Para dentro: Em T. Landauer, D. McNamara, S. Dennis e W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning (Um Caminho para o Significado)*. Laurence Erlbaum

Modelos Tópicos Probabilísticos

Mark Steyvers

University of California, Irvine

Universidade

Tom Griffiths

Brown

Envie Correspondência

para: Mark Steyvers
Departamento de Ciências
Cognitivas 3151 Social Sciences
Plaza University of California,
Irvine Irvine, CA 92697-5100
Email: msteyver@uci.edu

1. Introdução

Muitos capítulos deste livro ilustram que a aplicação de um método estatístico como a Análise Semântica Latente (LSA; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998) a grandes bases de dados pode produzir uma visão da cognição humana. A abordagem LSA faz três afirmações: que a informação semântica pode ser derivada de uma matriz de co-ocorrência de documentos de palavras; que a redução da dimensionalidade é uma parte essencial dessa derivação; e que palavras e documentos podem ser representados como pontos no espaço euclidiano. Neste capítulo, buscamos uma abordagem coerente com as duas primeiras dessas afirmações, mas diferente no terceiro, descrevendo uma classe de modelos estatísticos em que as propriedades semânticas das palavras e dos documentos são expressas em termos de tópicos probabilísticos.

Modelos temáticos (por exemplo, Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2002; 2003; 2004; Hofmann, 1999; 2001) baseiam-se na ideia de que os documentos são misturas de tópicos, em que um tópico é uma distribuição de probabilidade sobre as palavras. Um modelo temático é um *modelo generativo* para documentos: especifica um procedimento probabilístico simples pelo qual documentos podem ser gerados. Para fazer um novo documento, escolhe-se uma distribuição por tópicos. Então, para cada palavra desse documento, escolhe-se um tópico aleatoriamente de acordo com essa distribuição, e extrai-se uma palavra desse tópico. Técnicas estatísticas padrão podem ser usadas para inverter esse processo, inferindo o conjunto de tópicos que foram responsáveis pela geração de uma coleção de documentos. A Figura 1 mostra quatro exemplos de tópicos que foram derivados do corpus da TASA, uma coleção de mais de 37.000 passagens de texto de materiais educacionais (por exemplo, linguagem e artes, estudos sociais, saúde, ciências) coletados pela Touchstone Applied Science Associates (ver Landauer, Foltz, & Laham, 1998). A figura mostra as dezesseis palavras que têm a maior probabilidade sob cada tópico. As palavras nestes tópicos estão relacionadas ao uso de drogas, cores, memória e mente, e visitas médicas. Documentos com diferentes conteúdos podem ser gerados escolhendo diferentes distribuições ao longo dos tópicos. Por exemplo, ao dar probabilidade igual aos dois primeiros tópicos, pode-se construir um documento sobre uma pessoa que tomou muitas drogas, e como isso afetou a percepção das cores. Ao dar a mesma probabilidade aos dois últimos tópicos, pode-se construir um documento sobre uma pessoa que sofreu uma perda de memória, o que exigiu uma visita ao médico.

Tópico 247Topic 5Topic 43Topic 56

palavra	sonda.
DRUGS	.069
DRUG	.060
MEDICINA	.027
EFEITOS	.026
CARROÇARIA	.023
MEDICAMENTOS	.019
DOR	.016
PESSOA	.016
MARIJUANA	.014
LABEL	.012
ÁLCOOL	.012
PERIGOSO	.011
ABUSO	.009
EFEITO	.009
CONHECIDO	.008
PILLS	.008

palavra	sonda.
VERMELHO	.202
AZUL	.099
VERDE	.096
AMARELO	.073
BRANCO	.048
COLOR	.048
BRILHO	.030
COLORS	.029
ORANGE	.027
COROA	.027
PINK	.017
LOOK	.017
PRETO	.016
PROPRIEDADE	.015
CROSS	.011
CORADO	.009

palavra	sonda.
MENTE	.081
THOUGHT	.066
LEMBRE-SE	.064
MEMÓRIA	.037
PENSAGEM	.030
PROFESSOR	.028
FELT	.025
REMETIDO	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENTO	.020
PENSAGEM	.019
O QUE É	.016
MARAVILHA	.014
ESQUEÇA	.012
LEMBRE-SE	.012

palavra	sonda.
DOCTOR	.074
DR.	.063
PATIENTE	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NÚMERO	.031
PATIENTES	.029
DOUTORES	.028
SAÚDE	.025
MEDICINA	.017
NURSING	.017
DENTAL	.015
NOLSES	.013
PHYSICIAN	.012
HOSPITAIS	.011

Figura 1. Uma ilustração de quatro (de 300) tópicos extraídos do corpus da TASA.

