Fundação Universidade Federal do ABC
Pró reitoria de pesquisa
Av. dos Estados, 5001, Santa Terezinha, Santo André/SP, CEP 09210-580
Bloco L, 3ºAndar, Fone (11) 3356-7617
iniciacao@ufabc.edu.br

Relatório Final de Iniciação Científica
referente ao Edital: 01/2019.

Título do projeto: Modelagem de tópicos com o algoritmo LDA aplicado a obras representativas das ondas do feminismo.

Palavras-chave do projeto: LDA, Modelagem de Tópico, PLN, Ondas do Feminismo.

Área do conhecimento do projeto: Processamento de Linguagem Natural.

Bolsista: Não.

Santo André

11/2020

Sumário

1 Resumo 2
2 Introdução 2
3 Fundamentação teórica 5
3.1 Modelagem de Tópicos 5
3.2 *Latent Dirichlet Allocation* 5
3.3 Base de Dados 7
4 Metodologia 9
4.1 Materiais e Métodos 9
4.2 Etapas Práticas da pesquisa 9
5 Resultados e discussão dos resultados 13
6 Conclusões e perspectivas de trabalhos futuros 16
Referências 17

1 Resumo

Este projeto visa utilizar o algoritmo LDA (*Latent Dirichlet Allocation*) em uma base de dados de arquivos-texto selecionados entre obras de autoras representativas das quatro ondas ou gerações do feminismo. O algoritmo será usado para gerar automaticamente tópicos dos conjuntos de palavras mais recorrentes nas obras analisadas. Com esse trabalho, verifica-se ser possível classificar as principais demandas e reivindicações históricas de cada autora ou de cada onda ou geração de autoras feministas.

2 Introdução

A estreita conexão entre linguagem e pensamento tem colocado as técnicas de Processamento de Linguagem Natural (PLN) no centro do debate sobre máquinas inteligentes (JURAFSKY, 2009).

Dentre os avanços tecnológicos que nos proporcionam uma infinidade de facilidades, a chamada era da informação permitiu um aumento significativo na criação de conteúdos tanto impressos como digitais: vivemos em meio a um oceano de informação, sem, no entanto, ter uma definição consensual da comunidade científica sobre o que é informação. Esses avanços tecnológicos, em grande medida benéficos, tornaram quase impossível para o ser humano processar manualmente toda a informação de forma que a automatização desse processo se tornou necessária.

A Linguagem Natural, linguagem usada pelo ser humano, é uma linguagem não estruturada e com diversas particularidades que dificultam seu processamento computacional. De acordo com Frege, significado não pode ser equacionado como representação (SEARLE, 1967) e há que se distinguir dois significados em uma expressão de linguagem natural: o significado literal e o significado do locutor da expressão, como no exemplo "o tempo hoje está maravilhoso" (HAUSSER, 2001). Não é tarefa simples, com o uso apenas de máquinas, extrair o significado mais adequado de uma palavra ou expressão a partir da análise de seu contexto, ou compreender palavras digitadas de forma errada e que seriam facilmente compreendidas por seres humanos. Além disso, especificidades de cada linguagem natural podem se tornar obstáculos até hoje intransponíveis para máquinas.

Na Inteligência Artificial, o Processamento de Linguagem Natural (PLN) tem desempenhado um papel fundamental de incentivo à descoberta de novas técnicas para implementar algoritmos voltados à compreensão automática de linguagem natural. Uma dessas técnicas, o Processamento Estatístico de Linguagem Natural (PELN), compreende abordagens estatísticas para o processamento automático de linguagem, incluindo modelagem probabilística, teoria da informação e álgebra linear (MANNING, 1999). Exemplo desta abordagem estatística é o LDA (*Latent Dirichlet Allocation*), um algoritmo generativo probabilístico, que a partir de um documento faz uma distribuição em forma de tópicos de acordo com suas palavras-chave e respectivas frequências de uso. Com essa distribuição é possível descobrir [computacionalmente] os principais assuntos tratados no documento analisado (BLEI, 2003) e atribuir rótulos que os identifiquem adequadamente.

O conceito de ondas do feminismo é caracterizado por momentos históricos relevantes para as reivindicações feministas e pela forma como essas reivindicações foram feitas (ROCHA, 2017). Até há pouco tempo apenas três ondas do feminismo eram consideradas, porém com a disseminação generalizada de acesso a redes

sociais e outros recursos computacionais, a literatura especializada já aponta diversas publicações caracterizadas como a quarta onda do feminismo (ROCHA, 2017).

