# Capítulo 4

# Modelagem de Tópicos e Criação de Rótulos: Identificando Temas em Dados Semi-Estruturados e Não-Estruturados

Diogo Nolasco e Jonice Oliveira

***Abstract***

*Recently, the increase in the flow and quantity of information stored resulted in a bigger demand for solutions to identify and to interpret this data. In this scenario, one of the existing challenges is the one of thematic or subject detection in large document collections. Topic Modeling is a technique that has a group of algorithms capable of extracting topics from documents, making both the theme identification in this corpora and its posterior grouping easier. On the other hand, the Topic Labeling helps in the theme recognition by the users, with methods to present them in a clear and intuitive way. Then, through topic modeling, it is possible to separate documents in topics and group them by subject, afterwards utilizing the labeling to extract a better representation for the found topics. In this tutorial, we present: i) The main methods used in both approaches, ii) how to use and combine them to help in the solution of real problems, exemplifying through examples in different scenarios (for example, scholar and social media scenarios), iii) and finally, the opportunities of application in various knowledge areas.*

***Resumo***

*Atualmente, o aumento do fluxo e da quantidade de informações armazenadas resultou em uma demanda por soluções para identificar e interpretar tais informações. Neste cenário, um dos desafios que existem é o de identificar temas ou assuntos em grandes coleções de documentos. A modelagem de tópicos é um conjunto de algoritmos capaz de extrair tópicos de documentos, visando à identificação desta coleção e facilitando o posterior agrupamento destes. Já a rotulagem de tópicos auxilia no reconhecimento de temas, provendo métodos para representá-los aos usuários de forma intuitiva. Então, através da modelagem é possível dividir documentos em tópicos e, após agrupá-los por tema, pode-se utilizar a rotulagem para extrair uma melhor representação dos grupos. Neste curso, apresentaremos: i) os principais métodos utilizados em ambas as abordagens, ii) como utilizá-los juntos para resolver problemas*

*reais, exemplificando-os em diferentes cenários (por exemplo, o cenário acadêmico e o sensoriamento participativo através das mídias sociais), iii) e por fim, as oportunidades de aplicação em diversas áreas.*

## 4.1. Introdução

Com o aumento da quantidade das informações geradas, torna-se cada vez mais difícil buscarmos ou interpretarmos os documentos armazenados. Para essa tarefa foram desenvolvidas técnicas probabilísticas chamadas de modelagem de tópicos, que são utilizadas para descobrir, extrair e agrupar documentos de grandes coleções em estruturas temáticas [Blei 2012].

Um dos primeiros algoritmos mais representativos para este fim foi descrito em [Papadimitriou et al. 1998] e, desde então, muitos outros foram criados com o intuito de otimizar a tarefa da extração de tópicos de uma coleção. Na modelagem, cada documento é representado como uma combinação de tópicos e cada tópico é representado por um conjunto de termos, ambos com probabilidades associadas. Assim, cada tópico extraído da coleção possui termos mais relevantes. Analogamente, cada documento possui tópicos mais relevantes de acordo com as respectivas probabilidades.

Por exemplo, dada uma coleção de documentos, podemos esperar que termos como "foguete" e "espaço" estejam associados em maior quantidade (maior probabilidade) ao tópico de "viagens espaciais" do que ao tópico "genética". Por outro lado, talvez "genética" contenha mais termos como "gene" e "DNA", enquanto termos gerais como "planeta", "isso" e "desde" teriam igual chance de aparecer em ambos os tópicos. Assim, cada tópico extraído poderia ser inicialmente representado por seus termos mais comuns e poderíamos agrupar documentos de acordo com o respectivo tópico.

Esses algoritmos são muito utilizados na organização de documentos textuais [Blei 2010]. Atualmente, seu uso tem-se expandido em outros cenários, como documentos acadêmicos e técnicos [Nolasco e Oliveira 2016], blogs [Mei 2006], notícias [Gao et al. 2012] e redes sociais [Hong e Davison 2010]. Recentemente, com a ampla utilização das mídias sociais e seu uso pela população para relatar problemas, opiniões e seus anseios, a modelagem de tópicos se tornou um poderoso aliado no sensoriamento participativo.

Após a aplicação da modelagem dos tópicos, normalmente o resultado é um conjunto de termos que nos indicam ou induzem ao(s) tema(s) ou assunto(s) de uma coleção. Entretanto, a apresentação de um conjunto de termos fora de seu contexto original, por vezes, pode dificultar o reconhecimento dos temas por pessoas não especializadas no domínio ou na coleção. Com isto surge a necessidade de uma maior interpretação semântica dos tópicos para melhor identificação de um tema. Para isso, utilizamos a rotulagem de tópicos, onde é possível definir cada tópico com um conjunto de termos mais explicativos.

A rotulagem de tópicos é uma técnica que permite exibir aos usuários os tópicos semanticamente mais coerentes, diminuindo a dependência de conhecimentos especializados (sobre o domínio ou coleção) necessários para a interpretação de tais tópicos.

Após utilizar um algoritmo para modelagem e extração dos tópicos da coleção, ainda temos o desafio de identificar o tema encontrado. Por si só, o algoritmo separa os documentos em tópicos e representa-os como uma lista de termos em ordem de relevância [Blei et al. 2003; Blei et al. 2010; Hoffmann 1999]. Para quem tem pouco conhecimento acerca dos temas presentes ou pouca familiaridade com os documentos, pode ser difícil identificar um tópico pela lista de termos. Logo, a rotulagem pode auxiliar nesta tarefa, produzindo termos mais compreensíveis aos usuários que estão explorando a coleção ou o tópico em si, como por exemplo, o título do documento mais relevante para aquele tópico [Manning et al. 2008].

Na modelagem, os tópicos são representados como distribuições de termos. Consequentemente, técnicas que tiram proveito dessa estrutura particular foram criadas para a formação de rótulos mais informativos. Para criar rótulos em grandes coleções, as técnicas variam do uso de uma simples matriz de termos-documentos [Papadimitriou et al. 1998] até o uso da estrutura do documento e a relevância dos termos para cada seção [Nolasco e Oliveira 2016].

Assim, este capítulo visa apresentar as mais eficientes e recentes técnicas de modelagem de tópicos e de rotulagem de tópicos existentes. Serão apresentados diversos exemplos reais que buscam ilustrar como podemos identificar temas em coleções de diferentes tipos. Também serão apresentadas aplicações e como utilizá-las com vários formatos de dados.

Este capítulo está organizado da seguinte forma: A seção 4.2 apresenta uma explanação e fundamentação teórica para a modelagem de tópicos, com suas subseções apresentando introdução, principais técnicas, aplicações e exemplos e desafios na área respectivamente. Já a seção 4.3 apresentará os conceitos envolvidos na rotulagem de tópicos, da mesma forma que a seção anterior com introdução, principais técnicas, exemplos e desafios apresentados nas suas subseções respectivamente. Por fim, a seção 4.4 apresenta as considerações finais sobre o assunto.

## 4.2. Introdução à modelagem de tópicos

A modelagem de tópicos, apesar de ser um método estatístico para descobrir temas na estrutura de um corpus, também é vista como uma "clusterização"[1] *fuzzy* ou *soft* [De Oliveira et al. 2007]. A "clusterização" de dados ou análise de agrupamentos é uma técnica de mineração de dados multivariados que através de métodos numéricos e a partir somente das informações presentes nos dados, tem por objetivo agrupar automaticamente

[1] "clusterização" é um aportuguesamento do inglês *clustering* e significa "agrupamento". Como o termo é amplamente utilizado no cenário de mineração de dados, será utilizado aqui como equivalente.

os dados de uma coleção em grupos geralmente disjuntos denominados clusters ou agrupamentos. É considerada uma técnica de aprendizado não supervisionado que geralmente envolve dois parâmetros básicos: N, um número de casos da base de dados (por exemplo, documentos) e K, o número de grupos (por exemplo, o número de temas existentes).

Diferente do conceito de classificação (técnica de aprendizado supervisionado), a "clusterização" é uma técnica mais "primitiva" onde não há nenhuma suposição a respeito dos grupos. Na classificação, existem classes predefinidas e através de um treinamento com exemplos de execução, os algoritmos "aprendem" como alocar os dados em cada classe, daí o nome aprendizado supervisionado. A "clusterização", ao contrário, não conhece de antemão as classes existentes e nem possui exemplos de como distribuir os dados entre os grupos, por isso realiza um aprendizado não-supervisionado.

