Doi:10.1145/2133806.2133826

Levantamento de um conjunto de algoritmos que oferecem uma solução para a gestão de grandes arquivos de documentos.

Por DaviD m. Blei

Modelos temáticos probabilísticos

À medida que o conhecimento OUr COLLeCTive continua a ser digitalizado e armazenado - na forma de notícias, blogs, páginas Web, artigos científicos, livros, imagens, som, vídeo e redes sociais - torna-se mais difícil encontrar e descobrir o que procuramos. Precisamos de novas ferramentas computacionais para ajudar a organizar, pesquisar e compreender estas vastas quantidades de informação.

Neste momento, trabalhamos com informações online usando duas ferramentas principais - pesquisa e links. Digitamos palavras-chave em um motor de busca e encontramos um conjunto de documentos relacionados a elas. Nós olhamos para os documentos desse conjunto, possivelmente navegando para outros documentos vinculados. Esta é uma forma poderosa de interagir com o nosso arquivo online, mas falta algo.

Imagine pesquisar e explorar documentos com base nos temas que os atravessam. Podemos "ampliar" e "diminuir" para encontrar temas específicos ou mais amplos; podemos ver como esses temas mudaram com o tempo ou como estão ligados uns aos outros. Em vez de encontrar documentos apenas através da pesquisa por palavras-chave, podemos primeiro encontrar o tema que nós estão interessados, e depois examinam os documentos

relacionados com esse tema.

Por exemplo, considere o uso de temas para explorar a história completa do New York Times. Em um nível amplo, alguns dos temas podem corresponder às seções do jornal-por-em-ign policy, assuntos nacionais, esportes. Poderíamos ampliar um tema de in- terest, como política externa, para revelar vários aspectos da mesma - a política externa chinesa, o conflito no Oriente Médio, o relacionamento dos EUA com a Rússia. Poderíamos então navegar pelo tempo para revelar como esses temas específicos mudaram, acompanhando, por exemplo, as mudanças no conflito no Oriente Médio nos últimos 50 anos. E, em toda essa exploração, seríamos apontados para os artigos originais relevantes para os temas. A estrutura temática seria um novo tipo de janela através da qual se poderia explorar e digerir a coleção.

Mas não interagimos com os arquivos elec- tronic desta forma. Enquanto mais e mais textos estão disponíveis online, simplesmente não temos o poder humano de lê-los e estudá-los para proporcionar o tipo de experiência de navegação descrita acima. Para este fim, pesquisadores de aprendizado de máquina desenvolveram *a modelagem probabilis- tic*, um conjunto de algoritmos que visam descobrir e anotar grandes arquivos de documentos com informações temáticas. A modelagem tópica de algo- rithms são métodos estatísticos que ana- lyze as palavras dos textos originais para descobrir os temas que os atravessam, como esses temas estão ligados entre si, e como eles mudam

» principais contributos

- **Os modelos temáticos são algoritmos para descobrir os principais temas que permeiam uma grande coleção de documentos e, de outra forma, não estruturada. Os modelos temáticos podem organizar a coleção de acordo com os temas descobertos.**
- **algoritmos de modelagem temática podem ser aplicados a colecções massivas de documentos.**
  Os recentes avanços neste campo permitem-nos analisar colecções de streaming, como pode encontrar numa Web aPi.
- **Entre outras aplicações, eles têm sido usados para encontrar padrões em dados genéticos, imagens e redes sociais.**

tempo. (Veja, por exemplo, a Figura 3 para tópicos encontrados através da análise do *Yale Law Journal*). Algoritmos de modelagem de tópicos não requerem nenhuma anotação prévia ou etiquetagem dos documentos - os tópicos emergem da análise dos textos origi- nal. A modelagem de tópicos nos permite organizar e resumir arquivos eletrônicos em uma escala que seria impossivel - através de anotações humanas.

alocação de Dirichlet latente

Primeiro descrevemos as idéias básicas por trás da *alocação de Dirichlet latente* (LDA), que é o modelo temático mais simples. [8] A intuição por detrás da LDA é que os documentos exibem múltiplos tópicos. Por exemplo, considere o artigo da Figura 1. Este artigo, intitulado "Buscando as Necessidades (Genéticas) da Vida", trata do uso da análise de dados para determinar o número de genes que um organismo precisa para sobreviver (em um sentido evolucionário).

À mão, destacamos palavras diferentes que são usadas no artigo. Palavras sobre *análise de dados*, como "computador" e "predição", são destacadas em azul; palavras sobre *biologia evolutiva*, como "vida" e "organismo", são destacadas em rosa; palavras sobre *genética,* como "sequenciada" e "genes", estão destacados em amarelo. Se demorássemos algum tempo para destacar cada palavra do artigo, você veria que este arti- clemente mistura genética, análise de dados e biologia evolutiva em diferentes porções pró-. (Excluímos palavras, como "e" "mas" ou "se", que contêm pouco conteúdo tópico). Além disso, saber que este artigo mistura esses tópicos o ajudaria a situá-lo em uma coleção de artigos científicos.

LDA é um modelo estatístico de colecções de docu- ment que tenta captar esta intuição. É mais facilmente descrito pelo seu processo generativo, o processo imaginário aleatório pelo qual o modelo assume que os documentos surgiram. (A interpretação do LDA como um modelo probabilístico é explicada mais tarde).

Nós definimos formalmente um *tópico* para ser uma distribuição por um vocabulário fixo. Por exemplo, o tópico de *genética* tem palavras sobre genética com alta probabilidade e o tópico de *biologia evolutiva* tem palavras sobre biologia evolutiva com alta probabilidade. Assumimos que estes tópicos são especificados antes de qualquer dado ter sido gerado.[a] Agora, para cada documento do acervo, nós geramos - comemos as palavras num processo de duas etapas.

. Escolha aleatoriamente uma distribuição por tópicos.

. Para cada palavra do documento

a. Escolha aleatoriamente um tópico da distribuição em vez dos tópicos da etapa #1.

b. Escolha aleatoriamente uma palavra da distribuição correspondente sobre o vocabulário.