Nesse contexto, usamos em nosso trabalho de IC obras de autoras reconhecidas internacionalmente que representam as quatro ondas do feminismo. Essa análise idealmente deveria nos permitir classificar claramente as principais demandas de cada autora, a respectiva onda feminista em que elas se encaixam e situá-las em seu momento histórico. Na prática, porém, verificou-se que vários tópicos de diferentes ondas se sobrepõem e que algumas demandas, como a luta por direitos civis, se estendem pelas quatro ondas com maior ou menor ênfase.

3 Fundamentação teórica

Neste capítulo vamos discutir técnicas de modelagem de tópicos e sua aplicação em textos representativos das diferentes ondas feministas.

3.1 Modelagem de tópicos

O volume de informação gerado na sociedade em diferentes formatos e plataformas e a propagação da informação que por muito tempo foi realizada em arquivos impressos tomaram uma proporção infinitamente maior após o surgimento da internet. A quantidade e a qualidade dessa informação demandam a busca de ferramentas de processamento automático para tornar viável seu entendimento.

Um tópico é a representação de um assunto ou tema, formado por uma combinação de palavras que têm certa probabilidade de ocorrerem juntas e de forma frequente com o intuito de determinar o conteúdo latente (oculto) em grandes arquivos de texto (BLEI, 2011). Em outras palavras, "tópico é definido como um conjunto de palavras que frequentemente ocorrem em documentos semanticamente relacionados" (FRIGYIK; KAPILA; GUPTA, 2010).

A visualização desse grande volume de informação por meio de tópicos facilita a compreensão do conteúdo processado e a modelagem de tópicos é uma alternativa eficaz para esse processamento. Para este trabalho a técnica algorítmica de modelagem de tópicos utilizada foi o *Latent Dirichlet Allocation (LDA)*.

3.2 *Latent Dirichlet Allocation*

O LDA é um modelo probabilístico generativo, ou seja, é um modelo que inicialmente gera de forma aleatória uma distribuição de probabilidades de tópicos presentes num corpus a partir de variáveis latentes. Nos passos seguintes, ajustes são feitos para que o algoritmo possa convergir a valores compatíveis com a frequência das palavras no corpus. Ao analisar um texto, se eliminarmos as palavras não significativas e destacarmos as palavras mais recorrentes, é possível descobrir superficialmente do que se trata o texto. O LDA captura computacionalmente esse processo intuitivo para o ser humano. Em seu pré-processamento são excluídos pronomes, artigos, preposições e outros termos que não são essenciais para o entendimento do conteúdo do documento. Com o texto reduzido pela eliminação dos termos não relevantes, o LDA executa o processo generativo associando palavras em um determinado grupo ou tópico (BLEI, 2011).

O LDA recebe esse nome pois usa a alocação latente de *Dirichlet*, que é uma distribuição de probabilidade cujo intervalo é ele mesmo um conjunto de distribuição de probabilidades. É muito usado na inferência bayesiana para descrever conhecimento a priori sobre a probabilidade de que certas variáveis aleatórias sejam distribuídas de acordo com uma ou outra distribuição específica (Wikipedia Dirichlet, 2020).

