo.

Uma análise proposta por (BINKLEY et al., 2014) concluiu que a melhor contagem de tópicos pode ser aferida com um número K de tópicos, tal que K está no conjunto {5, 15, 25, 50, 75}. Para o hiperparâmetro passos, que representa o número de passos através do corpus durante o treinamento, (BINKLEY et al., 2014) sugere definir o valor de 50. Já para os hiperparâmetros α e β , o autor discorre que quando o valor de α for alto, os documentos provavelmente compreenderão uma maior mistura de tópicos, e no caso contrário, a mistura será de poucos tópicos. Quando o valor de β for alto, cada tópico terá uma maior probabilidade de possuir misturas de várias palavras, e no caso contrário, o tópico será formado por poucas palavras. O autor propõe a adoção do hiperparâmetro β com valor $\beta = 10^{-2}$.

No processo de treino, a cada geração de tópicos, variando dentre as valores possíveis de K e α , foi realizada uma avaliação por meio da métrica de Coerência C_v , com o objetivo de maximizar este valor. Para cada número K de tópicos presente em {5, 15, 25, 50, 75}, houve a variação no valor de α em {0.01, 0.31, 0.61, 0.909, *symmetric*, *asymmetric*} com a aspiração de definir o valor do α .

Para K igual ao número de tópicos, os valores *symmetric*, *asymmetric* são definidos conforme as Equações 4.1 e 4.2, respectivamente.

$$\alpha = \frac{1}{K} \quad (4.1)$$

$$\alpha = \frac{1}{\sqrt{K} + 1} \quad (4.2)$$

Nos baseamos na proposta de variação dos hiperparâmetro apresentada por (LI, 2019) para complementar o treinamento do modelo. Deste modo, foram selecionados os melhores

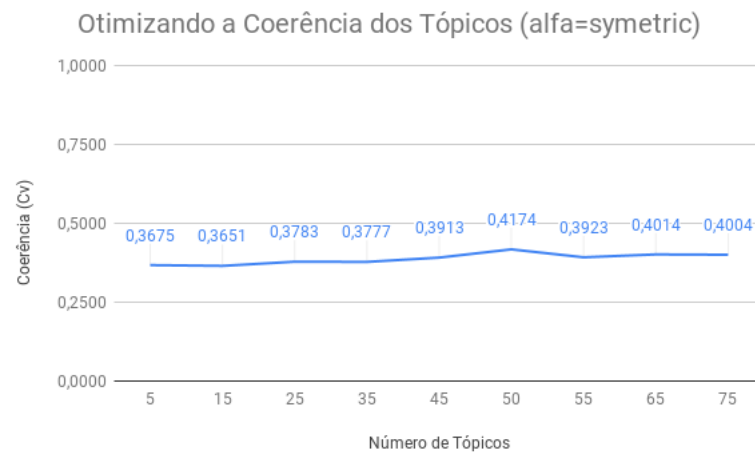


Figura 4.5 – Gráfico da coerência do modelo para hiperparâmetro α com valor "symetric".

resultados da combinação de K tópicos com os hiperparâmetros para o nosso modelo de tópicos com o LDA. Foram realizados 55 testes conforme as combinações para a otimização da métrica de coerência. A Figura 4.5 apresenta um gráfico com o desempenho de α com valor "symetric", onde a maior coerência observada foi com 50 tópicos. O valor da coerência foi 0.4174.

Sendo assim, os hiperparâmetros foram definidos com: número de tópicos $K = 50$; $\alpha = \text{"symetric"}$ e $\beta = 10^{-2}$. Com estes valores, a próxima etapa foi a extração dos tópicos propriamente dita.

4.4.3 Extração dos Tópicos

Utilizamos um algoritmo¹⁶ para a extração e obtenção dos tópicos, onde submetemos como parâmetros para o módulo da biblioteca Gensim que realiza a modelagem do LDA: o dicionário, o *corpus*, o número de passos, o número de tópicos, o valor α e o valor β . Dessa forma, resultando na obtenção do modelo LDA já treinado.

¹⁶ Código para testes de extração dos tópicos do modelo LDA obtido. Disponível em <https://github.com/o-eduardo/PGC-LDA/blob/master/AutomacaoExtracaoDados/testes/automacao_extracao_topicos.py>. Acesso em: 14 abr. 2021.

5 Resultados

5.1 Tópicos Extraídos

A extração de tópicos aplicada em coleções de documentos pode ser encontrada em diversos trabalhos na literatura. Este estudo visa a análise de artigos científicos, especialmente da área de Tecnologia da Informação e domínios correlatos, de modo que foram selecionados estudos similares, com a aplicação de modelagem de tópicos sobre textos científicos relacionados a Tecnologias da Informação.

Nesta seção apresentam-se os resultados preliminares obtidos, dadas as extrações de tópicos. A partir da execução do LDA, foram obtidos os tópicos com as palavras, e sua respectiva probabilidade de ocorrência no tópico. Todos os demais tópicos foram apresentados no Apêndice A.

Analizando os tópicos obtidos de maneira geral e examinando os termos contidos em cada tópico, é possível inferir algumas atribuições de rótulos em alguns tópicos. Esta rotulação sugestiva pode ser feita tanto pelo programador, quanto por votação em um grupo de pessoas envolvidas atuantes nas áreas correlatas aos assuntos citados na base de dados, conforme proposto por (NOLASCO, 2016).

Para a base de dados utilizada nesse trabalho, verifica-se, assim como nos tópicos extraídos no Apêndice A, que predominam assuntos ligados à **Sistemas de Informação**, como “sistemas”, “banco de dados”, “desenvolvimento”, “testes”, “software”, “caso de uso” etc. Os tópicos 2 e 42 representam esses assuntos, conforme a Tabela 5.1, apresentada abaixo.

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"

Tabela 5.1 – Quadro com os tópicos 2 e 42.