Este modelo estatístico reflete a intuição de que os documentos exibem tópicos com várias pontas. Cada documento exibe os tópicos em diferentes proporções (passo #1); cada palavra em cada document é retirada de um dos tópicos (passo #2b), onde o tópico selecionado é escolhido a partir da distri- buição por-documento sobre os tópicos (passo #2a). [b]

No artigo de exemplo, a distribuição sobre tópicos colocaria a capacidade de sondagem em *genética*, *análise de dados*, e

a Tecnicamente, o modelo pressupõe que osics de topo são gerados primeiro, antes dos documentos.

b Devíamos explicar o nome misterioso, "alocação Dirichlet latente". A distribuição que é usada para desenhar a distribuição temática por-documento no passo #1 (o histograma dos desenhos animados na Figura 1) é chamada de *distribuição Dirichlet*. No processo genera- tive para LDA, o resultado do Dirichlet é usado para *alocar* as palavras do documento para diferentes tópicos. Por que *latente*? Continue lendo.

dad 0.02
os
número 0.02
computa 0.01
dor

figura 1. as intuições por detrás da alocação de Dirichlet latente. Assumimos que existe algum número de "tópicos", que são distribuições por palavras, para toda a coleção (extrema esquerda). cada documento é assumido como segue. primeiro escolha uma distribuição por tópicos (o histograma à direita); depois, para cada palavra, escolha uma atribuição de tópico (as moedas coloridas) e escolha a palavra do tópico correspondente. os tópicos e atribuições de tópico nesta figura são ilustrativos - eles não se encaixam em dados reais. Veja a figura 2 para os tópicos que cabem a partir dos dados.

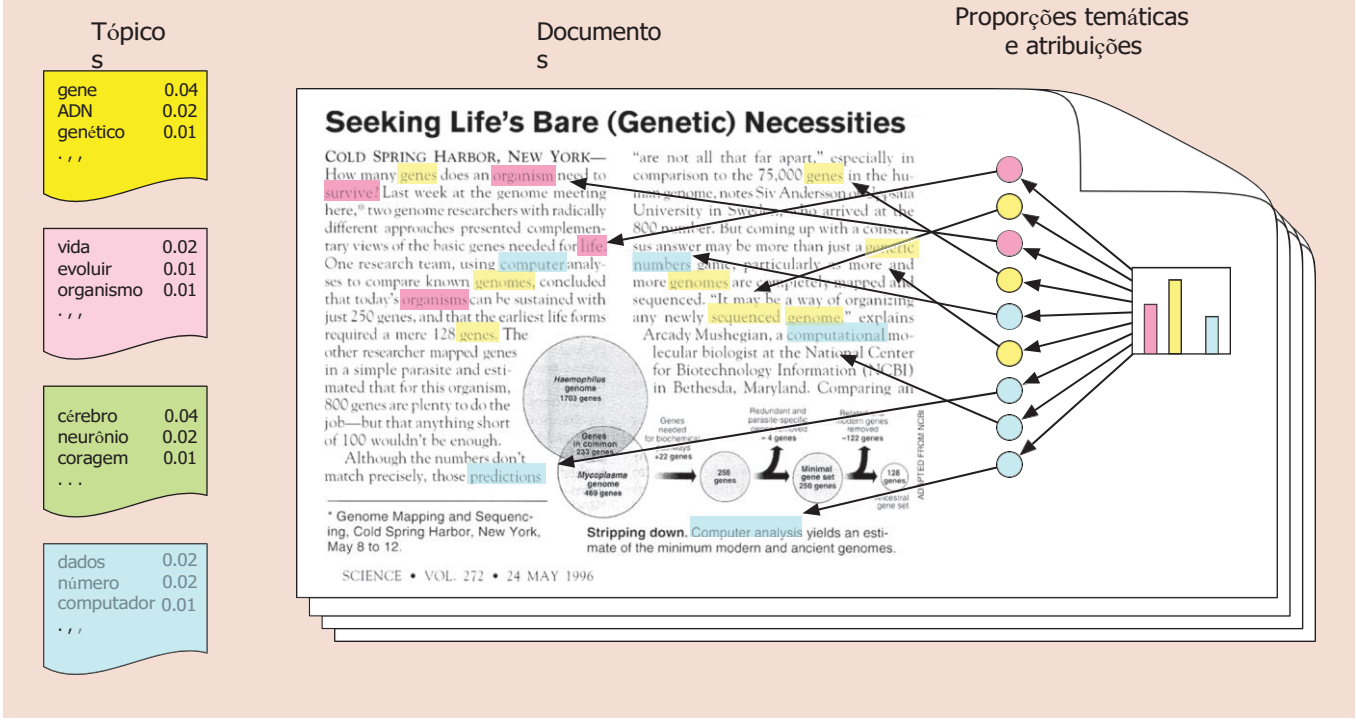
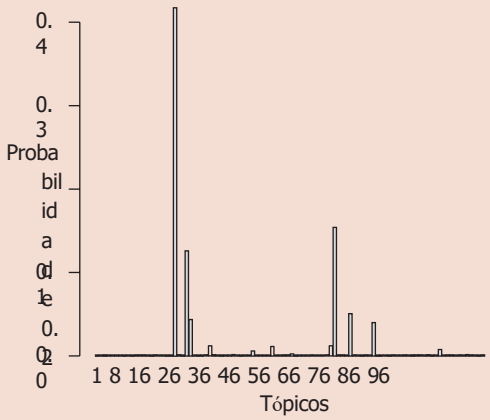


figura 2. Dedução real com LDA. Nós ajustamos um modelo 100-topic LDA a 17.000 artigos da revista *Science*. à esquerda estão as proporções dos tópicos inferidos para o artigo de exemplo na figura 1. à direita estão as 15 palavras mais frequentes dos tópicos mais frequentes encontrados neste artigo.



"Genética"	"Evolução"	"Doença"	"Computadores"
DNA do genoma humano Sequência de genes genéticos Sequência de genes genéticos Sequência molecular do mapa de informação genética Sequências de projectos de mapeamento genético	vida evolutiva das espécies de organismos evolutivos grupos de biologia de origem grupo de diversidade filogenética viva dois novos comuns	doença hospedeira de bactérias resistentes a doenças bacterianas novas estirpes controlam parasitas infecciosos da malária unidos tuberculose	modelos de computadores modelos de sistemas de rede de computadores de dados de informação métodos paralelos redes software novas simulações

biologia evolutiva, e cada palavra é extraída de um desses três top- ics. Observe que o próximo artigo da coleção pode ser sobre *análise de dados* e *neurociência*; sua distribuição sobre tópicos colocaria a capacidade de sondagem sobre esses dois tópicos. Esta é a característica distintiva da alocação Dirichlet latente - todos os documentos da coleção compartilham o mesmo conjunto de tópicos, mas cada docu- mento exibe esses tópicos em proporção diferente.