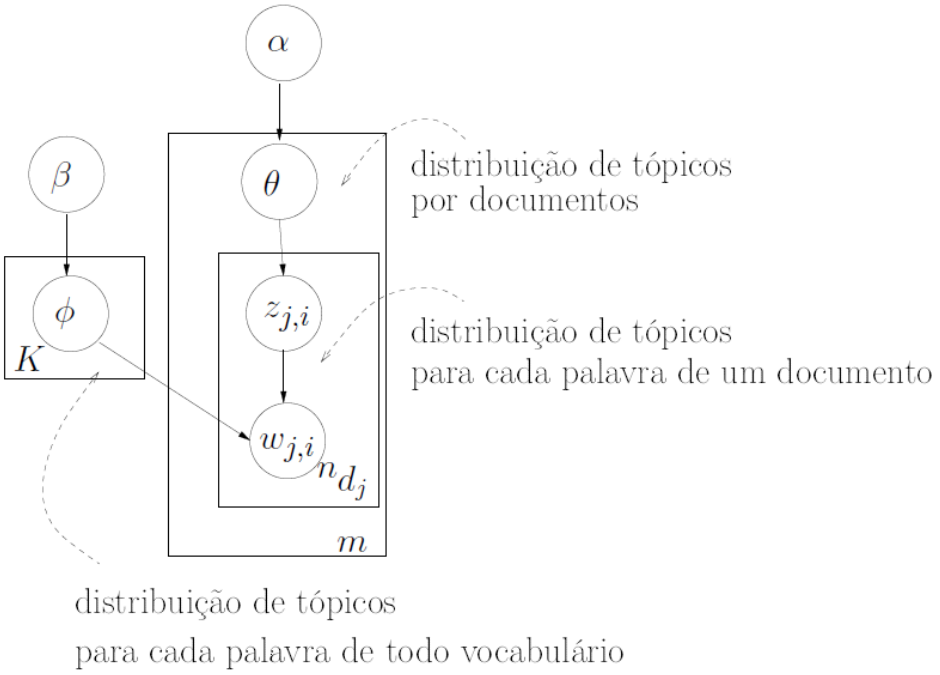
Segundo (FRIGYIK; KAPILA; GUPTA, 2010), na modelagem de distribuição de palavras em um documento podemos fazer a seguinte analogia: se tivermos um dicionário contendo *k* palavras possíveis, então um determinado documento pode ser representado por uma função de densidade discreta de comprimento *k* produzido pela normalização da frequência empírica de suas palavras. Um grupo de documentos produz uma coleção de funções de densidade discreta e podemos ajustar uma distribuição *Dirichlet* para capturar sua variabilidade. Diferentes distribuições de *Dirichlet* podem ser usadas para modelar documentos por diferentes autores ou documentos sobre diferentes tópicos.

Para o algoritmo LDA, cada documento é composto por um número de tópicos, 10, 15, 20 tópicos previamente definidos pelo usuário e cada tópico é constituído por um conjunto de palavras-chave, determinadas por um processo generativo.

O processo gerativo de parâmetros do algoritmo LDA abrange os seguintes passos:

- Para cada documento, selecione uma distribuição sobre tópicos;
- Para cada tópico, selecione uma distribuição sobre o vocabulário;
- Para cada palavra em cada documento,
    - Selecione um tópico;
    - Selecione uma palavra deste tópico.

A Fig. 1 ilustra os passos do processo gerativo de parâmetros do algoritmo LDA, sendo $w_{j,i}$ a representação das palavras dos documentos.



Fonte: (FALEIROS; LOPES, 2016).

Fig. 1 – Processo Gerativo de Parâmetros do Algoritmo LDA

3.3 Base de Dados

A história do feminismo pode ser dividida em períodos os quais são denominados "ondas". Cada período foi caracterizado por uma série de ações populares e reivindicações de movimentos feministas que marcaram de alguma forma a luta pelos direitos das mulheres. Para este projeto resolvemos usar como base de dados obras de autoras de projeção internacional que representassem cada um desses períodos.

Não existe consenso na literatura quanto ao momento exato em que surgiu o feminismo, porém compreende-se que a primeira onda surgiu no século XIX e durou até o início do século XX. Os movimentos desse período ocorreram em diversas partes do mundo, sendo importante lembrar que existiram antecessoras a esse estágio de luta e que elas são categorizadas como pré-feministas.

Para representar a primeira onda foram selecionadas duas autoras, Mary Wollstonecraft e sua obra "*A Vindication of the Rights of Woman*", que serviu de inspiração para o movimento das sufragistas, e as obras "*Marriage and love*", "*Woman Suffrage*" e "*The Tragedy of Woman's Emancipation*" da autora Emma Goldman.

A segunda onda ocorreu entre as décadas de 1960 até o início de 1980. Essa onda expandiu o debate dos direitos das mulheres englobando questões como sexualidade, mercado de trabalho, direitos legais, entre outros. As obras utilizadas para a segunda onda foram "O segundo sexo" da Simone de Beauvoir, "A mística feminina" da Betty Friedan e "The personal is political" da Carol Hanisch.

A terceira onda feminista teve início na década de 1990. A partir daqui os conceitos de feminismo estabelecidos começaram a ser debatidos de forma a abranger mulheres de várias condições sociais, visando diversificar as reivindicações para uma diversidade de mulheres de diferentes etnias, sexualidade e também mulheres trans. As obras escolhidas para esse período foram "Mulheres, Raça e Classe" da autora Angela Davis e "Problemas de Gênero" da Judith Butler.