Os tópicos extraídos pela modelagem podem ser então vistos como clusters e os dados agrupados como os casos. Além disso, pode-se dividir a "clusterização" em dois tipos principais: *hard clustering* e *soft clustering* [Arabie et al. 1996]. O primeiro é o mais usual onde cada caso é associado a um e somente um grupo. Já o último, onde se encaixa a modelagem de tópicos, pode atribuir a cada caso um ou mais grupos com diferentes proporções (que no caso da modelagem é representado pela probabilidade de cada grupo).

Assim, a modelagem probabilística de tópicos é uma abordagem para atacar o problema do agrupamento e organização de dados, principalmente de conteúdo textual e cujo objetivo principal é a descoberta de tópicos e a anotação de grandes coleções de documentos por classificação temática. Tais métodos analisam quantitativamente as palavras dos textos originais para descobrir os temas presentes nos mesmos. Os algoritmos de modelagem de tópicos não requerem nenhum conhecimento prévio dos elementos e os tópicos emergem da análise dos textos originais [Blei 2012].

### 4.2.1. Principais Técnicas

A pesquisa em modelagem de tópicos a partir de documentos de textos teve inicialmente um marco com o desenvolvimento da técnica conhecida como Análise de Semântica Latente (*Latent Semantic Analysis* ou LSA) [Landauer e Dumais 1997]. No LSA, utilizou-se do ferramental da álgebra linear para decompor um corpus nos seus temas constituintes, mais especificamente através da aplicação da decomposição SVD (*Singular Value Decomposition*) numa matriz com a contagem de frequência dos termos ao longo dos documentos de uma coleção. Na área de pesquisa em recuperação de informações, o LSA é utilizado para retornar documentos correspondentes a partir de uma busca por palavras-chave, categorizar documentos e generalizar resultados através de documentos equivalentes em diversas línguas [Chang et al., 2009]. Na modelagem de tópicos, o modelo LDA (Alocação Latente de Dirichlet, do inglês *Latent Dirichlet Allocation*) [Blei 2012] é um dos mais populares e serviu como base para a criação de muitos outros modelos probabilísticos. As fundações do modelo LDA foram baseadas no LSA e pLSI

(*Probabilistic Latent Semantic Indexing*, uma evolução do LSA com o uso de fórmulas probabilísticas [Steyvers e Griffiths 2007; Blei e Lafferty 2009].

No caso dos algoritmos de modelagem de tópicos, a abordagem baseia-se em criar uma distribuição de grupos para cada termo de um documento textual e uma distribuição de grupos para cada documento. Baseado nessas distribuições pode-se agrupar os documentos de acordo com as probabilidades associadas a cada grupo. Um exemplo deste tipo de técnica é ilustrado na Figura 4.1, que mostra quatro documentos associados aos tópicos "Genética", "Evolução", "Doenças" e "Computadores" (A largura das arestas que conectam os documentos aos grupos indicam a proporção do tópico presente no documento).

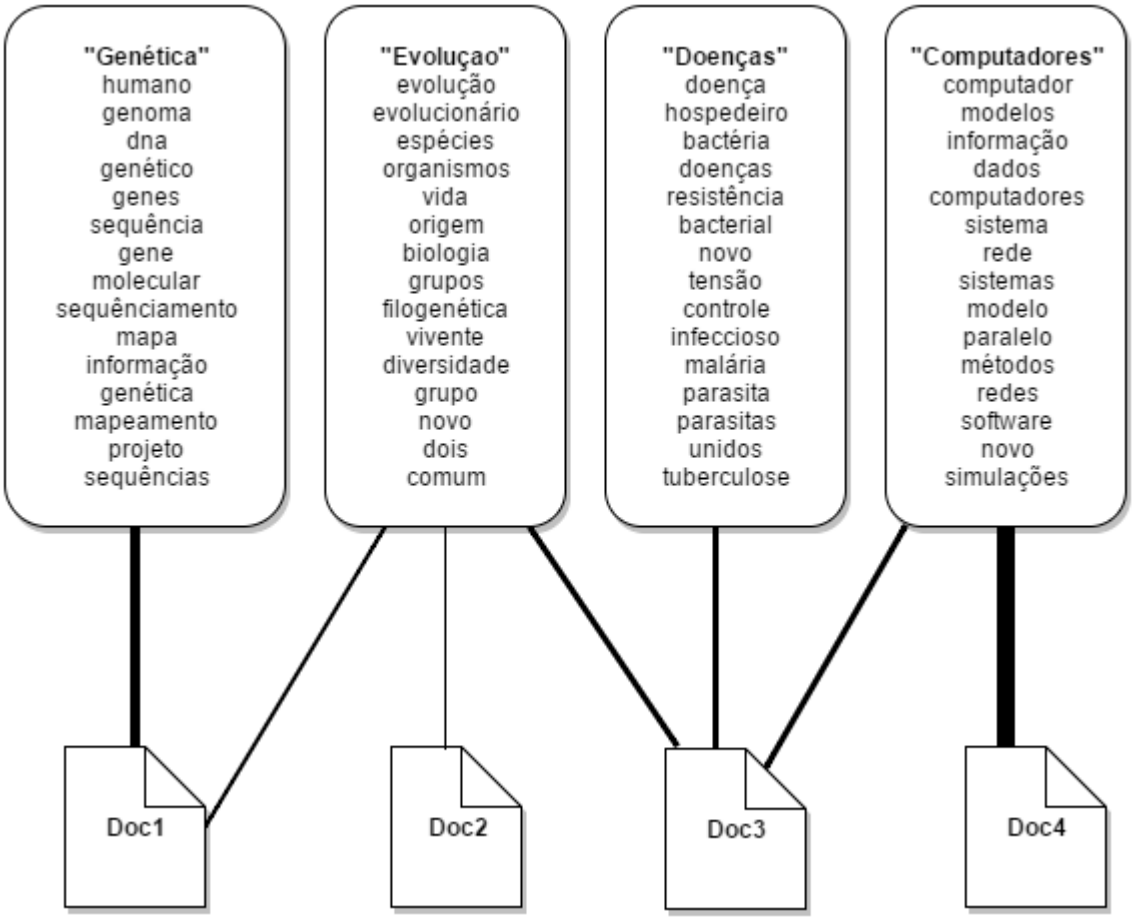


**Figura 4.1. Exemplo de associação entre tópicos e documentos**

A Alocação Latente de Dirichlet (LDA) e outros modelos de tópicos fazem parte do campo de pesquisa mais amplo em modelagem probabilística. Nesse tipo de modelagem, os dados são tratados como oriundos de um processo generativo que contém variáveis ocultas. Esse processo define uma distribuição de probabilidade conjunta sobre as variáveis aleatórias observadas e as ocultas, a qual é usada para computar a distribuição

condicional das variáveis ocultas dadas as variáveis observadas. Essa distribuição condicional também é chamada de distribuição posterior ou simplesmente "posterior". LDA se encaixa nesse *framework*. As variáveis observadas são as palavras nos documentos e as variáveis ocultas são a estrutura de tópicos (como mostra a Figura 4.2). O problema computacional de inferir a estrutura de tópicos oculta a partir de um conjunto de documentos é o problema de computar a distribuição posterior – a distribuição condicional das variáveis ocultas dados os documentos [Blei 2012].

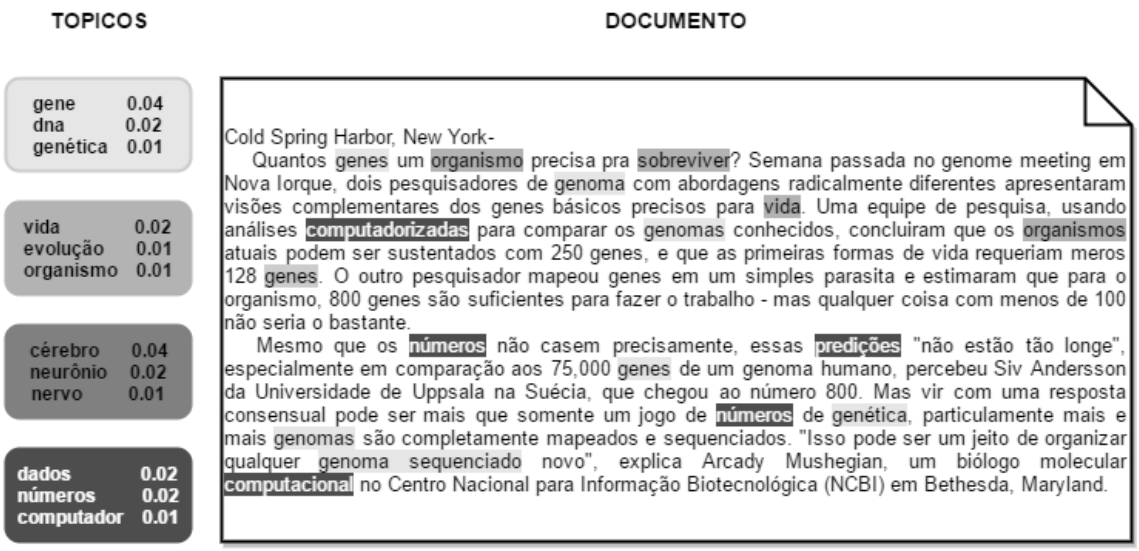


**Figura 4.2. Relação entre os termos contidos no documento e os tópicos existentes (A cor de fundo representa a ligação da palavra ao tópico correspondente)**