Como descrevemos na introdução, o objetivo da modelagem de tópicos é descobrir automaticamente os tópicos a partir de uma coleção de documentos. Os documentos em si são observados, enquanto a estrutura de tópicos - os tópicos, a distribuição de tópicos por-documento e a atribuição de tópicos por palavra - é uma *estrutura oculta*. O problema computacional central para a modelagem de tópicos é usar os documentos observados para inferir a estrutura de tópicos ocultos. Isto pode ser pensado como "revertendo" o processo generativo - qual é a estrutura oculta que provavelmente gerou a coleção observada?

A Figura 2 ilustra o exemplo inference usando o mesmo exemplo document da Figura 1. Aqui, pegamos 17.000 artigos da revista *Science* e usamos um algoritmo de modelagem de tópicos para inferir a estrutura de tópicos ocultos. (O algoritmo assumiu que existiam 100 tópicos). Calculamos então a distribuição de tópicos inferidos para o artigo de exemplo (Figura 2, à esquerda), a distribuição por tópicos que melhor descreve sua coleção par- ticular de palavras. Note que esta distribuição por tópicos, embora possa usar qualquer um dos tópicos, tem apenas "activated" um punhado deles. Além disso, podemos examinar os termos mais prováveis de cada um dos tópicos mais prováveis (Figura 2, à direita). No exame, vemos que esses termos são reconhecíveis como termos sobre genética, sobrevivência e análise de dados, os tópicos que são com- citados no artigo de exemplo.

Ressaltamos que os algoritmos não têm informações sobre esses subtemas e os artigos não são etiquetados com tópicos ou palavras-chave. As distribuições de tópicos entre tabelas surgem através do cálculo da estrutura oculta que provavelmente gerou a leitura col- letiva observada dos documentos.[c] Por exemplo, a Figura 3 ilustra tópicos descobertos a partir do *Yale Law Journal*. (Aqui o num- ber de tópicos foi definido para ser 20.) Tópicos

c De fato, chamar esses modelos de "modelos tópicos" é retrospectivo - os tópicos que emergem do

algoritmo de inferência são interpretáveis para quase todas as coleções que são analisadas. O fato de estes se parecerem com tópicos tem a ver com a estrutura estatística da linguagem observada e como esta interage com os pressupostos probabilísticos específicos da LDA.

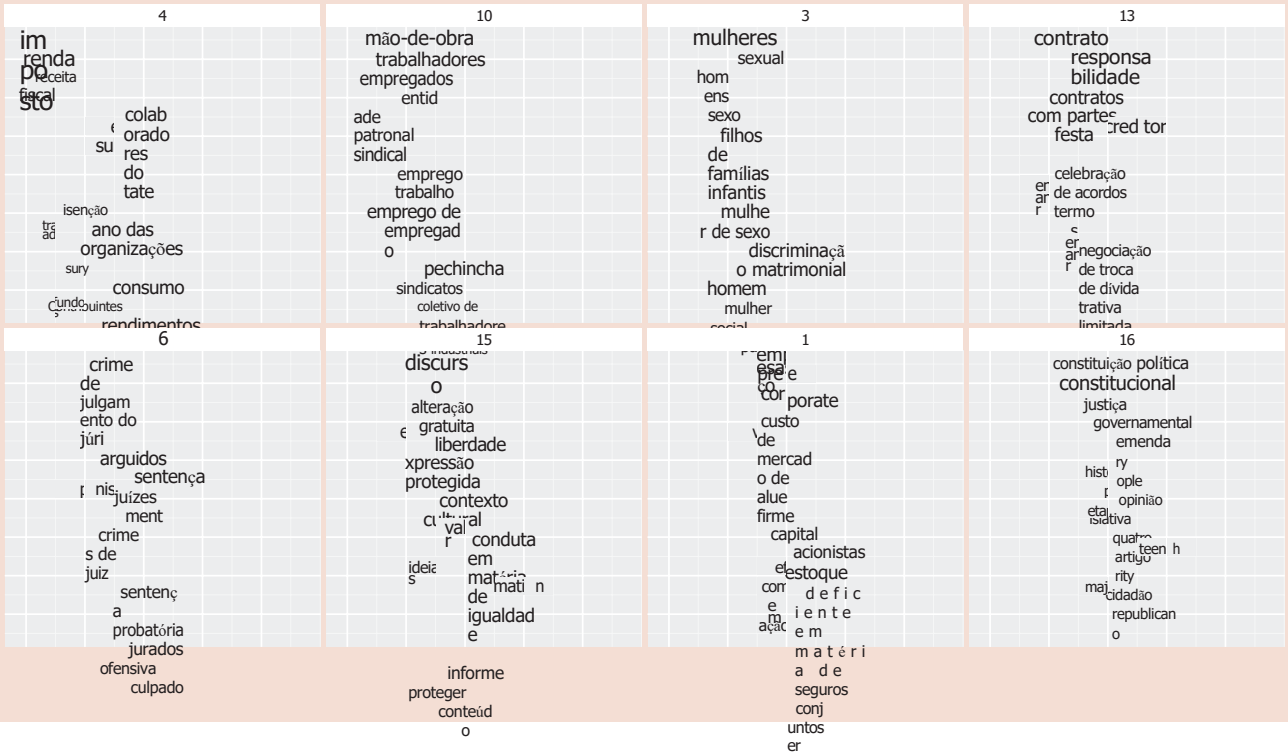
sobre temas como genética e análise de dados são substituídos por temas sobre discriminação e direito contratual.

A utilidade dos modelos temáticos deriva da propriedade de que a estrutura oculta inferida se assemelha à estrutura temática da coleção. Esta estrutura oculta inter- tabela anotará cada documento da coleção - uma tarefa que é cuidadosa de executar à mão - e estas anotações podem ser usadas para auxiliar tarefas como recuperação de informações, classificação e exploração de corpus. [d] Desta forma, o modelo de tópicos fornece uma solução algorítmica para gerenciar, organizar e anotar grandes arquivos de textos.