Representar o conteúdo de palavras e documentos com tópicos probabilísticos tem uma vantagem distinta sobre uma representação puramente espacial. Cada tópico é interpretável individualmente, proporcionando uma distribuição de probabilidade sobre palavras que escolhe um conjunto coerente de termos correlacionados. Enquanto a Figura 1 mostra apenas quatro dos 300 tópicos que foram derivados, os tópicos são tipicamente tão interpretáveis como os que são mostrados aqui. Isto contrasta com os eixos arbitrários de uma representação espacial, e pode ser extremamente útil em muitas aplicações (por exemplo, Griffiths & Steyvers, 2004; Rosen-Zvi, Griffiths, Steyvers, &

Smyth, 2004; Steyvers, Smyth, Rosen-Zvi, & Griffiths, 2004).

O plano deste capítulo é o seguinte. Primeiro, descrevemos com mais detalhes as idéias-chave por trás dos modelos temáticos e delineamos como é possível identificar os tópicos que aparecem em um conjunto de documentos. Em seguida, discutimos os métodos para

respondendo a dois tipos de semelhanças: avaliar a semelhança entre dois documentos, e avaliar a semelhança associativa entre duas palavras. Encerramos considerando como os modelos generativos têm o potencial de fornecer uma visão mais profunda da cognição humana.

2. Modelos Generativos

Um modelo generativo para documentos é baseado em regras simples de amostragem probabilística que descrevem como as palavras em documentos podem ser geradas com base em variáveis latentes (aleatórias). Ao encaixar um modelo generativo, o objetivo é encontrar o melhor conjunto de variáveis latentes que possam explicar os dados observados (ou seja, palavras observadas em documentos), assumindo que o modelo realmente gerou os dados. A Figura 2 ilustra a abordagem da modelagem temática de duas formas distintas: como modelo generativo e como um problema de inferência estatística. À esquerda, o processo generativo é ilustrado com dois tópicos. Os tópicos 1 e 2 são tematicamente relacionados com dinheiro e rios e são ilustrados como sacos contendo diferentes distribuições por palavras. Diferentes documentos podem ser produzidos escolhendo palavras de um tópico, dependendo do peso dado ao tópico. Por exemplo, os documentos 1 e 3 foram gerados por amostragem apenas dos tópicos 1 e 2 respectivamente, enquanto o documento 2 foi gerado por uma mistura igual dos dois tópicos. Observe que os números sobrescritos associados às palavras nos documentos indicam qual tópico foi usado para amostrar a palavra. Da forma como o modelo é definido, não há noção de exclusividade mútua que restrinja as palavras a fazer parte de apenas um tópico. Isto permite que os modelos de tópicos capturem polissemia, onde a mesma palavra tem múltiplos significados. Por exemplo, tanto o tópico dinheiro como o tópico rio podem dar alta probabilidade à palavra BANCO, o que é sensato dada a natureza polissêmica da palavra.

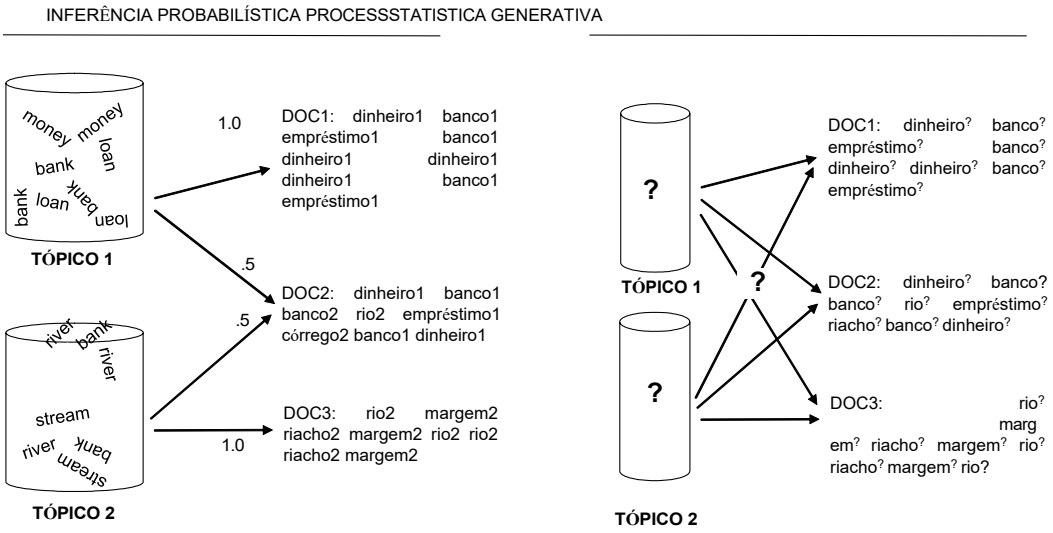


Figura 2. Ilustração do processo generativo e do problema de inferência estatística subjacente aos modelos temáticos