O processo generativo em LDA produz documentos de texto e os dados manipulados são as palavras que irão formar esses documentos. Trata-se de um processo imaginário, a partir do qual a estrutura de tópicos de uma coleção é obtida por inferência a partir da inversão daquele processo. Tecnicamente, o modelo assume que os tópicos são gerados antes dos documentos. Um tópico é definido como uma distribuição de probabilidade sobre um vocabulário fixo. Como exemplo, um tópico sobre genética será aquele que contém palavras relacionadas à genética com maior probabilidade de ocorrência. Em contraposição, um tópico que se relacione com qualquer outro assunto distinto conterá palavras sobre genética com probabilidade de ocorrência muito baixa ou zero. Todos os tópicos contêm distribuições com probabilidades sobre todo o vocabulário fixo, mas essas probabilidades só assumirão valores mais altos nos termos relacionados ao tópico.

O processo que gera os documentos em LDA é realizado em duas etapas. Para a geração de cada documento da coleção, tem-se que:

1. Uma distribuição sobre tópicos é escolhida aleatoriamente. Exemplo: num modelo com apenas 3 tópicos, uma distribuição sobre tópicos possível para um documento A pode exibir probabilidades 0.1, 0 e 0.9 de ocorrência dos tópicos x, y e z respectivamente.

2. Para cada palavra no documento:
   a. Um tópico é escolhido aleatoriamente a partir da distribuição obtida no passo 1.
   b. Uma palavra é escolhida aleatoriamente a partir do tópico (o qual é uma distribuição de probabilidade sobre o vocabulário) obtido em 2a.

Cada documento exibe tópicos em proporções distintas (passo 1), cada palavra em cada documento é obtida a partir de um dos tópicos (passo 2b), o qual por sua vez é escolhido a partir da distribuição sobre tópicos de um documento em particular (passo 2a). Esse modelo estatístico reflete a intuição de que documentos exibem múltiplos tópicos, um pressuposto que está por trás da formulação do modelo LDA.

O modelo LDA também pode ser descrito mais formalmente através da seguinte notação:

1. Dado os tópicos $\theta_{1:N}$, onde cada $\theta_n$ é uma distribuição sobre o vocabulário V.
2. As proporções dos tópicos para o d-ésimo documento são $\rho_d$, onde $\rho_{d,n}$ é a proporção do tópico n no documento d.
3. As atribuições de tópicos para o d-ésimo documento são $z_d$, onde $z_{d,i}$ é a atribuição do tópico para a i-ésima palavra no documento d.
4. Finalmente, as palavras observadas para o documento d são $w_d$, onde $w_{d,i}$ é a i-ésima palavra no documento d, a qual é um elemento do vocabulário V.

Com essa notação, o processo generativo em LDA corresponde à distribuição conjunta das variáveis observadas e ocultas representada pela expressão:

$$p(\theta_{1:N}, \rho_{1:D}, z_{1:D}, w_{1:D}) =$$
$$\prod_{j=1}^{N} p(\theta_d) \prod_{d=1}^{D} p(\rho_d) \left(\prod_{i=1}^{I} p(z_{d,i}|\rho_d) p(w_{d,i}|\theta_{1:N}, z_{d,i})\right) \quad (1)$$

Ainda que a modelagem LDA esteja no ativo campo de pesquisa em modelagem probabilística de tópicos e possibilite a segmentação automática de coleções de milhares de documentos, o que de outra forma não seria possível alcançar por anotação humana, é

preciso cautela no uso e interpretação dos resultados obtidos a partir desse modelo. Os tópicos e sua distribuição ao longo dos documentos obtidos a partir da modelagem LDA e de outros modelos de extração de tópicos não são "definitivos". O ajuste de um modelo de tópicos a uma coleção sempre irá produzir padrões a partir do corpus, ainda que os mesmos não estejam "naturalmente" presentes na coleção. Por isso é importante a utilização em conjunto com outros métodos que evidenciem claramente os assuntos presentes na coleção, como será visto mais adiante com a rotulagem dos tópicos. Portanto, modelos de tópicos devem ser vistos como uma ferramenta útil para exploração de dados, onde os tópicos provêm um resumo do corpus que seria impossível de obter manualmente. De qualquer forma, a análise de um modelo de tópicos pode revelar conexões entre documentos e no interior dos mesmos que não seriam óbvias a olho nu e pode ainda encontrar coocorrências inesperadas entre termos [Blei e Lafferty 2009].

### 4.2.1. Aplicações

Os algoritmos de modelagem de tópicos são bem flexíveis quanto ao seu uso, permitindo a identificação de temas em qualquer tipo de dados textuais e até mesmo em dados genéticos [Pritchard et al. 2000] e de imagens [Bart et al. 2011]. Sejam em artigos, blogs ou até mesmo mensagens é possível descobrir a estrutura temática que permeia os dados.

A Tabela 4.1 ilustra tópicos extraídos de uma coleção de artigos da conferência Knowledge Discovery and Data Mining (KDD) [Nolasco e Oliveira 2016]. Normalmente os tópicos são representados utilizando-se a lista de termos com probabilidade maior relacionado ao tópico.

Quanto maior a relação com o tema, maior é a probabilidade do termo estar relacionado ao tópico. Assim, na Tabela 4.1, o tópico 2 por exemplo tem como primeiro termo "otimização", portanto, sempre que um documento contiver esse termo sua associação com o tópico 2 aumenta. Seguindo o mesmo raciocínio, os termos de outras áreas ou assuntos apresentarão uma probabilidade ínfima se não houver relação temática com o tópico. Os termos comuns que estão presentes em todos os tópicos (como *stop-words*) normalmente serão diluídos pelas probabilidades, pois não são específicos de nenhum tema em particular.

**Tabela 4.1. Principais termos de quarto tópicos extraídos dos artigos do KDD**

| Artigos – KDD | | | |
|---|---|---|---|
| **Tópico 1** | **Tópico 2** | **Tópico 3** | **Tópico 4** |
| identify | optimization | knowledge | system |
| disease | methods | accuracy | mining |
| medical | proposed | detection | management |
| identifying | formulation | sample | techniques |
| health | show | available | analysis |
| study | functions | standard | systems |
| records | solve | given | application |
| clinical | linear | requires | large |
| features | regression | work | designed |
| patients | propose | performance | high |

Outro exemplo de um cenário totalmente distinto pode ser visto na Tabela 4.2. Neste caso, tem-se uma comparação entre tópicos extraídos da mídia comum (artigos do New York Times, no exemplo em questão) e da mídia social (Twitter) (Adaptado de [Zhao et al. 2011]). Apesar da natureza textual distinta, pode-se ver que foi possível extrair tópicos significantes de ambas as fontes. As redes sociais, mensagens, e-mails e qualquer tipo de fonte cujo conteúdo textual seja informal e sujeito a muita variação é um desafio para a modelagem de tópicos, porque esta necessita de uma consistência textual nos dados. Porém, ao longo do tempo, pode se dizer que todos esses meios de comunicação desenvolveram formas consistentes de transmitir sua mensagem, seja através de *tags* ou de vocabulário específico, abreviações, gírias, ainda há um padrão em sua comunicação e portanto a extração é possível, visto que é utilizado um método não-supervisionado.

Atualmente, estão disponíveis em diversas linguagens de programação, bibliotecas que implementam algoritmos de modelagem de tópicos para a extração de tópicos como nos exemplos vistos. Dentre elas, pode-se destacar a "lda-c"[2] para a linguagem C, "mallet"[3] para Java e "gensim"[4] para Python, porém muitas outras linguagens já possuem os algoritmos implementados. Devido ao seu caráter

[2] Disponível em: https://www.cs.princeton.edu/~blei/lda-c/index.html
[3] Disponível em: http://mallet.cs.umass.edu/
[4] Disponível em: https://radimrehurek.com/gensim/

probabilístico, não é necessário realizar nenhum pré-processamento complexo no corpus para a utilização, bastando uma simples tokenização do conteúdo textual. As palavras comuns são irrelevantes para o resultado da execução e quaisquer demais processamentos podem ser utilizados como forma de limpeza ou de melhorar desempenho, porém opcionais.