Lda e modelos probabilísticos. LDA e outros modelos tópicos fazem parte do campo mais vasto da *modelação probabilística*. Na modelagem probabilística generativa, tratamos os nossos dados como sendo provenientes de um processo generativo que inclui *variáveis escondidas*. Este processo generativo define uma *distribuição conjunta de probabilidade* sobre as variáveis aleatórias observadas e ocultas. Nós realizamos a análise dos dados usando essa distribuição conjunta para calcular a *distri- buição condicional* das variáveis ocultas, dada a

d Veja, por exemplo, o navegador da *Wikipedia* construído com um modelo de tópicos em http://www.sccs.swarthmore.edu/users/08/ajb/tmve/wiki100k/ browse/topic-list.html.

figura 3. um modelo de tópicos adequado ao *Yale Law Journal*. aqui, há 20 tópicos (os oito primeiros são traçados). cada tópico é ilustrado com suas palavras mais freqüentes. a posição de cada palavra ao longo do *eixo x* denota sua especificidade para os documentos. por exemplo, "patrimônio" no primeiro tópico é mais específico do que "imposto".



variáveis observadas. Esta distribuição condicional também é chamada de *distribuição posterior*.

A LDA cai precisamente neste quadro - trabalho. As variáveis observadas são as palavras dos documentos; as variáveis ocultas são a estrutura temática; e o processo gerativo é como descrito aqui. O problema computacional de inferir a estrutura temática oculta dos documentos é o problema de calcular a distribuição posterior, a distribuição condicional das variáveis escondidas dos documentos.

Podemos descrever a LDA de forma mais formal com a seguinte notação. Os tópicos são $b_{1:K}$, onde cada b_k é uma distribuição sobre o vocabulário (as distribuições sobre as palavras à esquerda na Figura 1). As proporções de tópicos para o documento *dth* são q_d , onde $q_{d,k}$ é a proporção de tópicos para o tópico *k* no documento *d* (o histograma dos desenhos do carro na Figura 1). As proporções de tópicos para o documento *dth* são z_d , onde $z_{d,n}$ é a proporção de tópicos para a *n-ésima* palavra no documento *d* (a moeda colorida na Figura 1). Finalmente, as palavras observadas para o documento *d* são w_d ,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right). \quad (1)$$

Com esta notação, o processo generativo para LDA corresponde à distribuição das variáveis ocultas e observadas na articulação dobrada,

Note que esta distribuição especifica uma série de dependências. Por exemplo, a atribuição de tópicos $z_{d,n}$ depende das proporções por-documento de tópicos q_d . Como outro exemplo, a palavra observada $wd_{,n}$ depende do tópico assign- ment $zd_{,n}$ e de *todos* os tópicos $b_{1:K}$. (Operacionalmente, esse termo é definido pela pesquisa de qual tópico $zd_{,n}$ se refere e pela pesquisa da probabilidade da palavra $wd_{,n}$ dentro desse tópico).

Estas dependências definem a LDA. Elas são codificadas nas suposições estatísticas por trás do processo generativo, na forma matemática particular da distribuição conjunta, e...

A linguagem para descrever famílias de distribuições de probabilidade. [e] O modelo gráfico cal para LDA está na Figura 4. Estas três representações são formas equivalentes de descrever as suposições probabilísticas por trás da LDA.

Na próxima seção, descrevemos os algoritmos de inferência para LDA. No entanto, primeiro fazemos uma pausa para descrever a curta onde $w_{d,n}$ é a *enésima* palavra no documento *d*, que é um elemento do vocabulário fixo.

história destas ideias. O LDA foi desenvolvido para corrigir um problema com um modelo *probabilístico* de *análise semântica latente* (pLSI) desenvolvido previamente. [21] Esse modelo foi em si uma versão probabilística do trabalho seminal de *análise semântica latente*,[14] que revelou a utilidade da decomposição do valor singular da matriz documentária-terminal. A

de uma terceira maneira - na *probabilística*
modelo gráfico para LDA. Os modelos gráficos probabilísticos fornecem um modelo gráfico

partir dessa perspectiva de factorização da matriz, a LDA também pode ser vista como um tipo de análise de componentes principais para dados discretos. [11, 12]

Computação posterior para a Lda. Passamos agora ao cálculo computacional

e O campo dos modelos gráficos é na verdade mais do que uma linguagem para descrever famílias de distribuições. É um campo que ilumina o

ligações matemáticas profundas entre probabi independência lística, teoria gráfica, e algo...

rithms para computação com distri- buições de probabilidade. [35]

problema, calculando a distribuição condicional da estrutura temática, tendo em conta os documentos observados. (Como mencionamos, isto é chamado de *posterior*.) Usando a nossa notação, o posterior é

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \quad (2)$$

O numerador é a distribuição conjunta de todas as variáveis aleatórias, que pode ser facilmente computada para qualquer configuração das variáveis ocultas. O denomina- tor é a *probabilidade marginal* das observações, que é a probabilidade de ver o corpus observado sob qualquer modelo tópico. Em teoria, ele pode ser computado pela soma da distri- buição conjunta sobre cada instante possível da estrutura do tópico oculto.

Esse número de estruturas temáticas possíveis, no entanto, é exponencialmente grande; essa soma é intratável para com- pute. [f] Como para muitos probabilis modernos - modelos de interesse de tic - e para grande parte das estatísticas Bayesianas modernas - não podemos calcular o posterior por causa do denominador, que é conhecido como a *evidência.* Um objetivo central da pesquisa do modelo probabilístico moderno é desenvolver métodos eficientes para aproximá-lo. Algoritmos de modelagem tópico - como os algoritmos usados para criar as Figuras 1 e 3 - são frequentemente adaptações de metáforas de propósito geral - ods para aproximação da distribuição posterior.

Algoritmos de modelagem tópicos formam uma aproximação da Equação 2, adaptando uma distribuição alternativa sobre a estrutura do tópico latente para estar próxima do verdadeiro posterior. Algoritmos de modelagem de tópicos geralmente se enquadram em duas categorias - algo- rithms baseados em amostragem e algoritmos variacionais.

Amostragem baseada em Os algoritmos tentam recolher amostras do posterior para o aproximar com uma distribuição empírica. O algoritmo de amostragem mais utilizado para modelagem de tópicos é A distribuição limitadora é a posterior. A cadeia de Markov é definida nas variáveis temáticas ocultas para um determinado corpus, e o algoritmo é executar a cadeia por um longo período de tempo, recolher amostras da distribuição limite, e depois aproximar a distribuição com as amostras coletadas. (Muitas vezes, apenas uma amostra é coletada como uma aproxi- mação da estrutura do tópico com

cada nó é uma variável aleatória e é rotulada de acordo com seu papel no processo generativo (ver figura 1). os nós ocultos - as proporções dos tópicos, atribuições e tópicos - não estão sombreados. os nós observados - as palavras dos documentos - estão sombreadas. os retângulos são notações de "placa", o que denota replicação. a placa denota as palavras da coleção dentro dos documentos; a placa *D* denota a coleção de documentos dentro da coleção.