O processo generativo aqui descrito não faz nenhuma suposição sobre a ordem das palavras como elas aparecem nos documentos. A única informação relevante para o modelo é o número de vezes que as palavras são produzidas. Isto é conhecido como a *suposição do saco de palavras*, e é comum a muitos modelos estatísticos de linguagem, incluindo o LSA. Naturalmente, a informação por ordem de palavras pode conter indicações importantes para o conteúdo de um documento e esta informação não é utilizada pelo modelo. Griffiths, Steyvers, Blei e Tenenbaum (2005) apresentam uma extensão do modelo tópico que é sensível à ordem de palavras e aprende automaticamente os factores sintácticos e semânticos que orientam a escolha das palavras (ver também Dennis, este livro para uma abordagem diferente a este problema).

O painel direito da Figura 2 ilustra o problema da inferência estatística. Dadas as palavras observadas em um conjunto de documentos, gostaríamos de saber qual modelo de tópico é mais provável que tenha gerado os dados. Isto envolve inferir a distribuição de probabilidade sobre palavras associadas a cada tópico, a distribuição sobre tópicos para cada documento e, muitas vezes, o tópico responsável pela geração de cada palavra.

3. Modelos Tópicos Probabilísticos

Uma variedade de modelos tópicos probabilísticos tem sido utilizada para analisar o conteúdo de documentos e o significado das palavras (Blei et al., 2003; Griffiths e Steyvers, 2002; 2003; 2004; Hofmann, 1999; 2001). Todos estes modelos utilizam a mesma ideia fundamental - que um documento é uma mistura de tópicos - mas fazem pressupostos estatísticos ligeiramente diferentes. Para introduzir a notação, vamos escrever $P(z)$ para a distribuição sobre tópicos z *num* determinado documento e $P(w \mid z)$ para a distribuição de probabilidade sobre palavras w dado tópico z. Várias distribuições de palavras tópico $P(w \mid z)$ foram ilustradas nas Figuras 1 e 2, cada uma dando um peso diferente às palavras tematicamente relacionadas. Cada palavra w_i em um documento (onde o índice se refere ao símbolo da i-ésima palavra) é gerada primeiro por uma amostragem de um tópico da distribuição de tópicos, depois escolhendo uma palavra da distribuição de palavras do tópico. Escrevemos $P(z_i = j)$ como a probabilidade de que o *jth* tópico foi amostrado para o ith word token e $P(w_i \mid z_i = j)$ como a probabilidade da palavra w_i sob o tópico j. O modelo especifica a seguinte distribuição sobre as palavras dentro de um documento:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j) \tag{1}$$

onde T é o número de tópicos. Para simplificar a notação, deixe $\phi^{(j)} = P(w \mid z=j)$ referir-se à distribuição multinomial sobre palavras para o tópico j e $\theta^{(d)} = P(z)$ referir-se à distribuição multinomial sobre tópicos para o documento d. Além disso, suponha que a coleção de textos consiste em documentos D e cada documento d consiste em N_d fichas de palavras. Que N seja o número total de fichas palavras (i.e., $N = N_d$). Os parâmetros ϕ e θ indicar quais palavras são importantes para cada tópico e quais tópicos são importantes para um determinado documento, respectivamente.