**Tabela 4.2. Exemplos de tópicos extraídos de mídias tradicionais e sociais**

| New York Times | | Twitter | |
|---|---|---|---|
| **Tema** | **Tópico** | **Tema** | **Tópico** |
| **Arte** | book,novel,story,life,writes<br>world,century,history,culture<br>art,museum,exhibition<br>war,history,world,civil,time | **Arte** | rob,moon,love,twilight<br>gaga,lady,#nowplaying<br>adam,lambert,fans,kris<br>chirs,brown,song,beyonce<br>download,live,mixtape,<br>music |
| **Negócios** | cars,car,ford,toyota,vehicles<br>media,news,magazine,ads | **Negócios** | #ebay,auction,closing<br>#jobs,job,#ukjobs |
| **Educação** | project,money,group,center<br>percent,study,report,rate | **Família & Vida** | dog,room,house,cat,door,<br>good,night,hope,tonight<br>life,#quote,success,change<br>god,love,lord,heart,jesus<br>smiles,laughs,hugs,kisses |
| **Estilo** | french,paris,luxury,swiss,<br>watch | **Twitter** | tweet,follow,account<br>lmaoo,smh,jus,aint,lmaooo |
| **Ciência** | space,moon,station,spirit,<br>earth | | |
| **Mundo** | case,charges,prison,trial,court<br>officials,announced,news,<br>week,<br>department,agency,federal,<br>law,south,north,korea,korean,<br>power | | |

### 4.2.1. Desafios e Oportunidades

A superação das limitações do modelo LDA é uma área de pesquisa ativa e duas abordagens importantes no desenvolvimento de novos modelos de tópicos podem ser destacadas: a criação de novos modelos através do relaxamento de alguns pressupostos assumidos em LDA e a incorporação de metadados do corpus para enriquecer a

modelagem dos documentos. Descrevendo com mais detalhes pressupostos que motivaram o desenvolvimento de modelos estendidos de tópicos os quais serão apresentados a seguir, um primeiro pressuposto assumido na modelagem LDA está relacionado ao conceito de "*bag of words*", o qual parte do princípio de que a ordem das palavras num documento não é relevante. Apesar de não ser um pressuposto realista, ele é razoável se o único objetivo da aplicação for revelar a estrutura semântica de textos. Um segundo pressuposto assumido em LDA é o de que a ordem dos documentos não importa. Esse pressuposto pode não ser realista ao se analisar coleções que atravessam anos ou séculos, pois nesses casos é importante considerar que existem alterações nos tópicos ao longo do tempo. O terceiro pressuposto em destaque é o de que o número de tópicos é conhecido e fixo. Na modelagem LDA, um dos parâmetros que deve ser definido a priori é justamente o número de tópicos a serem extraídos. O quarto e último pressuposto considerado é o de que os tópicos são independentes, o que impede de modelar a correlação entre os mesmos [Blei 2012].

Outra consideração, agora de ordem prática, para a execução da modelagem é a atualização do modelo. Uma vez que se extrai os tópicos, para qualquer mudança nos documentos é necessária uma nova execução do algoritmo. Uma segunda consideração é o tamanho do vocabulário. Quanto maior o número de termos utilizados, mais processamento será necessário para a execução. Técnicas que removam termos insignificantes, como TF-IDF ou remoção por frequência nominal podem ser úteis para otimizar o desempenho. Uma última consideração é a escolha do número de tópicos. Esta é uma variável de entrada para a modelagem, porém muitas vezes a quantidade de temas presentes na coleção é desconhecida por parte dos usuários. O número de tópicos procurado pode influenciar muito a saída do algoritmo, um número pequeno de tópicos pode forçar a fusão indesejada de temas e um número grande pode deixá-los com uma granularidade muito fina [Berry 2010].

Para os pressupostos, existem linhas de pesquisa que continuam expandindo a capacidade de remover ou contornar as limitações. Para o primeiro, temos em [Wallach 2006] que ao invés de usar unigramas na modelagem insere bigramas, assim, de certa forma se considera a ordem das palavras no resultado. Já [Wang 2007] insere n-gramas através de bigramas para o mesmo fim. O segundo pressuposto abriu a pesquisa de modelagem de tópicos dinâmica, que visa tanto modelar quanto descobrir temas presentes na coleção ao longo do tempo [Blei e Lafferty 2006]. Para o terceiro, que também apresenta um problema nas considerações, pode-se destacar a criação de algoritmos para modelagem de tópicos hierárquica [Teh et al. 2006; Pujara e Skomoroch 2012], a qual não necessita de um número de tópicos como entrada da mesma maneira que a "clusterização" hierárquica não necessita do número inicial de clusters a priori. Outra forma de evitar o problema de ter que "adivinhar" o número de tópicos presentes na coleção é o uso da análise de estabilidade que é capaz de verificar qual o número de tópicos mais estável para a coleção, ou seja. Com quantos tópicos perturbações nos dados não geram grandes perturbações nos resultados [Greene et al. 2014]. O quarto pressuposto

abre a pesquisa de modelagem de tópicos correlacionados [Blei e Lafferty 2007] que visa extrair ou descobrir correlações entre tópicos utilizando o fato da modelagem ser considerada uma "clusterização" *fuzzy*. Por fim, [Nolasco 2016] aborda os três últimos pressupostos criando um processo capaz de identificar o número de tópicos ótimo para a coleção (terceiro pressuposto), identificar a evolução de tópicos de pesquisa ao longo do tempo (segundo pressuposto) e correlacionar tópicos de pesquisas multidisciplinares (quarto pressuposto) no âmbito acadêmico e com grandes massas de dados.

As considerações de ordem prática, visando melhor desempenho e praticidade, resultaram na criação de algoritmos paralelos e que podem trabalhar com fluxo de dados atualizando-se continuamente, como em [Hoffman et al. 2010], que apresenta o "Online LDA" que é capaz de processar um grande fluxo de informações inclusive em *stream*.

Outras formas de melhorar os resultados do algoritmo é combiná-lo com outras técnicas ao invés de modificar o algoritmo em si, já que realizar processamentos repetitivos não é desejado computacionalmente. Uma dessas maneiras de tornar os resultados mais relevantes é o uso da rotulagem dos tópicos que é um assunto apresentado na próxima seção.

## 4.3. Rotulagem de Tópicos

A rotulagem de tópicos é uma técnica que permite exibir aos usuários os tópicos semanticamente mais coerentes, diminuindo a dependência de conhecimentos especializados (sobre o domínio ou coleção) necessários para a interpretação de tais tópicos.

Após sabermos o número de áreas existentes na coleção podemos realizar o agrupamento nestas áreas e gerar uma representação por tópico da coleção.

Normalmente, o resultado do agrupamento representa cada grupo com uma distribuição probabilística das palavras mais relevantes para cada um. Um desses resultados pode ser visto na Figura 4.3.

Podemos utilizar essa lista para que o usuário entenda o assunto e consequentemente a área descrita, o que atualmente é feito na maioria dos trabalhos da literatura [Blei 2012; Blei et al. 2003; Papadimitriou 1998]. Outra maneira de se descrever a área é utilizar termos que a expressem ou conceitos intimamente relacionados com ela. Esses termos podem ser palavras específicas (bioinformática, "clusterização"), pequenas frases com duas a três palavras (redes sociais, mineração de dados) ou até mesmo sentenças (Teoria dos Dois Fatores de Frederick Herzberg). Classificadores humanos frequentemente preferem o uso de frases de duas palavras [Chang et al. 2009].