Além disso, notamos tópicos relativos a **Ensino, Aprendizagem e Interação Humano-Computador**, possuindo termos como: “ambientes de aprendizagem”, “lousa digital”, “alunos”, “ensino” entre outros, conforme os tópicos 7, 20 e 43 na Tabela 5.2.

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"

Tabela 5.2 – Quadro com os tópicos 7, 20 e 43.

Tópicos com termos relacionados à **Inteligência Artificial** também foram notados, sob os vocábulos: “neural”, “aprendizagem por reforço”, “q-learning”, “treinamento”, “inteligencia artificial”, modelados nos tópicos 5 e 34, de acordo com a Tabela 5.3.

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"

Tabela 5.3 – Quadro com os tópicos 5 e 34.

Nos tópicos 10, 16 e 31, apresentados na Tabela 5.4, temos termos relacionados à **Internet das coisas** (IoT), **Arduíno** e etc, um conceito que se refere à interconexão digital de objetos cotidianos com a internet.

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"

Tabela 5.4 – Quadro com os tópicos 10, 16 e 31.

Os tópicos 39 e 41, presente na Tabela 5.5, modelaram o assunto de **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"

Tabela 5.5 – Quadro com os tópicos 39 e 41.

Já nos tópicos 6, 26 e 40, ilustrados na Tabela 5.6, trataram dos artigos relacionados a **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"

Tabela 5.6 – Quadro com os tópicos 6, 26 e 40.

Os tópicos 29 e 49 apresentaram termos relacionados a **LIBRAS** (Linguagem Brasileira de Sinais) e **Acessibilidade**, de acordo com a Tabela 5.7.

Tópico	Termos e pesos do tópico
29	0.014*"libras"; 0.005*"língua"; 0.003*"surdas"; 0.003*"comunicativa"; 0.003*"portuguesa"; 0.002*"indivíduo"; 0.002*"sinais"; 0.002*"sinais"; 0.002*"áudio"; 0.002*"eeg"
49	0.005*"online"; 0.004*"monitoramento"; 0.004*"interação"; 0.004*"interação"; 0.003*"wi-fi"; 0.004*"cores"; 0.003*"acessibilidade"; 0.003*"imagens"; 0.003*"deficiência"; 0.003*"monitoramento online"

Tabela 5.7 – Quadro com o tópico 29.

A área de **Gestão de Projetos e Riscos** poderia ser relacionada com os tópicos 36 e 44, extraídos neste trabalho e apresentados na Tabela 5.8

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";

Tabela 5.8 – Quadro com os tópicos 36 e 44.

Outro possível tema de tópico notado na extração foi encontrado no tópico 24, expondo o tema de Árvore Geradoras Mínimas da área de **Teoria dos Grafos**, de acordo com a Tabela 5.9.

Tópico	Termos e pesos do tópico
24	0.006*"problema"; 0.005*"geradora"; 0.005*"árvore geradora"; 0.005*"árvore geradora mínima"; 0.005*"geradora mínima"; 0.005*"mínima"; 0.005*"busca"; 0.005*"árvore"; 0.004*"algoritmo"; 0.004*"generalizado"

Tabela 5.9 – Quadro com o tópico 24.

Por fim, destacamos os tópicos 8 e 15, que respectivamente, poderiam modelar os assuntos de **Programação Paralela** e **Computação em Nuvem**, tendo em vista as Tabelas 5.10 e 5.11.

Tópico	Termos e pesos do tópico
8	0.008*"trabalho"; 0.005*"gpu"; 0.005*"redes"; 0.005*"algoritmo"; 0.005*"cpu"; 0.004*"rede"; 0.004*"cuda"; 0.004*"mpi"; 0.004*"opencl"; 0.004*"opencl"

Tabela 5.10 – Quadro com o tópico 8.

Tópico	Termos e pesos do tópico
15	0.008*"nuvem"; 0.004*"openstack"; 0.003*"infraestrutura"; 0.003*"idoso"; 0.002*"serviço"; 0.002*"recursos computacionais"; 0.002*"computação nuvem"; 0.002*"computacionais"; 0.002*"idosos"; 0.002*"media queries"

Tabela 5.11 – Quadro com o tópico 15.

Também verificamos algumas ferramentas mais específicas presentes em certos tópicos, como por exemplo: *CUDA*¹, *MPI*, *OpenMP*², *Python*, *Hadoop*³, *Cocos2d*⁴, *Openstack*⁵, etc., onde tais ferramentas e *frameworks* de mercado estiveram completamente relacionados aos seus respectivos tópicos.

¹ <<https://developer.nvidia.com/cuda-zone>>

² <<https://www.openmp.org/>>

³ <<https://hadoop.apache.org/>>

⁴ <<https://www.cocos.com/>>

⁵ <<https://www.openstack.org/>>

5.2 Sobreposição e Disjunção dos Tópicos

Outro aspecto observado nos resultados foi a presença de sobreposição entre os termos dos tópicos extraídos. Como tópicos com sobreposições, notamos: "Sistema", "Dados", "Redes", "Desenvolvimento", "software", "informação" etc.

Podemos supor tal comportamento devido ao fato de que as áreas do conhecimento em que os projetos desenvolvidos nos eventos estão ligada, compartilhando conceitos em suas fundamentações.

Uma outra maneira de visualizar a disposição dos tópicos está apresentada na Figura 5.1, obtida através da biblioteca LDAVis. Nos baseamos no *script* de sugestão de uso da documentação da biblioteca proposto em (SIEVERT; SHIRLEY, 2014 apud MABEY, 2015). A única preparação necessária para obter a Figura 5.1 foi o envio do modelo de tópicos (gerado com o Gensim) como parâmetro do módulo da biblioteca LDAVis.