A quarta onda surgiu a partir de 2012 e emergiu através de campanhas das redes sociais em oposição e exposição de casos de violência e assédio sexual, entre outros.

Para essa onda foram escolhidas quatro obras "O mito da beleza: Como as imagens de beleza são usadas contra as mulheres" da Naomi Wolf, "Má Feminista" da Roxane Gay, "O feminismo é para todo mundo" da Bell Hooks e "Men explain things to me" da autora Rebecca Solnit.

Todas as obras foram escolhidas com o critério de representação da época em que foram escritas e de sua influência, e foram baixadas em pdf na versão da língua inglesa.

4 Metodologia

A apresentação da Metodologia está dividida em duas partes: Materiais & Métodos e Etapas da Pesquisa.

4.1 Materiais e Métodos

A pesquisa se iniciou com as leituras técnicas sobre o LDA e paralelamente sobre as ondas do feminismo na Europa e nos Estados Unidos. Muitas das fontes consultadas aparecem neste relatório e estão listadas na seção Referências.

Foram reunidos quatro conjuntos de obras de textos feministas em formato pdf relativos às ondas 1, 2, 3 e 4 e convertidos para o formato Json, mais apropriado para o processamento dos dados com um programa em Python. O algoritmo LDA foi rodado diversas vezes até que alguns resultados práticos coerentes começassem a aparecer. O ajuste de inúmeros parâmetros mostrou-se bem mais trabalhoso do que supúnhamos. A literatura técnica menciona este aspecto computacional extremamente trabalhoso, assim como as dificuldades da etapa de pré-processamento ressaltando que a maior parte do esforço e de recursos serão usados nestas fases, mas raramente traz exemplos práticos realistas.

4.2 Etapas Práticas da Pesquisa

As subseções a seguir representam os passos da metodologia empregada.

4.2.1 Obras literárias escolhidas

Na tabela 1 podemos ver a relação das obras em inglês escolhidas, das autoras e dos respectivos períodos. As obras foram baixadas em formato de arquivo pdf, depois separadas em ondas e então convertidas em arquivos txt. Foi usada uma ferramenta online para converter os arquivos em txt em formato Json. Esse processo foi feito para facilitar a importação do arquivo de cada onda feminista no corpus utilizado.

Tabela 1

Período	Obra	Autora
Primeira Onda	A Vindication of the Rights of Woman	Mary Wollstonecraft
	Marriage and love	Emma Goldman
	Woman Suffrage	Emma Goldman
	The Tragedy of Woman's Emancipation	Emma Goldman
Segunda Onda	The Second Sex	Simone de Beauvoir
	The Feminine Mystique	Betty Friedan
	The personal is political	Carol Hanisch
Terceira Onda	Women, Race, & Class	Angela Davis
	Gender Trouble	Judith Butler
Quarta Onda	The Beauty Myth	Naomi Wolf
	Bad Feminist	Roxane Gay
	Feminism is for everybody	Bell hooks
	Men Explain Things To Me	Rebecca Solnit

4.2.2 Ambiente e Algoritmo

O ambiente escolhido para rodar o código foi o Jupyter Notebook. Essa plataforma permite unir texto e código, permitindo que cada funcionalidade seja detalhada e explicada. Além disso, é possível dividir o código em blocos e rodá-los separadamente, tornando a estrutura mais flexível e amigável para gerar gráficos em tempo real.

Quanto ao modelo do LDA foi utilizado um código em Python disponível na Internet, sendo necessárias apenas algumas alterações para adequar o script aos requisitos do projeto. O modelo utilizado foi extraído do site *Machine Learning Plus*, de um artigo intitulado "*Topic Modeling with Gensim (Python)*", de Selva Prabhakaran.

4.2.3 Pré-processamento dos dados

O pré-processamento consiste na importação e limpeza do texto, e é nessa etapa que retiramos todo o excesso de caracteres irrelevantes, tornando o texto mais fácil de processar. Podemos dividir esse processo nas seguintes fases:

- Retirada de caracteres especiais e de pontuação. Neste projeto também foram retirados os números porque eles não contribuem para definir o tópico sendo discutido. Para se obter os melhores resultados é preciso avaliar cada texto em questão para decidir o que deve ser retirado.