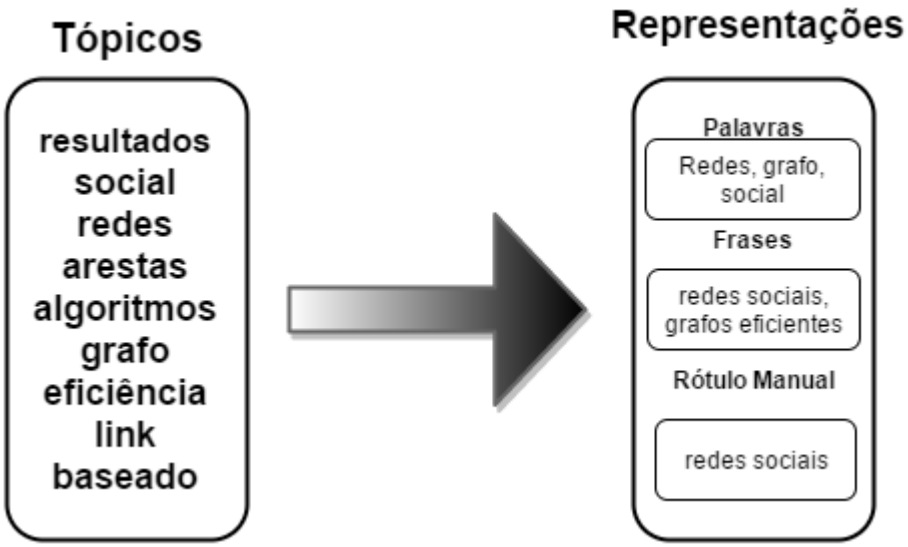


**Figura 4.3. Exemplo de um tópico e suas possíveis representações.**

O uso da lista resultante do agrupamento muitas vezes é útil para a identificação do assunto. Porém, exige familiaridade com cada área e com o domínio da coleção. Alguém que não domine a temática do domínio pode encontrar dificuldades em interpretar a lista e identificar os conceitos presentes, o assunto principal ou ligar as palavras para formar termos significantes que representem o assunto em questão. Isso é comum principalmente em áreas de pesquisa ou domínios técnicos onde comumente um tópico é facilmente reconhecido por um grupo familiarizado enquanto dificilmente será reconhecido por outros que não trabalhem especificamente no tema.

Um exemplo pode ser visto na Figura 4.3 que mostra uma lista resultante com palavras de uma área em ordem de relevância. Para pessoas da área de computação, principalmente os que trabalhem em áreas relacionadas a sistemas de informação e bancos de dados, pode-se inferir que se trata de documentos relacionados à mineração de dados devido à presença dos termos “mineração” e “algoritmos”. Para quem não vêm da área e portanto não está acostumado a usar essa terminologia pode ser difícil relacionar os termos presentes na lista porque cada um contém mais de um sentido dependendo do contexto. Por exemplo, “mineração” pode ser usado tanto no contexto de dados quanto no geológico, econômico ou militar. Outros termos como “modelo” e “sistema” sozinhos não transmitem muita informação útil porque isolados se tornam genéricos.

Assim, o desafio da geração de rótulos para os tópicos extraídos é de representar cada tópico de forma automática ao usuário de maneira que se identifique melhor o assunto em questão. Auxiliando na interpretação tanto ajudando o trabalho de quem conhece o domínio quanto facilitando a definição para quem não é familiarizado ao tema.

### 4.3.1. Principais Técnicas

A maioria dos trabalhos que utilizam a modelagem de tópicos para agrupamento usam a distribuição de termos de cada tópico extraído como própria representação [Blei 2012; Hoffman 1999].

Como a lista necessita de uma interpretação que por vezes não é trivial por parte dos usuários, outra opção utilizada é deixar o processo de rotulagem nas mãos de especialistas capazes de gerar rótulos manualmente de acordo com a lista ou com os documentos contidos [Chang 2009]. Essa é uma opção bem confiável visto que os especialistas têm o conhecimento necessário para interpretar e transmitir de forma correta os conteúdos, contudo também não está isenta de interpretações particulares dependendo do *background* do especialista e dificilmente pode ser aplicada em ambientes com grandes volumes de dados e de assuntos correlacionados, além de ser custosa e demandar bem mais tempo.

Outras técnicas utilizam abordagens semi-supervisionadas para criação dos rótulos. Nessas técnicas, normalmente o sistema gera rótulos genéricos ou simples e vai refinando o resultado com a ajuda humana.

Como exemplo, temos o uso combinado de classificação [Lau et al. 2011; Ramage et al. 2011] que utiliza a modelagem de tópicos como passo não-supervisionado e a própria classificação como passo supervisionado resultando numa abordagem semi-supervisionada. Neste caso, após o agrupamento o sistema gera os rótulos automaticamente baseado em um treinamento da coleção realizado pelos especialistas, que por sua vez devem conhecer as classificações possíveis para o efetivo uso da técnica.

Outra abordagem do tipo seria o aprendizado ativo [Downey et al. 2014], onde o sistema extrai termos para representar a área de forma simples (por exemplo usando algum termo da lista dos mais relevantes) e os especialistas dão um retorno ao sistema de quão bom está aquele rótulo ou melhorando-o e assim se vai modificando o rótulo até que esteja satisfatório.

Portanto, as principais técnicas existentes envolvem o uso de abordagem não-supervisionada, manual ou semi-supervisionada. As não-supervisionadas visam tornar o processo mais rápido as custas da falta do conhecimento especializado. As manuais são as tradicionais, que teoricamente dão melhores resultados utilizando mais recursos tanto pessoal quanto de tempo e as semi-supervisionadas tentam aliviar os problemas das manuais reduzindo o montante de trabalho especialista necessário através da introdução de um passo automático antes do trabalho manual.

Além dos métodos semi-supervisionados, existem também os que não necessitam de conhecimento prévio para sua execução. As principais abordagens totalmente automáticas existentes usam a própria lista para a geração dos termos (por exemplo, os 10 termos mais relevantes) [Lau et al. 2010] ou utilizam alguma estatística dentre todas

as palavras presentes na coleção [Mei et al. 2007] ou uma combinação de ambos [Nolasco e Oliveira 2016].

Para a maioria das principais técnicas existentes é necessário realizar um processo para a criação dos rótulos. Embora este processo tenha métricas ou formas diferentes, sua estrutura está presente nas principais técnicas existentes. A seguir, são apresentadas nas próximas seções uma fundamentação teórica desse processo geral com as principais técnicas utilizadas em cada etapa do processo.

### 4.3.2. Definições

Dada uma coleção de documentos C = $\{d_1, d_2, \dots, d_{|C|}\}$, onde $d_i$ é o documento número $i$, um vocabulário V = $\{w_1, w_2, \dots, w_{|V|}\}$ onde $w_j$ é o termo número j do vocabulário da coleção e um conjunto de tópicos extraídos de C, o objetivo da rotulagem é gerar rótulos compreensíveis para cada tópico que facilitem o entendimento da área.

**DEFINIÇÃO 1.** Um tópico $\theta$ de C é uma distribuição de probabilidades de termos tal que $\theta = \{p(w_1|\theta), p(w_2|\theta), \dots, p(w_{|V|}|\theta)\}$ e $\sum_{w \in V} p(w|\theta) = 1$. Assim, termos mais relevantes para a área teriam maior probabilidade e termos comuns para todas as áreas baixas probabilidades.

**DEFINIÇÃO 2.** Um rótulo l de $\theta$ é uma palavra ou conjunto de palavras que expressam o conteúdo de $\theta$. Por conseguinte, temos que é possível haver mais de um rótulo possível para cada área já que qualquer palavra usada para exprimir seu conteúdo pode ser utilizada. Isso é visível quando utilizamos sinônimos, embora sejam termos diferentes eles podem ser usados para representar a mesma coisa sem perda de informação.

Finalmente, para selecionarmos rótulos para $\theta$ podemos dividir o processo nas seguintes etapas:

1. Identificar um conjunto de candidatos $L = \{l_1, l_2, \dots, l_n\}$;
2. Calcular S(l, $\theta$), onde S é uma função da relevância do rótulo l para $\theta$;
3. Ordenar os candidatos baseado na função S;
4. Selecionar o(s) rótulo(s) mais relevante(s) para $\theta$ da lista ordenada;

Ao final, temos para cada tópico sua representação por meio de rótulos através desses passos.

### 4.3.3. Processo de geração de rótulos

De acordo com as etapas para criação de um rótulo podemos dividir o processo em três subtarefas principais: Seleção de candidatos, Ranqueamento, e Seleção de rótulos. A Figura 4.4 ilustra esse processo.

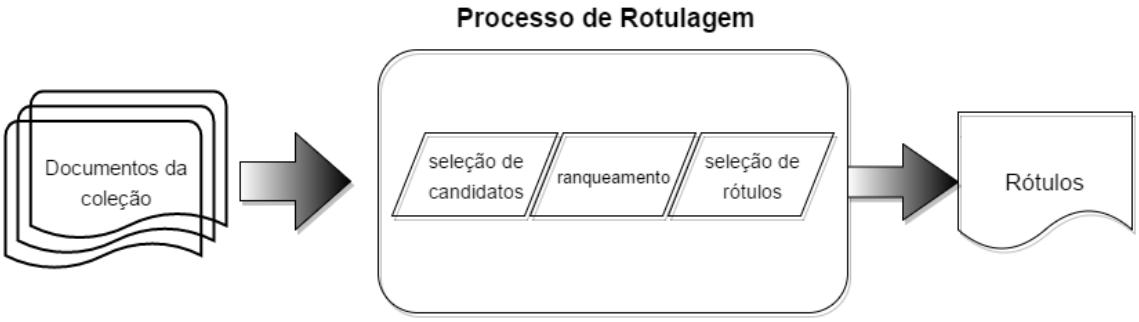


**Figura 4.4. Etapas do processo de Rotulagem**

Assim, selecionamos termos candidatos dos documentos, ordenamo-los por ordem de relevância (segundo uma função para este fim) e selecionamos os melhores. Esses três passos são descritos em detalhes a seguir.