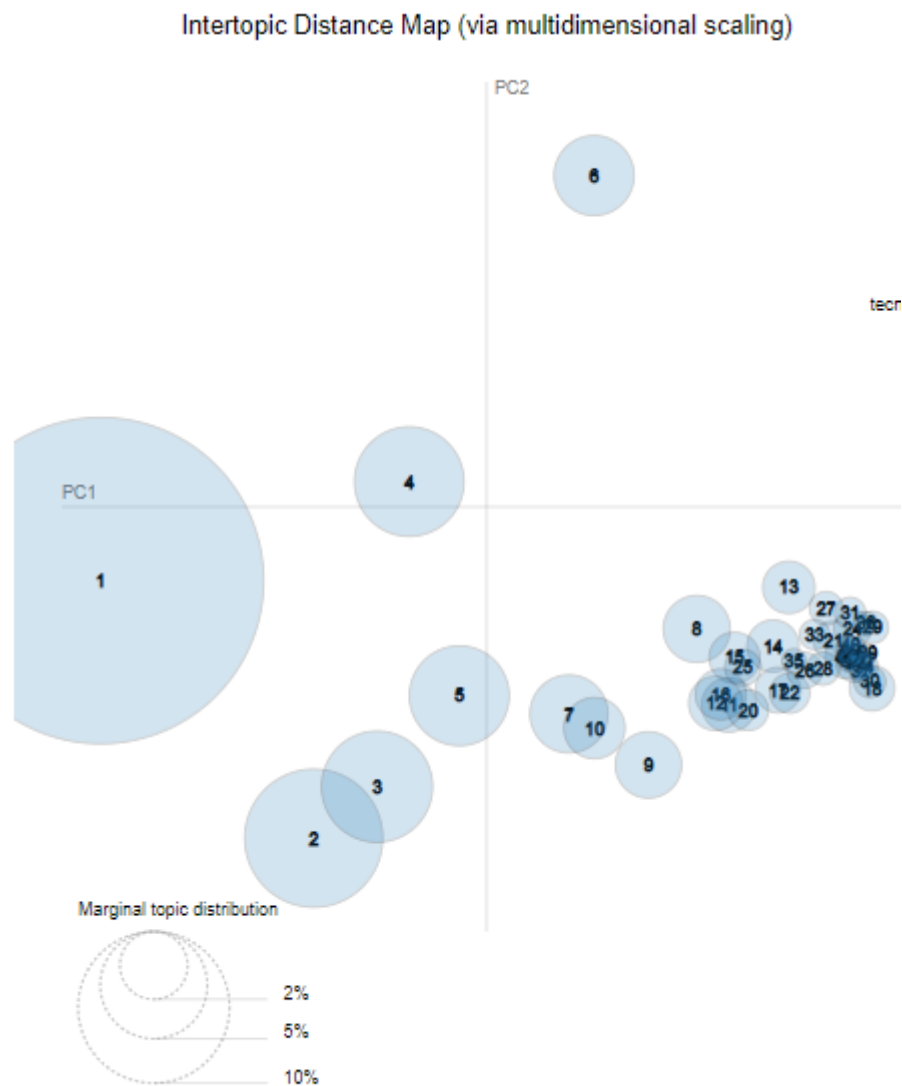


Figura 5.1 – Mapa da distância entre os tópicos.

A Figura 5.1 representa o mapeamento das distâncias entre os tópicos do modelo LDA em um plano bidimensional. Podemos notar que alguns tópicos se concentraram em regiões mais próximas uns dos outros, evidenciando as sobreposições citadas anteriormente.

De maneira oposta, observamos a disjunção de alguns tópicos, com um maior distanciamento entre os mesmos. Este comportamento pode estar relacionado à segmentação dos tópicos por assuntos que não compartilham os mesmos termos.

5.3 Análise dos Resultados

Na análise dos tópicos, foi realizada a segmentação dos documentos, dividindo os mesmos de acordo com o ano da edição do evento, com o objetivo de obter os principais tópicos

apresentados ao longo das edições do evento. Além disso, verificou-se os termos mais frequentes dos documentos para uma comparação com os tópicos mais frequentes.

Os 10 termos mais frequentes dentre todos os artigos do EATI foram: “dados”, “sistema”, “desenvolvimento”, “informação”, “trabalho”, “tecnologia”, “processo”, “aplicação”, “software” e “web”. A distribuição destes termos e suas respectivas frequências ao longo das edições do EATI pode ser visualizada pelo gráfico da Figura 5.2.

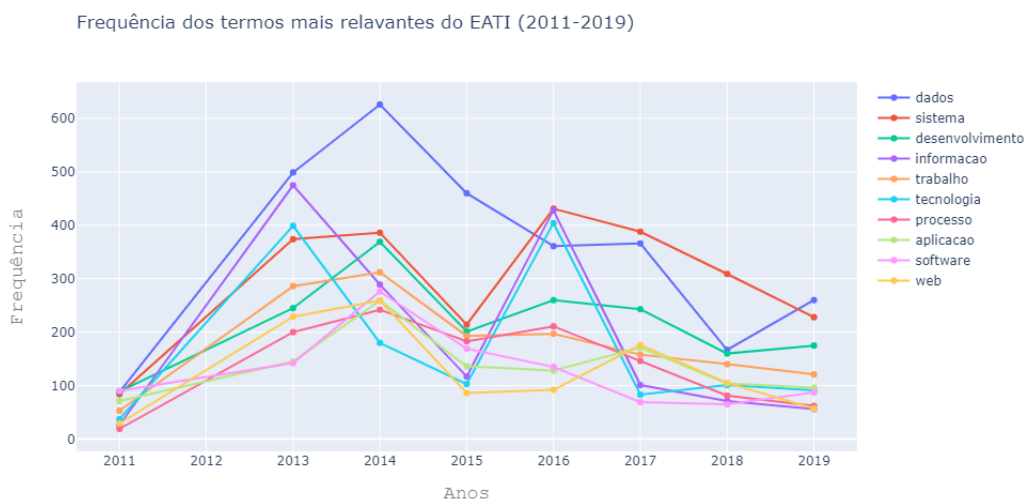


Figura 5.2 – Frequência dos termos mais relevantes do EATI (2011-2019).