Hofmann (1999; 2001) introduziu a abordagem temática probabilística da modelagem de documentos no seu método de Indexação Semântica Latente Probabilística (pLSI; também conhecido como o modelo de aspecto). O modelo pLSI não faz nenhuma suposição sobre como os pesos de mistura são gerados, tornando difícil testar a generalização do modelo para novos documentos. Blei et al. (2003) estenderam este modelo, introduzindo um Dirichlet prévio em θ, chamando o modelo generativo resultante de Latent Dirichlet Allocation (LDA). Como um prior conjugado para a multinomial, a distribuição Dirichlet é uma escolha conveniente como anterior, simplificando o problema da inferência estatística. A densidade de probabilidade de uma distribuição Dirichlet dimensional T sobre a distribuição multinomial $p=(p_1, ..., p_T)$ é definida por:

$$\mathrm{Dir}(\alpha) = \frac{\Gamma\left(\sum_{j=1}^{T} \alpha_j\right)}{\prod_{j=1}^{T} \Gamma(\alpha_j)} \prod_{j=1}^{T} \pi_j^{\alpha_j - 1} \tag{2}$$

Os parâmetros desta distribuição são especificados por $\alpha_1 ... \alpha_T$. Cada hiperparâmetro α_j pode ser interpretado como uma contagem de observação prévia para o número de vezes que o tópico j é amostrado em um documento, antes de ter observado qualquer palavra real desse documento. É conveniente usar uma distribuição Dirichlet simétrica com um único hiperparâmetro α tal que $\alpha_1 = \alpha_2 = ... = \alpha_T = \alpha$. Ao colocar um Dirichlet prévio sobre a distribuição temática θ, o resultado é uma distribuição temática suavizada, com a quantidade de suavização determinada pelo parâmetro α. A Figura 3 ilustra a distribuição de Dirichlet para três tópicos em um simplex bidimensional. O simplex é um sistema de coordenadas conveniente para expressar todas as distribuições de probabilidade possíveis -- para qualquer ponto $p = (p_1, ..., p_T)$ no simplex, nós temos $\sum_j p_j = 1$. O Dirichlet anterior sobre as distribuições de tópicos pode ser interpretado como forças sobre as combinações de tópicos com o α superior afastando os tópicos dos cantos do simplex, levando a uma maior suavização (compare o painel esquerdo e direito). Para α <1, os modos da distribuição Dirichlet estão localizados nos cantos do simplex. Neste regime (muitas vezes utilizado na prática), há um viés para a esparsidade, e a pressão é para escolher distribuições temáticas que favoreçam apenas alguns tópicos.

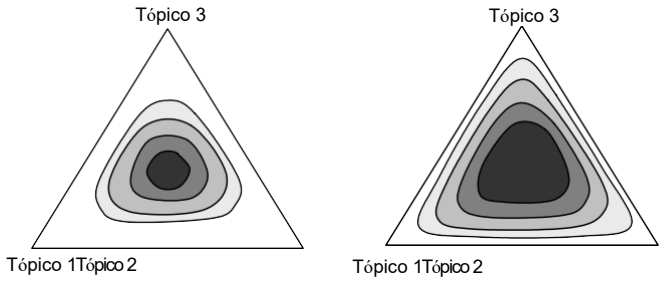


Figura 3. Ilustrando a distribuição simétrica de Dirichlet para três tópicos em um simplex bidimensional. Cores mais escuras indicam maior probabilidade. Esquerda: $\alpha = 4$. Direita: $\alpha = 2$.

Griffiths e Steyvers (2002; 2003; 2004) exploraram uma variante deste modelo, discutida por Blei et al. (2003), colocando também um Dirichlet simétrico *(*∀prior on"). O hiperparâmetro pode ser interpretado como a contagem de observação prévia sobre o número de palavras que são amostradas de um tópico antes de qualquer palavra do corpus ser observada. Isto suaviza a distribuição das palavras em cada tópico, com a quantidade de suavização determinada por Boas escolhas para os hiperparâmetros e dependerá do número de tópicos e do tamanho do vocabulário. A partir de pesquisas anteriores, encontramos =50/T e = 0,01 para trabalhar bem com muitas coleções de textos diferentes.

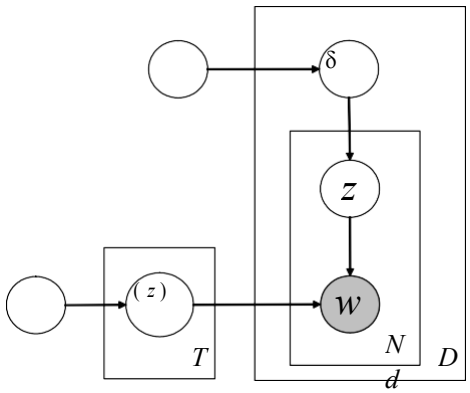


Figura 4. O modelo gráfico para o modelo tópico usando notação de placa.