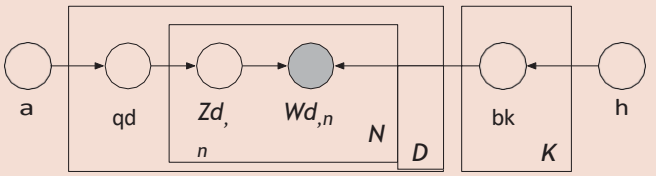
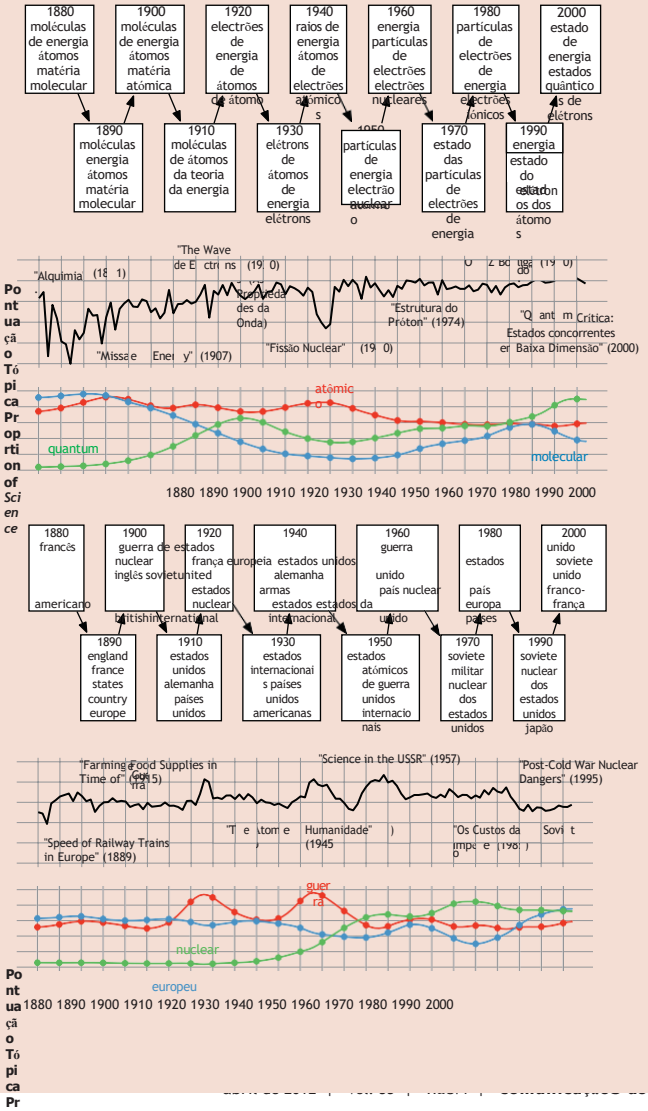


figura 5. dois tópicos de um modelo temático dinâmico. este modelo foi adequado à *Ciência* de 1880 a 2002. Temos ilustrado as palavras de topo em cada década.



o *Gibbs sampling*, onde construímos uma *cadeia de Markov* - uma sequência de variáveis aleatórias, cada uma dependente da anterior - que

f Mais tecnicamente, a soma está sobre todas as formas possíveis de atribuir cada palavra observada da coleção a um dos tópicos. As legendas dos documentos col- leções geralmente contêm palavras observadas pelo menos na ordem de milhões.

probabilidade máxima). Veja Steyvers e Griffiths33 para uma boa descrição da amostragem Gibbs para LDA, e veja http:// CRAN.R-project.org/package=lda para uma rápida implementação open-source.

Os métodos variáveis são uma alternativa minística e dissuasiva aos algoritmos baseados em amostragem. [22,35] Ao invés de aproximadamente acasalar o posterior com as amostras, os métodos variacionais postulam uma família parametrizada de distribuições sobre a estrutura oculta e depois encontram o membro dessa família que está mais próximo do posterior. [g] Assim, o problema de inferência é transformado em um problema de opti- mização. A medição variável - ods abre a porta para inovações em otimização para ter impacto prático na modelagem probabilística. Veja Blei et al. [8] para um algoritmo de inferência de variatação coordenada para LDA; veja Hoffman et al. [20] para um algoritmo online muito mais rápido (e software de código aberto) que lida facilmente com milhões de documentos e pode acomodar coleções de texto em streaming.

Falando vagamente, ambos os tipos de algoritmos realizam uma pesquisa sobre a estrutura do tópico. Uma coleção de docu- mentos (as variáveis aleatórias observadas no modelo) são mantidas fixas e servem como um guia para onde pesquisar. Qual a melhor abordagem depende do modelo tópico específico que está sendo usado - até agora temos focado na LDA, mas veja abaixo para outros modelos tópicos - e é uma fonte de debate acadêmico. Para uma boa discussão sobre os méritos e desvantagens de ambos, ver Asuncion et al. [1]

Pesquisa em modelagem de tópicos

O modelo simples de LDA fornece uma ferramenta errática para descobrir e explorar a estrutura temática oculta em grandes arquivos de texto. No entanto, uma das principais vantagens de formular LDA como um modelo probabilístico é que ele pode ser facilmente usado como um módulo em modelos mais complicados para objetivos mais com- plicados. Desde a sua introdução, o LDA foi ampliado e adaptado de muitas maneiras.

relaxando os pressupostos da Lda. LDA é definida pelas suposições estatísticas que faz sobre os

uma direção para a modelagem de tópicos é desenvolver métodos de avaliação que combinam como os algoritmos são usados. como podemos comparar modelos temáticos com base na sua capacidade de interpretação?

corpo. Uma área ativa de pesquisa de modelos temáticos é como relaxar e ampliar esses pressupostos para descobrir uma estrutura mais sofisticada nos textos.