Dito isso, realizamos uma extração de tópicos para os artigos das edições em que o EATI apresentou um tema transversal, de modo que as edições que apresentaram tais temas nos eventos foram as de 2013, 2014 e 2015.



Figura 5.3 – Palavras mais frequentes nos artigos da edição IV do EATI em 2013.

O IV EATI apresentou 51 artigos e teve o seguinte tema: “A Educação na Área de TI e seus Desafios”. As palavras mais frequentes nos artigos da edição de 2013 estão apresentadas na

Figura 5.3, ao passo que os 6 principais assuntos perante a extração de tópicos estão apresentados na Figura 5.4.

Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5	Tópico 6
<ul style="list-style-type: none"> • "web" • "informação" • "dados" • "documento" • "usuário" • "forma" • "segurança" • "xml" • "trabalho" • "conteúdo" 	<ul style="list-style-type: none"> • "informação", • "tecnologia", • "tecnologia informação", • "estratégico", • "alinhamento", • "dados", • "sistema", • "informações", • "governança", • "alinhamento estratégico". 	<ul style="list-style-type: none"> • "dados" • "riscos" • "sistema" • "projeto" • "informação" • "tecnologia" • "projetos" • "identificação" • "análise" • "processo"; 	<ul style="list-style-type: none"> • "tempo" • "trabalho" • "desempenho" • "threads" • "speedup" • "dados" • "gpu" • "algoritmo" • "openmp" • "cpu"; 	<ul style="list-style-type: none"> • "sistema" • "jogos" • "desenvolvimento" • "forma" • "ensino" • "dados" • "processo" • "aprendizagem" • "pesquisa" • "tecnologia" 	<ul style="list-style-type: none"> • "dados" • "alunos" • "aprendizagem" • "utilização" • "serviço" • "web" • "ambiente" • "ensino" • "sistema" • "informação"

Figura 5.4 – Tópicos da edição IV do EATI em 2013.

Para os 6 tópicos extraídos, destacamos os tópicos 5 e 6, que apresentaram palavras bastante relacionadas ao tema do evento, como “aprendizagem”, “pesquisa”, “ensino”, “jogos” e “alunos”. Por outro lado, as palavras mais frequentes nos textos pouco exprimem algum significado relativo ao tema, devido ao fato de serem mais genéricas, como “sistema”, “informação” etc.

O tema da edição de 2014 foi: “As Tecnologias de Informação e Comunicação, a Academia e o Mercado de Trabalho”, onde nesta edição foram apresentados 63 trabalhos. As palavras mais frequentes nos artigos da edição de 2014 estão apresentadas na Figura 5.5. Na Figura 5.6 temos os 6 principais assuntos presentes na edição V do EATI.



Figura 5.5 – Palavras mais frequentes nos artigos da edição V do EATI em 2014.

Dentre estes tópicos extraídos, da edição de 2014, foi interessante notar novamente que alguns deles se relacionam com o tema do evento, que por sua vez cita o “Mercado de Trabalho” e a “Academia” na área de Tecnologia da Informação e Comunicação.

Os tópicos 1, 5 e 6 exprimem alguns aspectos do mercado de trabalho na área de desenvolvimento de sistemas, como “requisitos”, “desenvolvimento”, “usuário”, “ambiente”, “testes mutantes”, “casos de teste” e “rede”. Já para a parte da Academia, conforme o tema do evento,

que neste tópico também foram encontrados termos relativos à mineração de dados, também relacionada aos aspectos da população. As áreas de Pesquisa e Ensino citadas no tema também puderam ser mapeadas pelo tópico 3.

Outro ponto a se destacar está presente no tópico 6, em que temos aspectos de recursos humanos, recrutamento e seleção de profissionais. Estes termos podem se relacionar com a ideia de que o desenvolvimento regional está ligado à uma maior concentração de empresas e profissionais da área de Tecnologia da Informação e Comunicação, conforme o tema do evento.

6 Considerações Finais

No presente capítulo, as considerações finais serão apresentadas com o objetivo de elucidar o desfecho dos objetivos descritos na introdução, e também a continuidade deste trabalho. Separamos as considerações finais em duas etapas: Conclusão e Trabalhos Futuros.

6.1 Conclusão

No presente projeto de graduação foi realizada uma análise exploratória em uma base de dados composta por artigos científicos, apresentados nas edições do Encontro Anual de Tecnologia da Informação (EATI). Para possibilitar a execução do projeto, foram efetuadas as etapas de coleta de dados, com a extração dos artigos e consolidação destes em formato de texto; pré-processamento, envolvendo a limpeza e normalização dos dados; treinamento do modelo de tópicos com a abordagem do LDA e teste do modelo LDA com os hiperparâmetros obtidos no treinamento para a extração dos tópicos.

Os hiperparâmetros foram obtidos a partir da adoção de uma métrica de coerência do modelo, ao passo que após o treinamento para maximizar a coerência, o modelo LDA com 50 tópicos foi considerado o melhor.

Para avaliação dos 50 tópicos extraídos, foram analisadas as principais palavras de cada tópico, constatando relações entre elas e relacionando-as a um possível assunto determinado. Assim, foram observados os tópicos mais recorrentes dentre os artigos, atrelados principalmente a banco de dados, desenvolvimento de software, e processos. Para complementar a análise dos tópicos, a distribuição dos termos mais frequentes da base de dados foi apresentada. Além disso, foi executada a extração dos principais tópicos nas edições do EATI que possuíam um tema central definido.

Devido à análise dos tópicos extraídos, constatou-se que os mesmos refletiram os assuntos relacionados ao EATI e as áreas correlatas de Tecnologia da Informação, Sistemas de Informação e Ciências da Computação. Não obstante, percebeu-se que os tópicos obtidos poderiam agrupar, resumir e organizar os artigos através de assuntos, e que os tópicos coletados nas edições com tema central apresentaram alguma relação com os respectivos temas.