- Tokenização: consiste em segmentar parágrafos em frases e depois quebrar frases em palavras. Uma das técnicas computacionais utilizadas para interpretar um texto consiste em analisar a sequência de palavras.

- Lematização: deflexiona palavras, reduzindo-as ao lema.

- Remoção de StopWords: retira palavras vazias, sem conteúdo informacional para o significado de um texto, como artigos, pronomes, etc.

4.2.4 Aplicação do algoritmo LDA

A implementação do algoritmo LDA é feita através da criação de um corpus de textos e de um dicionário das palavras presentes nos textos. Para criá-los foi utilizada a biblioteca Gensim, que tem diversas funcionalidades ligadas à criação e ao treinamento de modelos para o algoritmo LDA.

A biblioteca Gensim possibilita a criação de um id para cada palavra do dicionário, facilitando o processamento numérico, em vez de simbólico, do corpus. Além disso, também podemos determinar livremente o número de tópicos e visualizar palavras-chaves e seus pesos (probabilidade) para cada tópico; quanto maior a probabilidade de uma palavra, mais importante é aquela palavra para determinado tópico. Complementando os recursos da biblioteca Gensim, foi usado o pacote pyLDAvis para plotar os resultados desses tópicos de forma gráfica.

5 Resultados e discussão dos resultados

Os resultados práticos obtidos até o momento são tecnicamente coerentes, mas ainda não nos permitem identificar claramente tópicos ligados às temáticas feministas de cada onda. O problema não parece ser do algoritmo LDA pois os tópicos encontrados efetivamente aparecem nas obras escolhidas. Na prática, o que se constatou é que muitos tópicos se repetem em diferentes ondas feministas, com maior ou menor ênfase. Ou talvez o problema possa ser a escolha do número de tópicos definidos: 20. Talvez se fossem definidos 3, 5 ou 6 tópicos a sobreposição de tópicos dentro de um mesmo corpus ou a repetição de tópicos em diferentes ondas não ocorresse. Há uma dificuldade prática e teórica em encontrar o número ideal de tópicos para um determinado corpus. Além disso, se num determinado corpus os tópicos forem claramente delimitados por conjuntos de palavras distintas, por tentativa e erro é possível encontrar estes tópicos. Mas, e se os tópicos de um determinado corpus não forem claramente delimitados, o que é possível fazer com o LDA? Possivelmente muito pouco, já que na base de dados não existe a clara separação que se busca.

Apresentamos abaixo o resultado parcial de uma de nossas inúmeras simulações para ilustrar o formato dos resultados produzidos pelo LDA:

[(0,

0.035*"word" + 0.031*"appear" + 0.026*"interest" + 0.025*"use" + 0.023*"real" + 0.022*"death" + 0.021*"offer" + 0.020*"satisfy" + 0.019*"moment" + 0.017*"physical"'),

(1,

0.064*"desire" + 0.038*"show" + 0.037*"individual" + 0.030*"put" + 0.027*"reality" + 0.025*"example" + 0.022*"hold" + 0.021*"career" + 0.020*"victim" + 0.017*"realize"'),

(2,

0.159*"would" + 0.110*"give" + 0.050*"sexual" + 0.033*"remain" + 0.022*"refuse" + 0.019*"history" + 0.017*"mistress" + 0.016*"failure" + 0.016*"soul" + 0.014*"reach"'),

(3,

0.469*"woman" + 0.027*"other" + 0.026*"seek" + 0.025*"sometimes" + '0.016*"social" + 0.014*"alone" + 0.013*"suffer" + 0.010*"least" + 0.008*"superiority" + 0.008*"hard"'),

(4,

0.117*"take" + 0.076*"try" + 0.033*"good" + 0.032*"lose" + 0.031*"destiny" + 0.030*"pleasure" + 0.024*"bring" + 0.023*"turn" + 0.016*"problem" + 0.015*"glory"'),

Para simplificar a visualização dos resultados, vamos suprimir os valores das porcentagens e escolher apenas 5 tópicos significativos, dentre 20, para cada uma das ondas feministas.

A tabela abaixo representa cinco dos tópicos mais significativos gerados pela leitura da **primeira onda feminista**. A partir dele podemos observar que esses tópicos refletem demandas relacionadas à participação política, relação marido-esposa, ao comportamento das mulheres entre outras características dessa onda.