Uma suposição que a LDA faz é a suposição do "saco de palavras", que a ordem das palavras no documento não importa. (Para ver isto, note que a distribuição conjunta da Equação 1 permanece invariável à permutação das palavras dos documentos). Embora essa suposição seja irrealista, ela é re - sonível se nosso único objetivo for descobrir a estrutura semântica do curso dos textos. [h] Para objetivos mais sofisticados - como a geração de linguagem - ela não é evidentemente apropriada. Tem havido uma série de extensões à LDA que modelam palavras sem troca. Por exemplo, Wallach36 desenvolveu um modelo de tópicos que relaxa a suposição de palavras assumindo que os tópicos geram palavras condicionais à palavra anterior; Griffiths et al. [18] desenvolveram um modelo de tópicos que alterna entre LDA e um HMM padrão. Esses modelos ampliam significativamente o espaço de parâmetros, mas mostram um melhor desempenho na modelagem da linguagem.

Outra suposição é que a ordem dos documentos não importa. Mais uma vez, isto pode ser visto notando-se que a Equação 1 permanece invariável às permutações da ordenação de documentos na coleção. Esta suposição pode ser irrealista ao analisar coleções de longa duração que se estendem por anos ou séculos. Em tais coleções, podemos querer assumir que os *tópicos* mudam ao longo do tempo. Uma abordagem a este problema é o modelo temático dinâmico5 - um modelo que respeita a ordenação dos documentos - e dá uma estrutura tópica posterior mais rica do que a LDA. A Figura 5 mostra um tópico que resulta da análise de toda a revista *Science* sob o modelo de tópicos dinâmicos. Em vez de uma única distribuição por palavras, um tópico é agora uma sequência de distribuições por palavras. Podemos encontrar um tema subjacente da coleção e acompanhar como ele mudou ao longo do tempo.

Uma terceira suposição sobre a LDA é que o número de tópicos é assumido

g A proximidade é medida com a *divergência Kullback-Leibler*, uma medida teórica da informação - a distância entre duas distribuições de probabilidade.

h Como uma experiência de pensamento, imagine embaralhar as palavras do artigo da Figura 1. Mesmo quando baralhado, você seria capaz de perceber que o artigo tem algo a ver com genética.

conhecido e fixo. O modelo de tópicos não paramétricos Bayesianos34 oferece uma solução elegante: o número de tópicos é determinado pela coleção durante a inferência posterior e, além disso, novos documentos podem exibir tópicos nunca antes vistos. Os modelos Bayesianos de tópicos não paramétricos foram estendidos para hierarquias de tópicos, que encontram uma árvore de tópicos, passando de mais gerais para mais concretos, cuja estrutura particular é inferida a partir dos dados. [3]

Existem ainda outras extensões da LDA que relaxam várias suposições feitas pelo modelo. O modelo de tópicos correlatos6 e a máquina de alocação pachinko24 permitem a ocorrência de tópicos para exibir correlação (por exemplo, um documento sobre *geologia* é mais provável que seja também sobre *química* do que sobre *esportes*); o modelo de tópicos esféricos28 permite que palavras sejam *improváveis* em um tópico (por exemplo, "chave inglesa" será particularmente improvável em um tópico sobre *gatos*); modelos de tópicos esparsos reforçam a estrutura nas distribuições de tópicos; [37] e modelos de tópicos "estourados" fornecem um modelo mais realista de contagem de palavras. [15]

incorporando metadados. Em muitas configurações de análise de texto, os documentos contêm informações adicionais - tais como autor, título, localização geográfica, links e outras - que podemos querer ter em conta quando nos adequamos a um modelo tópico. Tem havido uma enxurrada de pesquisas sobre a adaptação de modelos de tópicos para incluir metadados.

O modelo autor-tópico29 é uma história de sucesso inicial para este tipo de pesquisa. As proporções do tópico são anexadas aos autores; trabalhos com múltiplos autores são supostos anexar cada palavra a um autor, extraída de um tópico extraído de suas proporções do tópico. O modelo autor-tópico permite inferências sobre autores, assim como docu- mentos. Rosen-Zvi et al. mostram exemplos de similaridade de autores com base em suas proporções do tópico - tais cálculos não são possíveis com LDA.

Muitas coleções de documentos estão vinculadas - por exemplo, artigos científicos estão vinculados por citação ou páginas da Web estão vinculadas por hiperlinks - e vários modelos de tópicos foram desenvolvidos para dar conta desses links ao estimar o top- ics. O *modelo temático relacional* de Chang e Blei[13] assume que cada documento é modelado como em LDA e que os links entre documentos depende da des- tância entre as suas proporções temáticas. Este é tanto um novo modelo de tópico como um novo modelo de rede. Ao contrário dos modelos estatísticos tradicionais de redes, o modelo temático rela- tional leva em conta os atributos dos nós (aqui, as palavras dos documentos) na modelagem dos links.

Outros trabalhos que incorporam meta- dados em modelos temáticos incluem mod- els de estrutura linguística,[10] modelos que contabilizam as distâncias entre corpora,[38] e modelos de entidades nomeadas. [26] Geral
-Os métodos para incorporar metadados em modelos tópicos incluem Dirichlet-multinomial regression mod- els25 e modelos tópicos supervisionados. [7]

Outros tipos de dados. Na LDA, os top- ics são distribuições sobre palavras e esta distribuição discreta gera observações (palavras em documentos). Um advan- tage do LDA é que estas escolhas para o parâmetro tópico e distribuição geradora de dados podem ser adaptadas a outros tipos de observações com apenas pequenas alterações nos algoritmos de inferência correspondentes. Como uma classe de modelos, o LDA pode ser pensado como um *modelo* de *associação mista* de dados agrupados, em vez de associar cada grupo de observações (documento) a um componente (tópico), cada grupo exibe múltiplos componentes em diferentes proporções. Modelos semelhantes ao LDA foram adaptados a muitos tipos de dados, incluindo dados sur- vey, preferências de usuário, áudio e música, código de computador, registros de rede e redes sociais. Descrevemos duas áreas onde o mod- els de membros mistos tem sido particularmente bem sucedido.

Na genética populacional, o mesmo modelo probabilístico foi inventado independentemente para encontrar populações ancestrais (por exemplo, originárias da África, Europa, Médio Oriente, entre outras) na ancestralidade genética de uma amostra de indivíduos.[27] A idéia é que o genótipo de cada índio-vídeo descende de uma ou mais populações ancestrais. Usando um modelo muito parecido com o LDA, os biólogos podem tanto caracterizar os padrões genéticos dessas populações (os "tópicos") como identificar como cada indivíduo os expressa (as "proporções dos tópicos"). Este modelo é poderoso porque os padrões genéticos nas populações ancestrais podem ser hipotéticos, mesmo quando os "puros" ples sam- ples deles não estão disponíveis.