6.2 Trabalhos Futuros

Em relação ao trabalhos futuros, conseguimos citar algumas oportunidades para abranger a análise realizada neste trabalho. Pode ser realizada a modelagem de tópicos com base em outros algoritmos, e dessa forma, comparar os resultados analisando outras métricas além da coerência

do modelo.

Desse modo, pode ser realizada a rotulação por assunto dos tópicos extraídos, baseada em discussões com pessoas envolvidas nas áreas de Tecnologia da Informação, Sistemas de Informação e Ciências da Computação (discentes, docentes, profissionais do mercado, etc).

Para aprofundar as análises, pode ser feita a utilização de algoritmos para a identificação dos tópicos por documento. Por fim, pode ser aplicado um algoritmo de aprendizado de máquina para a definição dos tópicos rotulados por documentos de maneira automática.

7 Cronograma

Neste capítulo, o planejamento das atividades necessárias para realização deste trabalho será apresentado sob o formato de um cronograma.

Atividades	Meses - 2020											
	2	3	4	5	6	7	8	9	10	11	12	
Escolha do Orientador	x											
Apresentação de Propostas para Orientador	x											
Definição do Tema do Trabalho	x	x										
Pesquisa e Levantamento Bibliográfico	x	x	x	x								
Redação das Referências Bibliográficas	x	x	x	x								
Definição do Método de Coleta de Dados		x										
Redação do Cronograma		x	x									
Redação da Descrição da Metodologia		x	x									
Redação dos Objetivos		x	x									
Redação da Introdução			x									
Redação da Identificação			x									
Redação da Justificativa			x									
Revisão do Relatório Consolidado PGC I			x									
Submissão do Relatório Consolidado PGC I			x									
Modelagem da Automação da Extração dos Artigos da Base de Dados					x	x						
Desenvolvimento e Execução da Automação da Extração dos Artigos					x	x						
Rastreabilidade e Mapeamento dos Artigos em uma Planilha					x	x						
Modelagem da Consolidação e União dos Artigos no Formato de Texto					x	x						
Desenvolvimento e Execução da Modelagem da Consolidação dos Artigos no Formato de Texto					x	x						
Pré-Processamento dos Dados						x	x	x				
Extração dos Tópicos com Modelagem LDA								x	x			
Redação da Fundamentação Teórica										x		
Redação do Detalhamento da Metodologia										x		
Redação dos Resultados Preliminares										x		
Revisão do Relatório Consolidado PGC II											x	
Submissão do Relatório Consolidado PGC II											x	

Tabela 7.1 – Cronograma de atividades 2020.

Atividades	Meses - 2021											
	2	3	4	5	6	7	8	9	10	11	12	
Análise Exploratória dos Tópicos	x	x										
Apresentação dos Resultados Experimentais	x	x										
Redação da Análise dos Resultados		x	x									
Redação da Conclusão e Trabalhos Futuros			x									
Revisão do Relatório Consolidado PGC III			x									
Submissão do Relatório Consolidado PGC III			x									

Tabela 7.2 – Cronograma de atividades 2021.

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*. New York, NY, USA: Association for Computing Machinery, 2014. (ICPC 2014), p. 26–36. Disponível em: <<https://doi.org/10.1145/2597008.2597150>>. Acesso em: 18 abr. 2021.
- BLEI, D. M. Probabilistic topic models. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 55, n. 4, p. 77–84, abr. 2012. Disponível em: <<https://doi.org/10.1145/2133806.2133826>>. Acesso em: 18 abr. 2021.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 993–1022, mar. 2003. Disponível em: <<https://dl.acm.org/doi/10.5555/944919.944937>>. Acesso em: 18 abr. 2021.
- BRUNIALTI, L.; PERES, S.; FREIRE, V.; LIMA, C. Aprendizado de máquina em sistemas de recomendação baseados em conteúdo textual: Uma revisão sistemática. In: *Anais do XI Simpósio Brasileiro de Sistemas de Informação*. Porto Alegre, RS, Brasil: SBC, 2015. p. 203–210. Disponível em: <<https://sol.sbc.org.br/index.php/sbsi/article/view/5818>>. Acesso em: 18 abr. 2021.
- 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*. Las Vegas, NV, USA, 1994. v. 161175. Disponível em: <<https://www.osti.gov/biblio/68573>>. Acesso em: 18 abr. 2021.
- 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) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016. Disponível em: <https://www.teses.usp.br/teses/disponiveis/55/55134/tde-10112016-105854/publico/ThiagodePauloFaleiros_revisada.pdf>. Acesso em: 18 abr. 2021.
- FONSECA, F. P. C. d. *Inferência das áreas de atuação de pesquisadores*. Dissertação (Mestrado) — Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018. Disponível em: <<https://www.teses.usp.br/teses/disponiveis/100/100131/tde-02032018-102111/en.php>>. Acesso em: 18 abr. 2021.
- HUGGINS, J. *Selenium webdriver Documentation*. 2013. Selenium HQ. Disponível em: <<https://www.selenium.dev/documentation/en/>>. Acesso em: 18 abr. 2021.
- KUHLING, A. Regular expression howto. *Regular Expression HOWTO—Python*, v. 2, n. 10, 2014. Disponível em: <<https://fossies.org/linux/python-docs-pdf-a4/howto-regex.pdf>>. Acesso em: 18 abr. 2021.
- KUHLMAN, D. *A python book: Beginning python, advanced python, and python exercises*. Cambridge, MA, USA: Dave Kuhlman Lutz, 2009.