#### 4.3.3.1 Seleção de Candidatos

Para extrair uma lista de candidatos L, antes precisamos de uma forma de selecionar termos dos documentos de uma área.

Em uma modelagem de tópicos cada documento da coleção tem uma probabilidade associada com cada tópico. Os documentos mais relevantes terão uma maior probabilidade associada com o tópico em questão e uma menor com tópicos não relacionados ou pouco presentes.

Deste modo, [Nolasco e Oliveira 2016] utilizam uma seleção baseada em uma amostra dos documentos mais significativos dependendo da probabilidade associada a cada documento em relação ao tópico visando diminuir os ruídos e aumentar o desempenho para coleções grandes. [Mei et al. 2007; Lau et al. 2010] utilizam todo o conteúdo da coleção como base. [Syed et al. 2008; Lau et al. 2011] utilizam também bases externas como Wikipédia e ontologias para auxiliar na seleção.

Com base nos dados utilizados por cada uma das técnicas se selecionam os candidatos. A principal parte do algoritmo é a extração de termos e para isso podemos dividir a extração em duas abordagens: **Textual**, que utiliza o corpo de texto como matéria –prima para extração dos termos; e por **Palavras-chave,** que utiliza classificações, termos do autor como forma de descrever o todo.

A extração textual é a forma mais utilizada para a seleção de candidatos. A mais simples é considerar todos os termos presentes na coleção como candidatos. Também há a possibilidade de se utilizar apenas bigramas ou trigramas [Mei et al. 2007].

Uma abordagem mais sofisticada é a extração de frases nominais por processamento de linguagem natural ou utilizar um extrator de palavras-chave. [Nolasco e Oliveira 2016] apresentam uma extração baseado no algoritmo *fast keyword extraction algorithm* [Berry 2010], que é baseado no fato de que os rótulos frequentemente contêm múltiplas palavras mas raramente contém pontuação ou *stopwords*. A entrada para o

algoritmo é uma lista de *stopwords* e delimitadores de frases (como pontos e vírgulas). Todas as palavras entre os delimitadores e *stopwords* são consideradas um termo ou rótulo inicial para uso na seleção de candidatos. Enquanto [Hammouda et al. 2005] utilizam a própria extração como agrupamento.

Ao contrário da textual, na extração por palavras-chave extraímos palavras-chave descritas pelos autores dos documentos assim como descritores e classificações quando presentes

É importante frisar que nem todas as coleções possuirão rótulos providos pelos autores muito menos descritores. Inicialmente podemos pensar que as palavras-chave são a melhor forma de descrever a área já que são relevantes aos documentos. Infelizmente, como queremos rotular um tema ou área, as palavras-chave muitas vezes são específicas demais para essa tarefa já que se aplicam ao documento e os descritores muito genéricos e amplos, reduzindo sua capacidade de descrever os conceitos da área.

#### 4.3.3.1 Ranqueamento

Após a produção de termos candidatos para uso como rótulos dos tópicos, o passo seguinte é ordená-los de acordo com a relevância de cada um para o grupo. Da mesma maneira que a seleção de candidatos, aqui serão apresentados alguns dos principais métodos de ranqueamento existentes.

As principais técnicas apresentadas aqui são: A **Frequência de termos** (*tf*), a **Relação Grau/Frequência** (*deg*/*tf*) [Berry 2010] e o **Grau modificado de rótulo** [Nolasco e Oliveira 2016].

A **frequência** de termos (normalmente representada pela sigla tf, do inglês *term frequency*) é uma forma tradicional e muito utilizada para atribuir uma pontuação aos termos dependendo de sua relevância em relação ao corpus. Se baseia na suposição de que o peso de um termo que ocorre em um documento é diretamente proporcional à sua frequência.

Assim, para calcular a frequência, basicamente se conta a quantidade de vezes que um termo aparece em um documento. Para o uso na geração de rótulos, toma-se todos os termos selecionados como candidatos e calcula-se a frequência de cada um utilizando a fonte textual da modelagem.

Normalmente a frequência, como pode-se esperar, dá pontuações altas para *stop words* e termos que não são muito descritivos (como verbos comuns). Por isso, costuma-se usar o **inverso da frequência nos documentos** (*idf*) ou a relação ***tf-idf***. O primeiro dá mais peso a termos que ocorrem mais raramente, enquanto o segundo dá valor ao número de ocorrências, no entanto, esse valor é equilibrado pela frequência da palavra no corpus.

Se por um lado a frequência favorece termos mais curtos, que tendem a aparecer mais, a relação **grau/frequência** (*deg*/*tf*) pode favorecer termos mais longos combinando os conceitos de grau (*deg*) e frequência (*tf*).

O **grau** (normalmente representado pela sigla *deg*, do inglês *degree*) de uma palavra é definido como a quantidade de vezes em que ela aparece isolada (neste caso, palavra = termo) somado com a quantidade de vezes em que ela aparece incluída em um termo (para termos com mais de uma palavra). Para termos, o grau é calculado como a soma dos graus de suas palavras [Berry 2010].

Como o grau tende a favorecer as palavras que ocorrem em candidatos mais extensos (já que para termos é a soma dos graus das palavras) e a frequência termos com alta ocorrência, que costumam ser mais curtos, a relação grau/frequência favoreceria termos que ocorrem predominantemente em candidatos mais longos.

Essa métrica é uma forma de beneficiar termos frequentes tanto isoladamente quanto quando aparecem como parte de um termo maior.

Temos que o grau de uma palavra é a sua frequência como termo somada a frequência em que aparece dentro de termos compostas. O grau de um termo é simplesmente a soma dos graus de suas palavras constituintes. Estendendo esses conceitos e adaptando-o para o uso com os rótulos temos o conceito de **grau do rótulo** (*ldeg*) [Nolasco e Oliveira 2016] que basicamente é a frequência com que um rótulo candidato aparece na lista de candidatos somada a frequência com que esse rótulo aparece incluído em outros candidatos. A diferença entre o grau do rótulo e do termo seria que para os rótulos não tratamos das palavras que o constituem. Assim, o grau de um termo seria a soma do grau de suas palavras enquanto o grau do rótulo seria a quantidade de ocorrências de um termo isoladas ou como parte de outro, como se o próprio termo fosse uma palavra.

Essa definição beneficia principalmente unigramas (termos compostos por uma palavra) pois tendem a aparecer mais frequentemente. Por exemplo, um termo como "dados" pode aparecer em "mineração de dados", "visualização de dados", "análise de dados", mesmo que cada um represente um conceito diferente.

Uma solução para esse problema seria balancear as pontuações tanto de unigramas quanto de n-gramas (termos compostos por n palavras), de preferência fazendo um ajuste tal que se diminuísse o peso dos unigramas e se aumentasse o peso dos n-gramas, principalmente dos que aparecem isolados e não como parte de outro termo.

Então, baseado na extensão da definição de graus para rótulos e levando em conta o balanceamento entre termos simples e compostos, [Nolasco e Oliveira 2016] criaram a métrica de **grau modificado do rótulo**. Ela pode ser definida para um rótulo l como:

1. mdeg(l) = ldeg(l) + tf(l), ou
2. mdeg(l) = Número de ocorrências como parte de um termo composto + 2*tf(l)

Ou seja, o grau modificado de um rótulo é a soma do grau do rótulo e de sua frequência. Para cada ocorrência do termo como parte de outro atribui-se um ponto. Se os dois termos comparados são iguais atribui-se dois pontos. Agora, comparando os termos "dados" e "mineração de dados" com um termo "mineração de dados" daria uma pontuação de um ponto para "dados" (*match* parcial) e dois para "mineração de dados" (*match* perfeito).

Finalmente, temos na Tabela 4.3 uma comparação entre as formas de ranqueamento apresentadas junto com as respectivas pontuações em um caso exemplo.