Tópico 9	Tópico 10	Tópico 14	Tópico 17	Tópico 18
woman	make	woman	girl	man
work	husband	great	beauty	produce
world	moral	TRUE	early	form
law	order	understand	expect	effect
pay	place	human	day	sexual
year	slave	country	young	virtuous
suffrage	vice	soul	man	high
superior	depend	emancipation	continue	FALSE
political	build	strong	educate	reputation
class	easy	freedom	boy	act

A análise da **segunda onda feminista** retorna palavras como: fight, individual, class, emancipate, work, equal, sex, etc.

Essas palavras podem ser interpretadas por demandas da segunda onda como Igualdade e liberdade sexual.

Tópico 3	Tópico 5	Tópico 12	Tópico 14	Tópico 19
real	individual	lead	work	also
fight	idea	political	come	demand
personal	put	set	equal	sex
lot	old	thought	right	body
courage	class	send	realize	begin
feel	early	commitment	struggle	bring
situation	stand	essay	love	home
even	group	demonstrate	collect	possible
self	challenge	love	even	role
value	brain	emancipate	hope	claim

A **Terceira Onda Feminista** trata de diversas formas de opressão como gênero, raça e classe. No tópico 0 podemos ver a questão de gênero, no tópico 3 a questão de raça e no 13 podemos interpretar como a violência contra a mulher.

Tópico 0	Tópico 3	Tópico 5	Tópico 12	Tópico 13
gender	worker	law	female	sexual
sex	racism	language	male	rape
trouble	struggle	culture	lynch	position
notion	year	part	kind	call
binary	race	symbolic	process	assume
distinction	class	prior	year	masculine
subject	world	lesbian	production	begin
norm	campaign	reality	reveal	difference
femininity	historical	discourse	feel	violence
limit	issue	understand	turn	anti

Para a **quarta onda**, as distribuições de probabilidades dos cinco tópicos mais significativos foram:

Tópico 4	Tópico 11	Tópico 15	Tópico 17	Tópico 19
male	rape	female	give	woman
class	culture	sexual	magazine	percent
power	story	body	leave	issue
race	learn	part	talk	model
struggle	early	experience	put	report
fact	care	pleasure	skin	american
privilege	support	freedom	student	study
group	act	healthy	effect	fashion
character	set	future	idea	press
individual	fantasy	doctor	matter	suffer

Essa onda é a mais atual e isso se reflete nesses tópicos: demandas como o padrão de beleza imposto pela sociedade e sua relação com a mídia, o empoderamento feminino, assédio, entre outras coisas.

6 Conclusões e perspectivas de trabalhos futuros

Considerando a complexidade computacional do algoritmo LDA, sua compreensão demandou mais tempo e esforço do que previámos. Some-se a isto os contratempos relacionados à impossibilidade de trabalhar nas dependências e nas máquinas da universidade e também a sobrecarga que a interação virtual com o orientador significou. Nestas circunstâncias excepcionais que estamos enfrentando, consideramos que parte essencial do projeto de IC foi completada.

Como desafio teórico, estamos tentando desenvolver um método para determinar o número ideal de tópicos para um corpus em que não há uma delimitação clara entre os conjuntos de palavras. A ideia inicial é estabelecer eixos temáticos em torno dos quais haveria tópicos com maior ou menor aderência. Tendo estabelecido eixos temáticos de balizamento para tópicos semelhantes porém não idênticos, talvez seja possível estabelecer uma região de transição semântica entre os eixos que permita organizar a sequência de tópicos.

Como trabalho futuro para 2021, pretendemos submeter os resultados obtidos a revistas indexadas.

Referências

BEAUVOIR, Simone De. **The second sex**. 1. ed. United States of America: Vintage books, 2011. p. 1-873.

BIANCHINI, Leonardo. **Análise Exploratória dos tópicos no Stack Overflow usando LDA (Latent Dirichlet Allocation)**. 2018. Trabalho de conclusão de curso - UNIVERSIDADE FEDERAL DA FRONTEIRA SUL CAMPUS CHAPECÓ, Chapecó, SC - Brasil, 2018. Disponível em:

<https://rd.uffs.edu.br/bitstream/prefix/2096/1/BIANCHINI.pdf> Acesso em: 30/08/2020.

BIRD, S. et al. **Natural Language Processing with Python**. O'Reilly, Sebastopol, CA, EUA, 2009.

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. ISSN 1532-4435. Disponível em: <https://dl.acm.org/citation.cfm?id=944919.944937> Acesso em:15/04/2019.