- LI, S. *Topic Modeling and Latent Dirichlet Allocation (LDA) in Python*. 2019. Towards Data Science. 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. *CoRR*, cs.CL/0205028, 2002. Disponível em: <<https://arxiv.org/abs/cs/0205028>>. Acesso em: 18 abr. 2021.
- MABEY, B. *Welcome to pyLDavis's documentation*. 2015. Read The Docs. Disponível em: <<https://pyldavis.readthedocs.io/en/latest/>>. Acesso em: 14 abr. 2021.
- MATTMANN, C. A.; ZITTING, J. L. Tika in action. Manning, 2012. Disponível em: <<https://sisis.rz.htw-berlin.de/inh2012/12422815.pdf>>. Acesso em: 18 abr. 2021.
- 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. *Encontro Nacional de Pesquisa em Ciência da Informação*, v. 24, n. 2, 2018. Disponível em: <<http://hdl.handle.net/20.500.11959/brapci/103396>>. Acesso em: 18 abr. 2021.
- NOLASCO, D. *Identificação automática de áreas de pesquisa em C&T*. Dissertação (Mestrado) — Instituto de Matemática e Instituto Tércio Pacciti, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016. Disponível em: <<http://objdig.ufrj.br/15/teses/880244.pdf>>. Acesso em: 18 abr. 2021.
- REHUREK, 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. Disponível em: <<http://is.muni.cz/publication/884893/en>>. Acesso em: 18 abr. 2021.
- SIEVERT, C.; SHIRLEY, K. Ldavis: A method for visualizing and interpreting topics. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. Baltimore, MD, USA, 2014. p. 63–70. Disponível em: <<https://www.aclweb.org/anthology/W14-3110.pdf>>. Acesso em: 18 abr. 2021.
- SOUZA, M.; 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*, v. 9, n. 2, 2018. Disponível em: <<http://hdl.handle.net/20.500.11959/brapci/137081>>. Acesso em: 18 abr. 2021.
- SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. *Perspectivas em ciência da informação*, Scielo, v. 11, p. 161 – 173, 2006. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362006000200002&nrm=iso>. Acesso em: 18 abr. 2021.
- STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. *Handbook of latent semantic analysis*, v. 427, n. 7, p. 424–440, 2007. Disponível em: <<https://cocosci.princeton.edu/tom/papers/SteyversGriffiths.pdf>>. Acesso em: 18 abr. 2021.
- VIEIRA, L. C. et al. *Organização e disseminação da produção científica dos docentes do CESH/UFSM em um repositório digital*. Dissertação (Mestrado) — Centro de Ciências Sociais e Humanas, Universidade Federal de Santa Maria, Rio Grande do Sul, 2013. Disponível em: <<https://repositorio.ufsm.br/handle/1/4622>>. Acesso em: 18 abr. 2021.

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. Disponível em: <<https://www.redalyc.org/articulo.oa?id=465645954004>>. Acesso em: 18 abr. 2021.

ZIMAN, J. *Conhecimento público*. 4. ed. Olympia, WA, USA: Itatiaia Belo Horizonte, 1971. v. 26. 338-345 p.

Apêndices

APÊNDICE A – Tópicos Extraídos

Tópico 0	Tópico 1	Tópico 2	Tópico 3
0.013*"dificuldade"	0.009*"lixo"	0.008*"sistema"	0.003*"media"
0.011*"jogo"	0.007*"eletrônico"	0.006*"uso"	0.003*"exame"
0.009*"jogador"	0.007*"lixo eletrônico"	0.005*"fuzzy"	0.006*"notas"
0.006*"documentos"	0.006*"ambiental"	0.004*"software"	0.002*"aprovação"
0.006*"servidor"	0.005*"eletrônicos"	0.004*"objetos"	0.002*"nota necessária"
0.005*"autenticação"	0.004*"vegetação"	0.003*"desenvolvimento"	0.002*"design participativo"
0.004*"cliente"	0.003*"reciclagem"	0.003*"casos uso"	0.002*"participativo"
0.004*"fluxo"	0.003*"materiais"	0.003*"casos"	0.002*"calculado"
0.004*"ajuste"	0.003*"vendas"	0.003*"tabela"	0.002*"semestre"
0.003*"modulo"	0.003*"resíduos"	0.003*"processo"	0.002*"nota"

Tabela A.1 – Tópicos 0-3 extraídos da base de artigos.

Tópico 4	Tópico 5	Tópico 6	Tópico 7
0.011*"áudio"	0.006*"algoritmo"	0.010*"rede"	0.010*"sistema"
0.007*"mp"	0.005*"dados"	0.010*"sistema"	0.008*"desenvolvimento"
0.007*"câncer"	0.005*"treinamento"	0.006*"dados"	0.006*"forma"
0.006*"mdct"	0.005*"neural"	0.005*"teste"	0.006*"usuário"
0.005*"coeficientes"	0.005*"base"	0.005*"software"	0.005*"trabalho"
0.004*"colo"	0.004*"aprendizagem"	0.004*"servidor"	0.005*"uso"
0.004*"compressão"	0.004*"rede"	0.004*"comunicação"	0.005*"projeto"
0.004*"câncer colo"	0.004*"rede neural"	0.004*"protocolo"	0.005*"ambiente"
0.004*"risco"	0.004*"artificial"	0.004*"intrusão"	0.005*"aprendizagem"
0.004*"colo útero"	0.004*"neural artificial"	0.004*"redes"	0.005*"web"

Tabela A.2 – Tópicos 4-7 extraídos da base de artigos.

Tópico 8	Tópico 9	Tópico 10	Tópico 11
0.008*"trabalho"	0.005*"encaminhamentos"	0.012*"dados"	0.010*"informação"
0.005*"gpu"	0.004*"estudantil"	0.007*"sensores"	0.008*"forma"
0.005*"redes"	0.003*"coordenação"	0.006*"aplicações"	0.006*"dados"
0.005*"algoritmo"	0.003*"assistência estudantil"	0.005*"cidade"	0.006*"tecnologia"
0.005*"cpu"	0.003*"assistência"	0.005*"internet"	0.005*"rede"
0.004*"rede"	0.003*"coord. assistência"	0.004*"nota"	0.005*"alunos"
0.004*"cuda"	0.003*"coord. assistência estudantil"	0.004*"pagina"	0.005*"uso"
0.004*"mpi"	0.003*"discente"	0.004*"energia"	0.005*"resultados"
0.004*"opencl"	0.002*"farroupilha"	0.004*"iot"	0.005*"trabalho"
0.004*"openmp"	0.002*"opção"	0.003*"social"	0.005*"informações"

Tabela A.3 – Tópicos 8-11 extraídos da base de artigos.