**Tabela 4.3. Comparação das funções de Ranqueamento**

| **Candidatos** | **Pontuação para o candidato: "Redes Sociais"** | | |
|---|---|---|---|
| "redes sociais", "sistemas de redes sociais", "aplicações de redes sociais", "algoritmos de aprendizagem", "social", "classificador" | **tf** | **mdeg** | **deg/tf** |
| | **1** | **4** | **3** |

#### 4.3.3.1 Seleção de Rótulos

Depois de realizar o ranqueamento dos rótulos, o último passo é selecionar para cada área um deles e exibir como representante do assunto. Como a lista de rótulos já está ordenada, para pegar apenas um rótulo para a área basta selecionar o primeiro da lista. O problema na seleção de rótulos surge quando se usa múltiplos rótulos (mais de um rótulo descrevendo a área). Quando utiliza-se vários rótulos, cada um dos rótulos escolhidos necessita ao menos representar uma visão distinta dos conceitos englobados pela área, ao invés de serem sinônimos entre si.

Para solucionar as peculiaridades do uso de múltiplos rótulos, definimos dois tipos de seleção que podem ser utilizados em conjunto ou separadas para este caso: Seleções inter-tópico e intra-tópico [Nolasco e Oliveira 2016]. Essas duas formas de seleção são inspiradas nos conceitos de seleção inter e intra *cluster* [Manning et al. 2008], obviamente adaptando-se ao cenário de modelagem de tópicos que apresenta um outro paradigma de grupos.

A seleção **inter-tópico** é usada quando existem interseções nos rótulos de áreas diferentes, ou seja, o mesmo rótulo aparece em dois tópicos.

Como exemplo, suponha que existam dois tópicos extraídos pela modelagem chamados $\theta_1$ e $\theta_2$ com dois conjuntos de rótulos L1 = $\{l_1, l_2, l_4\}$ e L2 = $\{l_3, l_1, l_5\}$ respectivamente dos quais deseja-se selecionar dois rótulos para cada tópico. Neste caso, o rótulo $l_1$ se encontra na primeira posição do conjunto final de rótulos L1 e também na

segunda posição do conjunto L2 (rótulos de L1 e L2 ranqueados na ordem de leitura). Quando selecionados haveria o mesmo rótulo em ambos os tópicos, porém, devido ao ranqueamento, sabe-se que $l_1$ é mais relevante para $\theta_1$ do que para $\theta_2$(devido a posição na lista ordenada). Para evitar o problema da interseção na representação das áreas, diferenciando-as o máximo possível, a abordagem utilizada neste caso seria a de atribuir o rótulo ao tópico mais relevante correspondente selecionando um outro na ordem que distinga melhor as áreas. No caso de exemplo, atribuir-se-ia $l_1$ à $\theta_1$ devido a sua maior relevância ao tópico, selecionando $l_1$ e $l_2$ como rótulos para $\theta_1$. Para $\theta_2$, seriam selecionados $l_3$ e $l_5$, devido ao fato da relevância de $l_1$ser menor para este tópico e escolhendo então o rótulo seguinte respeitando-se a sequência do ranqueamento ($l_5$, neste caso). Quando há mais de um rótulo idêntico entre áreas, pode-se repetir o processo até que se ache rótulos suficientemente diferentes entre as áreas ou até que se termine a lista.

O caso de um mesmo rótulo aparecer em tópicos diferentes pode ocorrer pois cada documento é modelado como uma mistura de tópicos, então é possível que um rótulo de uma área apareça como sugestão de outra na seleção, principalmente se as áreas não forem suficientemente diferenciadas.

A desvantagem do uso de seleção inter-tópico é a de que a seleção de rótulos para um tópico depende dos rótulos de todos os outros tópicos. Essa característica torna este tipo de seleção mais trabalhoso para coleções onde estão presentes um grande número de áreas ou para quando se usa coleções dinâmicas, onde pode-se adicionar ou remover documentos da coleção após o processamento (neste caso seria necessário realizar todas as operações novamente para a coleção inteira).

Como pode-se imaginar, a seleção **intra-tópico** é realizada para selecionar a melhor sequência de rótulos dentro de um mesmo tópico visando facilitar ao máximo seu entendimento.

Suponha que sejam selecionados dois rótulos para um determinado tópico e por ordem estes sejam “mineração de dados” e “mineração” respectivamente. Claramente, um segundo rótulo “mineração” é redundante para uma área onde o assunto é mineração de dados, já que são quase que sinônimos. Outros rótulos, como “algoritmos” ou “aprendizado” podem oferecer outras perspectivas sobre os assuntos específicos da área. Esse é o objetivo deste tipo de seleção, eliminar sinônimos e termos redundantes favorecendo rótulos adicionais mais esclarecedores do assunto tratado.

Como um exemplo simples de seleção intra-tópico, suponha que temos um conjunto de rótulos L1 = $\{l_1, l_2, l_3\}$ associado a um tópico $\theta$, se $l_1$ contém $l_2$ como parte de si, substitui-se $l_2$ (pois possui menor relevância para $\theta$) por $l_3$ (o próximo rótulo de acordo com a ordem). Então, se aqui houvessem os rótulos “mineração de dados” e “mineração” o segundo seria substituído pelo próximo da lista por ter menos relevância que o primeiro de acordo com a pontuação e por estar contido também no primeiro.

Após feitas as seleções toda a geração de rótulo está finalizada com cada tópico associado a seu(s) rótulo(s).

### 4.3.4. Aplicações

A Tabela 4.4 mostra os resultados de uma rotulagem para tópicos de artigos científicos da área de computação. Comparando os resultados aqui contidos com os da Tabela 4.1 (ambos provêm da mesma base de dados, adaptado de [Nolasco e Oliveira 2016]) pode-se notar a diferença na representação e como a criação dos rótulos ajudou a facilitar compreensão acerca dos temas contidos na coleção.

**Tabela 4.4. Resultados de rotulagem aplicados a coleção de artigos**

| **Método** | **KDD** | |
|---|---|---|
| **Rótulo Manual** | social networks | Clustering |
| **Função tf com seleção textual** | social networks, nodes, connection subgraphs | clusters, clustering, algorithms |
| **Função mdeg com seleção textual** | social network, graph, networks | clustering, clusters, subspace cluster |
| **Função deg/tf com seleção textual** | social network, large network, data structures | real data, categorical objects, subspace clustering |
| **Função tf com seleção por palavras-chave** | social networks | clustering |
| **Função mdeg com seleção por palavras-chave** | user generated content | minimum description length principle |

Já a Tabela 4.5 mostra uma rotulagem em um outro cenário, o de notícias [Mei et al. 2007]. Os tópicos consistem de agrupamentos de reportagens da Associated Press. Aqui foi utilizada uma seleção de candidatos textual usando-se somente bigramas, os quais foram filtrados por frequência e aplicou-se uma função de ranqueamento baseado na divergência KL. Nota-se que apesar de neste caso os rótulos manuais diferirem bem mais dos automáticos, ainda assim a ideia da temática está presente e comparado com a lista de termos do tópico o entendimento do tema ainda parece mais simples com o uso dos bigramas.

Como visto, com as mais distintas abordagens é possível gerar rótulos mais fáceis de entender e que representam bem a temática contida no tópico. Como as técnicas dependem somente de texto, é possível utilizá-las com qualquer tipo de fonte e natureza de texto. Obviamente algumas podem se sair melhor que outras dependendo da natureza do domínio e dos dados contidos na coleção, porém é muito provável que a interpretação do tema melhore com o uso dos rótulos.

**Tabela 4.5. Resultados de rotulagem aplicados a coleção de notícias**

| **Método** | **Associated Press** | | | |
|---|---|---|---|---|
| **Função tf com divergência KL** | air force | court appeals | dollar rates | iran contra |
| **Rótulo Manual** | air plane crash | death sentence | international stock trading | iran contra trial |
| **Tópicos** | plane<br>air<br>flight<br>pilot<br>crew<br>force<br>accident<br>crash | court<br>judge<br>attorney<br>prison<br>his<br>trial<br>case<br>convicted | dollar<br>1<br>yen<br>from<br>late<br>gold<br>down<br>london | north<br>case<br>trial<br>iran<br>documents<br>walsh<br>reagan<br>charges |

### 4.3.4. Desafios e Oportunidades

A geração de rótulos para tópicos é um campo de pesquisa ainda pouco explorado. As vertentes que usam algum tipo de método supervisionado após a modelagem já estão mais estabelecidas, como o aprendizado ativo. O desafio atualmente é gerar rótulos mais próximos aos manuais levando em consideração a semântica dos rótulos.

Como a rotulagem normalmente envolve mais de uma etapa, há muitas possibilidades de avanços em cada uma delas. Desde uma seleção de candidatos mais criteriosa, eliminando os ruídos nos dados, como funções de ranqueamento que eliminem ambiguidades na interpretação do tópico. [Nolasco 2016] desenvolve diversas técnicas automáticas para cada uma das etapas de uma rotulagem e as compara com técnicas tradicionais, com resultados que mostram que a rotulagem não-supervisionada pode alcançar a qualidade manual até mesmo sem utilizar toda a coleção, tornando-a mais eficiente para cenários de *big data*.