BUTLER, Judith. **Gender Trouble**: Feminism And the Subversion of Identity. 1. ed. Reino Unido: Routledge, 1999. p. 1-272.

DAVIS, Angela Y.. **Women, Race, & Class**. 1. ed. New York: Vintage Books, 1981. p. 1-166.

FALEIROS, T. P.; LOPES, A. de A.. **Modelos probabilísticos de tópicos: desvendando o Latent Dirichlet Allocation**. 2016. – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP. Disponível em:

<http://repositorio.icmc.usp.br/handle/RIICMC/6647> Acesso em: 30/08/2020.

FRIEDAN, Betty. **The Feminine Mystique**. 1. ed. New York : W. W. NORTON & COMPANY New York London, 1963. p. 1-239.

FRIGYIK, B. A.; KAPILA, A; GUPTA, M. R. **Introduction to the Dirichlet Distribution and Related Processes**. UWEE Technical Report, University of Washington, 2010. Disponível em:

<http://mayagupta.org/publications/FrigyikKapilaGuptaIntroToDirichlet.pdf> Acesso em: 30/08/2020.

GAY, Roxane. **Bad Feminist**. 1. ed. United States: Corsair, 2014. p. 1-339.

GOLDMAN, Emma. **Marriage and love**: Emma Goldman's Anarchism and Other Essays. 2. ed. New York & London: Mother Earth Publishing Association, 1911.

GOLDMAN, Emma. **Woman Suffrage**: Emma Goldman's Anarchism and Other Essays. 2. ed. New York & London: Mother Earth Publishing Association, 1911. p. 201-217.

GOLDMAN, Emma. **The Tragedy of Woman's Emancipation**: Emma Goldman's Anarchism and Other Essays. 2. ed. New York & London: Mother Earth Publishing Association, 1911. p. 219-231.

HAUSER, R. **Foundations of Computational Linguistics**. Springer, Berlin, Alemanha, 2001.

HOOKS, Bell. **Feminism is for everybody**: Passionate Politics. 1. ed. Canada: Library of Congress, 2000. p. 1-136.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Second Edition, Pearson/Prentice Hall, Upper Saddle River, USA, 2009.

MANNING, C. D. e SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. The MIT Press, Cambridge, MA, EUA, 1999.

ROCHA, Fernanda de Brito Mota. **A quarta onda do movimento feminista: o fenômeno do ativismo digital**. Agosto 2017. p.137. Dissertação de Mestrado - UNISINOS, São Leopoldo, RS - Brasil, 2017. Disponível em:

<http://www.repositorio.jesuita.org.br/handle/UNISINOS/6728> Acesso em: 15/04/2019.

RODRIGUES, Ana. **O Livro do Feminismo**. GLOBO LIVROS, Porto Alegre, Rio Grande do Sul, 2019.

Wikipedia, Dirichlet. **Processo de Dirichlet**. Wikipedia. Disponível em: <https://pt.wikipedia.org/wiki/Processo_de_Dirichlet> Acesso em: 12/09/2020.

SEARLE, J. R. **Teoria da Comunicação Humana e a Filosofia da Linguagem**. In: Teoria da Comunicação Humana, Frank E. X. Dance (org.), Editora Cultrix, São Paulo, SP, Brasil, 1967.

SELAVA, Prabhakaran. Machine Learning Plus. **Topic Modeling with Gensim (Python)**. Disponível em:

<https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>. Acesso em: 30/09/2020.

SOLNIT, Rebecca. **Men Explain Things To Me**: And Other Essays . 1. ed. United States: [s.n.], 2008. p. 1-8.

Wikipedia, Dirichlet. **Processo de Dirichlet**. Wikipedia. Disponível em: <https://pt.wikipedia.org/wiki/Processo_de_Dirichlet> Acesso em: 12/09/2020.

HANISCH, Carol. "**The Personal is Politica**l". Disponível em <http://www.carolhanisch.org> Acesso em 24/11/2020.

WOLF, Naomi. **The Beauty Myth**: How Images of Beauty Are Used Against Women. 1. ed. New York: Harper Perennial, 2002. p. 1-368.

WOLLSTONECRAFT, Mary. **A Vindication of the Rights of Woman**: with Strictures on Political and Moral Subjects. 1. ed. [S.l.: s.n.], 1792. p. 1-260.