Tópico 12	Tópico 13	Tópico 14	Tópico 15
0.009*"imagem"	0.008*"processo"	0.008*"jogo"	0.008*"nuvem"
0.008*"filtro"	0.007*"tráfego"	0.006*"processo"	0.004*"openstack"
0.007*"radio"	0.005*"seleção"	0.005*"solicitações"	0.003*"infraestrutura"
0.007*"identificação"	0.005*"fincão"	0.005*"alunos"	0.003*"idoso"
0.007*"frequência"	0.005*"recrutamento"	0.004*"linguagem"	0.002*"serviço"
0.006*"identificação radio"	0.004*"ferramenta"	0.004*"logo"	0.002*"recursos computacionais"
0.006*"identificação radio frequência"	0.004*"império"	0.004*"serviço"	0.002*"computação nuvem"
0.006*"radio frequência"	0.004*"recrutamento seleção"	0.003*"software"	0.002*"computacionais"
0.006*"tag"	0.004*"resultados"	0.003*"material"	0.002*"idosos"
0.005*"leitor"	0.004*"montagem"	0.003*"forma"	0.002*"media queries"

Tabela A.4 – Tópicos 12-15 extraídos da base de artigos.

Tópico 16	Tópico 17	Tópico 18	Tópico 19
0.011*"energy"	0.009*"sistema"	0.008*"sistema"	0.004*"veículo"
0.008*"system"	0.007*"software"	0.006*"dados"	0.003*"trilha"
0.007*"solar"	0.006*"processo"	0.005*"informação"	0.003*"ângulo"
0.007*"power"	0.005*"desenvolvimento"	0.005*"forma"	0.003*"linha"
0.005*"arduino"	0.005*"dados"	0.004*"tecnologia"	0.002*"veículos"
0.004*"figure"	0.005*"conhecimento"	0.004*"trabalho"	0.002*"protótipo"
0.004*"gsm"	0.004*"trabalho"	0.004*"processo"	0.002*"câmera"
0.004*"data"	0.004*"sistemas"	0.003*"desenvolvimento"	0.002*"raspberry"
0.003*"generation"	0.003*"metodologia"	0.003*"alunos"	0.002*"motores"
0.003*"grid"	0.003*"afirmações"	0.003*"usuário"	0.002*"direção"

Tabela A.5 – Tópicos 16-19 extraídos da base de artigos.

Tópico 20	Tópico 21	Tópico 22	Tópico 23
0.005*"docentes"	0.008*"métodos"	0.015*"modelo"	0.004*"godot"
0.004*"bloco"	0.007*"software"	0.007*"população"	0.004*"cadeira"
0.004*"lousa"	0.005*"sangue"	0.005*"numero"	0.004*"linietsky"
0.004*"mão"	0.004*"doação"	0.005*"taxa"	0.003*"teto"
0.003*"dedos"	0.004*"ágeis"	0.005*"epidemias"	0.003*"manzur"
0.003*"algoritmo"	0.004*"processos"	0.004*"crescimento"	0.003*"nodes"
0.003*"lousa digital"	0.004*"modelo"	0.004*"contaminação"	0.003*"jogo"
0.003*"distancia"	0.004*"doadores"	0.004*"proposto"	0.003*"linietsky manzur"
0.003*"forma"	0.004*"doação sangue"	0.004*"célula"	0.003*"python"
0.003*"moodle"	0.004*"aspectos"	0.004*"dinâmica"	0.002*"fig"

Tabela A.6 – Tópicos 20-23 extraídos da base de artigos.

Tópico 24	Tópico 25	Tópico 26	Tópico 27
0.006*"problema"	0.006*"sistema"	0.007*"flow"	0.006*"k-means"
0.005*"geradora"	0.006*"recomendação"	0.006*"network"	0.003*"map-reduce"
0.005*"arvore gera- dora"	0.005*"velocidade"	0.005*"module"	0.002*"centroides"
0.005*"arvore gera- dora mínima"	0.005*"fuzzy"	0.005*"neutrality"	0.002*"agrícolas"
0.005*"geradora mi- nima"	0.004*"padrão"	0.005*"router"	0.002*"paralelo"
0.005*"minima"	0.004*"hadoop"	0.004*"traffic"	0.002*"hadoop"
0.005*"busca"	0.004*"arquivos"	0.004*"metrics"	0.002*"amostras"
0.005*"árvore"	0.004*"desempenho"	0.004*"network neutrality"	0.002*"fluxo"
0.004*"algoritmo"	0.004*"imagem"	0.003*"figure"	0.002*"função"
0.004*"generalizado"	0.004*"oas"	0.003*"breaking"	0.001*"cluster"

Tabela A.7 – Tópicos 24-27 extraídos da base de artigos.

Tópico 28	Tópico 29	Tópico 30	Tópico 31
0.002*"ambiente virtual"	0.014*"libras"	0.007*"dados"	0.007*"arduino"
0.002*"lembretes"	0.005*"língua"	0.006*"desenvolvimento"	0.006*"jovens"
0.002*"aveas"	0.003*"surdas"	0.005*"forma"	0.004*"quedas"
0.002*"september"	0.003*"comunicativa"	0.005*"sistema"	0.004*"pulseira"
0.001*"inclass"	0.003*"portuguesa"	0.005*"usuário"	0.003*"eventos"
0.001*"organizer"	0.002*"indivíduo"	0.004*"web"	0.003*"oficinas"
0.001*"complete"	0.002*"sinais"	0.004*"tecnologia"	0.003*"evento"
0.001*"class"	0.002*"gif"	0.004*"uso"	0.003*"manifestações"
0.001*"ifrsbg"	0.002*"áudio"	0.003*"controle"	0.002*"empreendedorismo"
0.001*"cco"	0.002*"eeg"	0.003*"resultados"	0.002*"cocos2d"

Tabela A.8 – Tópicos 28-31 extraídos da base de artigos.