Outra possibilidade é considerar os campos de dados semi-estruturados na escolha do rótulo. Um título de um artigo ou uma *hashtag* de um tweet deveriam ter mais peso na seleção de rótulos? É necessário usar toda a coleção para extrair um rótulo ou seria possível até mesmo buscar o rótulo fora dela? Essas são apenas algumas perguntas a serem respondidas. Um grande problema para pôr essas questões a prova é a ausência de um método preciso para se avaliar o quão próximo um rótulo gerado está de um rótulo

manual. Esse também é um problema a ser considerado junto com os desafios do aprimoramento da rotulagem.

## 4.4. Conclusões

O crescimento no armazenamento de documentos e a velocidade com que as bases crescem impôs um limite na capacidade disponível para se organizar e analisar dados manualmente. A partir daí surgem as técnicas de aprendizado visando automatizar estas tarefas para poder realiza-las em maior escala ou mais rapidamente. Neste cenário surgem também as técnicas de aprendizado não-supervisionado com o pressuposto de que as informações que se possa querer já estão contidas nos próprios dados e por isso deve ser possível aprender somente utilizando eles.

A modelagem de tópicos aparece como uma forma de aprender dos próprios documentos o necessário para classificá-los e organizá-los em tópicos. Uma nova área de pesquisa se abre então com vários algoritmos e extensões do conceito probabilístico de modelar um tópico e um documento como uma mistura de tópicos.

Após muitos trabalhos e aplicações da modelagem de tópicos em diversas áreas do conhecimento, uma questão ficou em aberto: Pode-se modelar e agrupar os documentos em tópicos, mas agora com a imensa quantidade de corpus existentes como podemos identificar ou representar os tópicos de maneira que não seja complexa para quem utiliza a ferramenta.

Então a rotulagem de tópicos surge como meio através do qual é possível representar os temas e assuntos extraídos dos dados para analistas, pesquisadores e usuários. De forma a simplificar a organização dos dados, facilitar a exploração temática de coleções e bases desconhecidas para se encontrar o que se deseja ou até mesmo descobrir algo novo e impensável.

Muitos desafios ainda existem na área de modelagem, que é relativamente jovem em comparação com as outras técnicas de agrupamento. Ainda mais na área de rotulagem que possui poucas aplicações práticas e material.

Através deste capítulo, espera-se que a modelagem de tópicos e as técnicas envolvidas tenham sido bem compreendidas e que os desafios e oportunidades sejam amplamente discutidos na comunidade científica contribuindo para a solução dos problemas existentes, bem como a criação de novas vertentes e aplicações inovadoras na área

## Referências

Arabie, Phipps; Hubert, Lawrence J.; De Soete, Geert (Ed.). (1996). Clustering and classification. World Scientific, 1996.

Bart, E., Welling, M., e Perona, P. (2011). Unsupervised organization of image collections: taxonomies and beyond. IEEE transactions on pattern analysis and machine intelligence, 33(11), 2302-2315.

Berry, M. W. J. K. (2010). Text Mining Applications and Theory. West Sussex, UK: John Wiley & Sons. doi:10.1002/9780470689646

Blei, David M. (2012) Probabilistic topic models. Communications of the ACM, v. 55, n. 4, p. 77-84, 2012.

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. IEEE Signal Processing Magazine, 27, p. 55–65. doi:10.1109/MSP.2010.938079

Blei, D. M. e Lafferty, J. (2009). Topic models. Text mining: classification, clustering, and applications, 10:71.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. The Annals of Applied Statistics, pages 17–35.

Blei, D. M. e Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113–120. ACM.

Blei, D. M., Ng, a Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Advances in neural information processing systems, pages 288–296.

De Oliveira, J. Valente et al. (Ed.). (2007). Advances in fuzzy clustering and its applications. New York: Wiley, 2007.

Downey, D., Yang, Y., Pan, S., & Zhang, K. (2014). Active Learning with Constrained Topic Model. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 30–33. Retrieved from http://www.aclweb.org/anthology/W/W14/W14-3104

Gao, Wei; Peng Li; e Kareem Darwish. (2012) "Joint topic modeling for event summarization across news and social media streams." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012. p. 1173-1182

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How Many Topics? Stability Analysis for Topic Models. Machine Learning and Knowledge Discovery in Databases.

Hammouda, K. M., Matute, D. N., & Kamel, M. S. (2005). CorePhrase: Keyphrase extraction for document clustering. Machine Learning and Data Mining in Pattern Recognition Proceedinds, 3587, 265–274. Retrieved from http://www.springerlink.com/index/1a8adr1jmc756ajk.pdf

Hofmann, T. (1999). Probabilistic latent semantic indexing. SIGIR '99: Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval, 50–57. doi:10.1145/312624.312649

Hong, Liangjie; Davison, Brian D. (2010) Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics. ACM, 2010. p. 80-88.

Landauer, T. K.; Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2):211, 1997.

Lau, J. H., Grieser, K., Newman, D., & Baldwin, T. (2011). Automatic labeling of topic models. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 1536–1545.

Manning, Christopher D. et al. (2008) Introduction to information retrieval. Cambridge: Cambridge university press, 2008.

Mei, Q., Liu, C., Su, H., & Zhai, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. Proceedings of the 15th International Conference on World Wide Web - WWW, 2006, p. 533. doi:10.1145/1135777.1135857

Mei, Qiaozhu; Shen, Xuehua; Zhai, ChengXiang. (2007) Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007. p. 490-499.

Nolasco, D., "Identificação automática de áreas de pesquisa em C&T"(a ser defendida em setembro de 2016), Dissertação de Mestrado do Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro, 2016

Nolasco, Diogo; Oliveira, Jonice. (2016) Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data. In: 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016. p. 358-367.

Papadimitriou, Christos H. et al. (1998) Latent semantic indexing: A probabilistic analysis. In: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. ACM, 1998. p. 159-168.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.

Pujara, J., & Skomoroch, P. (2012). Large-Scale Hierarchical Topic Models. NIPS Workshop on Big Learning, 1–8.

Ramage, D., Manning, C. D., & Dumais, S. (2011). Partially labeled topic models for interpretable text mining. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 457–465. doi:10.1145/2020408.2020481

Steyvers, M. e Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7):424–440.

Syed, Z. S., Finin, T., & Joshi, A. (2008). Wikipedia as an Ontology for Describing Documents. ICWSM, (March), 136–144. doi:10.1.1.133.2711

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. Journal of the american statistical association, 101(476).

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning, pages 977–984. ACM.

Wang, X., McCallum, A., & Wei, X. (2007). Topical N-grams: Phrase and topic discovery, with an application to information retrieval. Proceedings - IEEE International Conference on Data Mining, ICDM, 697–702. doi:10.1109/ICDM.2007.86

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In European Conference on Information Retrieval (pp. 338-349). Springer Berlin Heidelberg.

## Sobre os Autores

**Diogo Nolasco** é estudante do Programa de Pós Graduação em Informática (PPGI) da Universidade Federal do Rio de Janeiro (UFRJ), sob orientação da Profª Jonice Oliveira. Atua na área de Big Scholar Data, trabalhando na identificação temporal, representação e correlação de áreas científicas com foco em inovação tecnológica. Sua pesquisa tem sido aplicada na área da Saúde e Cidades Inteligentes. Seus interesses de pesquisa incluem bancos de dados, métodos de aprendizado não-supervisionado, big data e mineração de dados.

**Jonice Oliveira** é professora obteve o seu doutorado em 2007 na área de Engenharia de Sistemas e Computação, ênfase em Banco de Dados, pela COPPE/UFRJ. Durante o seu doutorado recebeu o prêmio IBM Ph.D. Fellowship Award. Na mesma instituição realizou o seu Pós-Doutorado, concluindo-o em 2008. Desde 2009 é professora do Departamento de Ciência da Computação da UFRJ e atua no Programa de Pós-Graduação em Informática (PPGI-UFRJ). Tornou-se Jovem Cientista do Nosso Estado pela FAPERJ (desde 2013) e atuou como professora visitante no Insight Centre for Data Analytics (Irlanda) durante 3 meses (2015), do qual permanece como colaboradora. Coordena o Laboratório CORES (Laboratório de Computação Social e Análise de Redes Sociais), que conduz pesquisas multidisciplinares para o entendimento, simulação e fomento às interações sociais. Suas principais áreas de pesquisa são Gestão do Conhecimento, Análise de Redes Sociais, Big Data, Suporte à Decisão, Colaboração e Recomendação.