LDA tem sido amplamente utilizada e adaptada na visão por computador, onde o

Os algoritmos de inferência são aplicados a imagens naturais ao serviço da recuperação, classificação e organização de imagens. Os pesquisadores de visão por computador têm feito uma analogia direta das imagens aos documentos. Na análise de documentos, assumimos que os documentos exibem tópicos de várias pontas e a coleção de documentos exibe o mesmo conjunto de tópicos. Na análise de imagens, assumimos que cada imagem exibe uma combinação de padrões visuais e que os mesmos padrões visuais se repetem ao longo de uma coleção de imagens. (Em uma etapa de pré-processamento, as imagens são analisadas para formar coleções de "palavras visuais"). O modelo- ing tópico para visão computadorizada tem sido usado para classificar imagens,[16] conectar imagens e legendas,[4] construir hierar- chies de imagem,[2,23,31] e outras aplicações.

Direcções futuras

A modelagem tópica é um campo emergente na aprendizagem de máquinas, e há muitas novas e excitantes direções para a pesquisa.

avaliação e verificação de modelos. Há uma desconexão entre a forma como os modelos temáticos são avaliados e porque esperamos que os modelos temáticos sejam úteis. Tipicamente, os modelos temáticos são avaliados da seguinte forma. Primeiramente, segure um subconjunto do seu corpus como o conjunto de teste. Em seguida, ajuste uma variedade de modelos de tópicos ao resto do corpus e aproxime uma medida de ajuste do modelo (por exemplo, probabilidade) para cada modelo treinado no conjunto de teste. Por fim, escolha o modelo que melhor se adapte ao seu desempenho.

Mas modelos temáticos são freqüentemente usados para organizar, resumir e ajudar os usuários a explorar grandes corpora, e não há razão técnica para supor que a precisão corresponde a uma melhor organização ou interpretação mais fácil. Uma direção aberta para a modelagem de tópicos é desenvolver métodos de avaliação que combinem com a forma como os algoritmos são usados. Como podemos comparar modelos tópicos com base na sua capacidade de interpretação?

Este é o problema *de verificação do modelo*. Quando confrontado com um novo corpus e uma nova tarefa, que modelo de tópico devo usar? Como posso decidir quais dos muitos pressupostos de modelagem são importantes para os meus objetivos? Como eu devo me mover entre os muitos tipos de modelos temáticos que foram desenvolvidos? Estas questões têm merecido alguma atenção por parte dos estatísticos,[9,30] mas têm sido menos escrutinadas para a escala

de problemas que a aprendizagem de máquinas resolve. Novas respostas computacionais a estas questões seriam uma contribuição significativa para a modelagem de tópicos.

visualização e interfaces de usuário. Outra direção futura promissora para a modelagem de tópicos é desenvolver novos métodos de interação e visualização de tópicos e corpora. Modelos temáticos pró-vídeo nova estrutura exploratória em grandes coleções Como podemos explorar melhor essa estrutura para ajudar na descoberta e exploração?

Um problema é como exibir os tópicos. Normalmente, exibimos os tópicos listando as palavras mais frequentes de cada um (ver Figura 2), mas novas maneiras de rotular os tópicos - escolhendo palavras diferentes ou exibindo as palavras escolhidas de forma diferente - podem ser mais eficazes. Um outro problema é a melhor forma de exibir um documento com um modelo de tópico. No nível do documento, os modelos de tópicos fornecem informações potencialmente úteis sobre a estrutura do documento. Combinada com etiquetas de tópicos eficazes, esta estrutura poderia ajudar os leitores a identificar as partes mais interessantes do documento. Além disso, as proporções de tópicos ocultos conectam implicitamente cada documento com os outros documentos (considerando uma medida de distância entre as proporções dos tópicos). Como podemos exibir melhor essas conexões? O que é uma interface eficaz para todo o corpus e sua estrutura temática inferida?

Estas são questões de interface do usuário, e são essenciais para a modelagem de tópicos. Algoritmos de modelagem tópica mostram muita promessa para a descoberta de uma estrutura temática de média incipiente em grandes leções de documentos. Mas tornar esta estrutura útil requer uma cuidadosa atenção à visualização da informação e às interfaces de usuário correspondentes. **Modelos temáticos para a descoberta de dados**. Modelos temáticos foram desenvolvidos com aplicações de engenharia de informação em mente. Como modelo estatístico, porém, os modelos tópicos devem ser capazes de nos dizer algo, ou nos ajudar a formar uma hipótese, sobre os dados. O que podemos *aprender* sobre a linguagem (e outros dados) com base no modelo temático posterior? Alguns trabalhos nesta área têm aparecido na ciência política,[19] biblio- métrica,[17] e psicologia. [32] Este tipo de pesquisa adapta os modelos temáticos a uma variável externa de interesse, uma tarefa difícil para a aprendizagem não supervisionada que deve ser cuidadosamente validada.

Em geral, este problema é melhor abordado por cientistas da computação em equipe com outros estudiosos para usar modelos temáticos para ajudar a explorar, visualizar e desenhar hipóteses a partir de seus dados. Além das aplicações científicas, como genética e neurociência, pode-se imaginar modelos temáticos a serviço da história, sociologia, lingüística, ciência política, estudos jurídicos, literatura comparativa e outros campos, onde os textos são um objeto principal de estudo. Trabalhando com estudiosos de diversas áreas, podemos começar a desenvolver uma nova metodologia interdisciplinar computacional para trabalhar e desenhar conclusões a partir de arquivos de textos.

Sumário

Temos pesquisado *modelos temáticos probabilísticos*, um conjunto de algoritmos que fornecem uma solução estatística para o problema da gestão de grandes arquivos de documentos. Com os recentes avanços científicos em apoio à aprendizagem não supervisionada de máquinas - compo- nentes flexíveis para modelagem, algo- rithms escaláveis para inferência posterior e maior acesso a grandes conjuntos de dados - os modelos temáticos prometem ser um componente importante para resumir e entender nosso crescente arquivo digitalizado de informação.