Modelo gráfico. Modelos generativos probabilísticos com etapas de amostragem repetidas podem ser convenientemente ilustrados usando notação de placa (ver Buntine, 1994, para uma introdução). Nesta notação gráfica, as variáveis sombreadas e não sombreadas indicam variáveis observadas e latentes (ou seja, não observadas) respectivamente. As variáveis e , assim como *z* (a atribuição das palavras tokens aos tópicos) são os três conjuntos de variáveis latentes que gostaríamos de inferir. Como discutido anteriormente, tratamos os hiperparâmetros e como constantes no modelo. A Figura 4 mostra o modelo gráfico do modelo tópico utilizado em Griffiths & Steyvers (2002; 2003; 2004). As setas indicam dependências condicionais entre as variáveis enquanto as placas (as caixas na figura) referem-se a repetições de etapas de amostragem com a variável no canto inferior direito referindo-se ao número de amostras. Por exemplo, a placa interna sobre *z* e *w* ilustra a amostragem repetida de tópicos e palavras até que N_d palavras tenham sido geradas para o documento *d*. A placa ao redor de$^{(d)}$ ilustra a amostragem de uma distribuição por tópicos para cada documento *d* para um total de documentos *D.* A placa ao redor de$^{(z)}$ ilustra a amostragem repetida de distribuições de palavras para cada tópico *z até que* os tópicos *T tenham sido* gerados.

Interpretação Geométrica. O modelo temático probabilístico tem uma interpretação geométrica elegante como mostra a Figura 5 (seguindo Hofmann, 1999). Com um vocabulário contendo *W* tipos de palavras distintos, um espaço dimensional *W* pode ser construído onde cada eixo representa a probabilidade de observar um determinado tipo de palavra. O simplex dimensional *W-1* representa todas as distribuições de probabilidade sobre as palavras. Na Figura 5, a região sombreada é o simplex bidimensional que representa todas as distribuições de probabilidade sobre três palavras. Como uma distribuição de probabilidade sobre as palavras, cada

documento na coleção de textos pode ser representado como um ponto no simplex. Da mesma forma, cada tópico também pode ser representado como um ponto no simplex. Cada documento que é gerado pelo modelo é uma combinação convexa dos tópicos *T* que não só coloca todas as distribuições de palavras geradas pelo modelo como pontos no simplex dimensional *W-1*, mas também como pontos no simplex dimensional *T-1* englobados pelos tópicos. Por exemplo, na Figura 5, os dois tópicos abrangem um simplex unidimensional e cada documento gerado fica no segmento de linha entre as localizações dos dois tópicos. O Dirichlet anterior sobre as distribuições de palavras dos tópicos pode ser interpretado como forças sobre as localizações dos tópicos com maior deslocamento das localizações dos tópicos para longe dos cantos do simplex.

Quando o número de tópicos é muito menor do que o número de tipos de palavras (ou seja, $T << W$), os tópicos abrangem um sub-simplexo de baixa dimensão e a projeção de cada documento no sub-simplexo de baixa dimensão pode ser pensada como redução da dimensionalidade. Esta formulação do modelo é semelhante à da Análise Semântica Latente. Buntine (2002) apontou correspondências formais entre modelos temáticos e análise de componentes principais, um procedimento intimamente relacionado com a Análise Semântica Latente.

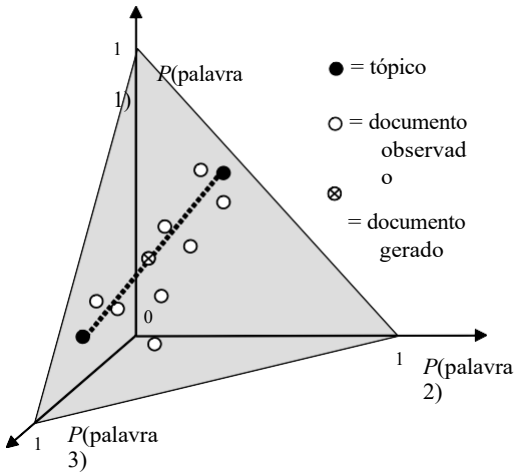


Figura 5. Uma interpretação geométrica do modelo temático.

Interpretação de Factorização Matricial. No LSA, uma matriz de co-ocorrência de documentos de palavras pode ser decomposta por decomposição de valores singulares em três matrizes (ver outro capítulo deste livro de Martin & Berry): uma matriz de vectores de palavras, uma matriz diagonal com valores singulares e uma matriz com vectores de documentos. A Figura 6 ilustra esta decomposição. O modelo temático também pode ser interpretado como factorização matricial, como apontado por Hofmann (1999). No modelo
descrita acima, a matriz de co-ocorrência do documento de trabalho é dividida em duas partes: uma matriz de tópicos e um documento

matriz . Note que a matriz diagonal D em LSA pode ser absorvida na matriz U ou V, tornando a semelhança entre as duas representações ainda mais clara.