Tópico 32	Tópico 33	Tópico 34	Tópico 35
0.012*"modelo"	0.004*"população"	0.012*"reforço"	0.011*"aplicativo"
0.007*"baterias"	0.004*"indivíduos"	0.009*"aprendizagem"	0.004*"desenvolvimento"
0.007*"vida"	0.003*"evolutiva"	0.007*"aprendizagem reforço"	0.004*"soluções"
0.006*"tempo"	0.003*"computação"	0.006*"agente"	0.004*"android"
0.005*"battery"	0.003*"solução"	0.004*"ação"	0.004*"moveis"
0.005*"malária"	0.003*"créditos"	0.003*"conexionista"	0.004*"repetição"
0.005*"resultados"	0.003*"grade"	0.003*"neural"	0.003*"algoritmos"
0.005*"ion"	0.002*"professor"	0.003*"exemplos"	0.003*"usuário"
0.004*"erro"	0.002*"problema"	0.003*"q-learning"	0.003*"aplicativos"
0.004*"bateria"	0.002*"computação evolutiva"	0.002*"ações"	0.003*"dispositivos"

Tabela A.9 – Tópicos 32-35 extraídos da base de artigos.

Tópico 36	Tópico 37	Tópico 38	Tópico 39
0.015*"dados"	0.020*"jogos"	0.012*"web"	0.011*"thread"
0.012*"informação"	0.015*"jogo"	0.009*"dados"	0.008*"documento"
0.009*"tecnologia"	0.010*"programação"	0.009*"aplicação"	0.008*"xml"
0.006*"tecnologia infor- mação"	0.010*"ensino"	0.009*"sistema"	0.007*"threads"
0.009*"conhecimento"	0.008*"aprendizagem"	0.008*"desenvolvimento"	0.007*"work"
0.004*"estratégico"	0.007*"desenvolvimento"	0.007*"aplicações"	0.006*"dados"
0.004*"alinhamento"	0.007*"alunos"	0.006*"dispositivos"	0.006*"execução"
0.004*"data"	0.007*"digitais"	0.006*"ferramenta"	0.006*"units"
0.004*"empresa"	0.005*"algoritmos"	0.005*"usuário"	0.006*"work units"
0.004*"projetos"	0.004*"engine"	0.005*"afirmações"	0.005*"thread vs"

Tabela A.10 – Tópicos 36-39 extraídos da base de artigos.

Tópico 40	Tópico 41	Tópico 42	Tópico 43
0.009*"sistema"	0.012*"cluster"	0.031*"dados"	0.008*"alunos"
0.006*"consumo"	0.007*"desempenho"	0.008*"afirmações"	0.005*"atividades"
0.005*"dnssec"	0.006*"tempo"	0.007*"sistema"	0.005*"dados"
0.005*"dados"	0.004*"threads"	0.007*"banco"	0.004*"resultados"
0.004*"informação"	0.004*"resultados"	0.006*"banco dados"	0.004*"projeto"
0.004*"estoque"	0.004*"computadores"	0.005*"data"	0.004*"programação"
0.004*"dns"	0.004*"núcleos"	0.004*"trabalho"	0.004*"desenvolvimento"
0.004*"segurança"	0.003*"sistema"	0.004*"forma"	0.003*"sistemas"
0.004*"web"	0.003*"beowulf"	0.004*"analise"	0.003*"ferramenta"
0.004*"indústria"	0.003*"número"	0.003*"teste"	0.003*"ensino"

Tabela A.11 – Tópicos 40-43 extraídos da base de artigos.

Tópico 44	Tópico 45	Tópico 46	Tópico 47
0.029*"informação"	0.004*"profissionais"	0.005*"informação"	0.006*"programação"
0.025*"tecnologia infor- mação"	0.004*"web services"	0.005*"tecnologia"	0.004*"alunos"
0.023*"tecnologia"	0.004*"método"	0.004*"textos"	0.004*"tempo"
0.011*"riscos"	0.004*"socket"	0.004*"ensino"	0.003*"ensino"
0.011*"processos"	0.004*"services"	0.004*"technology"	0.003*"dados"
0.010*"gerenciamento"	0.003*"internet"	0.004*"education"	0.003*"método"
0.010*"gestão"	0.003*"web"	0.004*"aprendizagem"	0.003*"computação"
0.009*"serviços"	0.002*"sockets"	0.003*"tecnologia infor- mação"	0.003*"musica"
0.006*"organização"	0.002*"volume"	0.003*"tecnologia infor- mação cs"	0.003*"android"
0.006*"segurança"	0.002*"comunicação"	0.003*"informação cs"	0.003*"superior"

Tabela A.12 – Tópicos 44-47 extraídos da base de artigos.

Tópico 48	Tópico 49
0.008*"comunicação"	0.005*"online"
0.008*"radio"	0.004*"monitoramento"
0.004*"desenvolvimento"	0.004*"interação"
0.004*"processo"	0.004*"visual"
0.004*"escola"	0.004*"cores"
0.004*"linguagem"	0.003*"acessibilidade"
0.004*"interação"	0.003*"wi-fi"
0.004*"deficiência"	0.003*"imagens"
0.003*"social"	0.003*"deficiência"
0.003*"uso"	0.003*"monitoramento online"

Tabela A.13 – Tópicos 48-49 extraídos da base de artigos.