Referências

1. asuncion, a., welling, m., smyth, P., teh, teh, y. on smoothing and inference for topic models. in *Uncertainty in Artificial Intelligence* (2009).
2. bart, e., welling, m., Perona, P. organização não supervisionada de colecções de imagens: taxonomias e mais além. *Trans. Reconhecimento de Padrões. Mach. Intell. 33*, 11 (2010) (2301–2315).
3. blei, D., griffiths, t. , Jordan, m. o processo de restaurante chinês aninhado e inferência bayesiana não-paramétrica de hierarquias de tópicos. *J. ACM 57*, 2 (2010), 1-30.
4. blei, D., Jordan, m. modelagem de dados anotados. em *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2003), acm Press, 127-134.
5. blei, D., lafferty, J. Dynamic topic models. in *International Conference on Machine Learning* (2006), acm, new york, ny, eua, 113-120.
6. blei, D., lafferty, J. um modelo temático correlacionado de ciência. *Ann. Appl. Stat., 1*, 1 (2007), 17-35.
7. blei, D., mcauliffe, J. modelos temáticos supervisionados. em *Sistemas de Processamento de Informação Neural* (2007).
8. blei, D., ng, a., Jordan, m. alocação de Dirichlet latente. *J. Mach. Aprenda. Res. 3* (Janeiro 2003), 993-1022.
9. caixa, g. amostragem e inferência de bayes na modelagem científica e robustez. *J. Roy. Stat. Soc. 143*, 4 (1980), 383-430.
10. Boyd-graber, J., blei, D. syntactic topic models. in *Sistemas de Processamento de Informação Neural* (2009).
11. buntine, w. Variational extensions to em and multinomial Pca. in *European Conference on Machine Learning* (2002).
12. buntine, w., Jakulin, a. Análise discreta de componentes. *Subespaço, Estrutura Latente e Seleção de Recursos*. c. saunders, m. grobelink, s. gunn, e J. shawe-taylor, eds. springer, 2006.

13. chang, J., blei, D. modelos relacionais hierárquicos para redes de documentos. *Ann. Aplic. Stat. 4*, 1 (2010).
14. Deerwester, s., Dumais, s., landauer, t., Furnas, g., harshman, r. indexação por análise semântica latente. *J. Am. Soc. Informar. Sci. 41*, 6 (1990), 391-407.
15. Doyle, g., elkan, c., accounting for burstiness in topic models. in *International Conference on Machine Learning* (2009), acm, 281-288.
16. Fei-Fei, l., Perona, P. a bayesian hierarchical model for learning natural scene categories. in *IEEE Computer Vision and Pattern Recognition* (2005), 524-531.
17. gerrish, s., blei, D. a language-based approach to measuring scholarly impact. in *International Conference on Machine Learning* (2010).
18. griffiths, t., steyvers, m., blei, D., tenenbaum, J. integrando tópicos e sintaxe. *Advances in Neural Information Processing Systems 17*. l. K. saul, y. weiss, e l. bottou, eds. mit Press, cambridge, ma, 2005, 537-544.
19. grimmer, J. a bayesian hierarchical topic model for political texts: measuring expressed agendas expressed expressed in senate press releases. *Polit. Anal. 18*, 1 (2010), 1.
20. hoffman, m., blei, D., bach, F. aprendizagem on-line para alocação de Dirichlet latente. em *Sistemas de Processamento de Informação Neural* (2010).
21. hofmann, t. Probabilistic latent semantic analysis. in *Uncertainty in Artificial Intelligence (UAI)* (1999).
22. Jordan, m., ghahramani, Z., Jaakkola, t., saul, l. introdução aos métodos variacionais para modelos gráficos. *Mach. Aprenda. 37* (1999), 183-233.
23. li, J., wang, c., lim, y., blei, D., Fei-Fei, l., construindo e usando uma hierarquia de imagem semantivisual. em *Computer Vision and Pattern Recognition* (2010).
24. li, w., mccallum, a. Pachinko allocation: Dag- structured mixture models of topic correlations. in *International Conference on Machine Learning* (2006), 577-584.
25. mimno, D., mccallum, a. modelos temáticos condicionados a características arbitrárias com regressão Dirichlet-multinomial. in *Uncertainty in Artificial Intelligence* (2008).
26. newman, D., chemudugunta, c., smyth, P. modelos de entidade estatística-tópicos. in *Knowledge Discovery and Data Mining* (2006).
27. Pritchard, J., stephens, m., Donnelly, P. inferência da estrutura da população usando dados do genótipo multilocus. *Genetics 155* (junho de 2000), 945-959.
28. reisinger, J., waters, a., silverthorn, b., mooney, r. spherical topic models. in *International Conference on Machine Learning* (2010).
29. rosen-Zvi, m., griffiths, t., steyvers, m., smith, P., o modelo autor-tópico para autores e documentos. em *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (2004), auai Press, 487-494.
30. Rubin, D. bayesianamente justificável e relevantes cálculos de frequência para o estatístico aplicado. *Ann. Stat. 12*, 4 (1984), 1151-1172.
31. sivic, J., russell, b., Zisserman, a., Freeman, w., efros, a., descoberta sem supervisão de hierarquias de classes de objetos visuais. em *Conference on Computer Vision and Pattern Recognition* (2008).
32. socher, r., gershman, s., Perotte, a., sederberg, P., blei, D., norman, K. a bayesian analysis of dynamics in free recall. in *Advances in Neural Information Processing Systems 22*. y. bengio, D. schuurmans, J. lafferty, c. K. i. williams, e a. culotta, eds, 2009.
33. steyvers, m., griffiths, t. Modelos temáticos probabilísticos. *Análise Semântica Latente: Um Caminho para o Significado*.
t. landauer, D. mcnamara, s. Dennis, e w. Kintsch, eds. lawrence erlbaum, 2006.
34. os, y., Jordan, m., beal, m., blei, D. processos hierárquicos Dirichlet. *J. Am. Stat. Assoc. 101*, 476 (2006), 1566-1581.
35. wainwright, m., Jordan, m. modelos gráficos, famílias exponenciais, e inferência variacional. *Encontrado. Trends Mach. Aprenda. 1*(1–2) (2008), 1–305.
36. wallach, h. topic modeling: beyond bag of words. in *Proceedings of the 23rd International Conference on Machine Learning* (2006).
37. wang, c., blei, D. Desacoplamento da esparsidade e suavidade no discreto processo hierárquico Dirichlet. *Avanços nos Sistemas de Processamento de Informação Neural 22*. y. bengio, D. schuurmans, J. lafferty, c. K. i. williams, e a. culotta, eds. 2009, 1982-1989.
38. wang, c., thiesson, b., meek, c., blei, D. markov modelos temáticos. em *Artificial Intelligence and Statistics* (2009).

David M. Blei (blei@cs.princeton.edu) é professor associado no departamento de informática da Universidade de Princeton, Princeton, n. J.