Esta factorização evidencia uma semelhança conceptual entre a LSA e os modelos temáticos, encontrando ambos uma representação de baixa dimensão para o conteúdo de um conjunto de documentos. No entanto, também mostra várias diferenças importantes entre as duas abordagens. Nos modelos temáticos, a palavra e os vetores dos documentos das duas matrizes decompostas são distribuições de probabilidade com a restrição de que os valores das características sejam não-negativos e somados a um. No modelo LDA, restrições adicionais a priori são colocadas sobre as distribuições de palavras e tópicos. Não existe tal restrição nos vetores LSA, embora existam outras técnicas de factorização de matrizes que requerem valores de característica não negativos (Lee & Seung, 2001). Em segundo lugar, a decomposição de LSA fornece uma base orto-normal que é computacionalmente conveniente porque uma decomposição para as dimensões *T* dará simultaneamente todas as aproximações dimensionais inferiores também. No modelo temático, as distribuições de palavras-palavra são independentes mas não ortogonais; a inferência do modelo tem de ser feita separadamente para cada dimensionalidade.

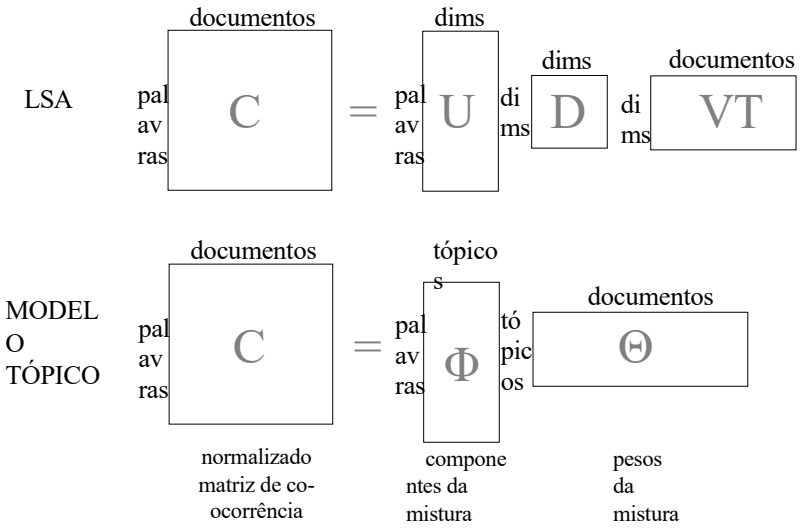


Figura 6. A factorização matricial do modelo LSA em comparação com a factorização matricial do modelo tópico

Outras aplicações. O modelo estatístico subjacente à abordagem de modelagem temática foi ampliado para incluir outras fontes de informação sobre documentos. Por exemplo, Cohn e Hofmann (2001) estenderam o modelo pLSI, integrando conteúdo e informações de links. Em seu modelo, os tópicos estão associados não apenas a uma distribuição de probabilidade sobre termos, mas também sobre hiperlinks ou citações entre documentos. Recentemente, Steyvers, Smyth, Rosen- Zvi, e Griffiths (2004) e Rosen-Zvi, Griffiths, Steyvers, e Smyth (2004) propuseram o modelo autor-tópico, uma extensão do modelo LDA que integra informação de autoria com conteúdo. Ao invés de associar cada documento a uma distribuição por tópicos, o modelo autor-tópico associa cada autor a uma distribuição por tópicos e assume que cada documento de múltiplos autores expressa uma mistura das misturas de tópicos dos autores. O modelo estatístico subjacente ao modelo temático também tem sido aplicado a outros dados que não texto. Os modelos de grade-of-membership (GoM) desenvolvidos pelos estatísticos na década de 1970 são de forma semelhante (Manton, Woodbury, & Tolley, 1994), e Erosheva (2002) considera um modelo de GoM equivalente a um modelo tópico. O mesmo modelo tem sido usado para análise de dados em genética (Pritchard, Stephens, & Donnelly, 2000).

4. Algoritmo para Tópicos de Extração

As principais variáveis de interesse no modelo são as distribuições de palavras-palavras temáticas e as distribuições temáticas para cada documento. Hofmann (1999) usou o algoritmo de expectativa-maximização (EM) para obter estimativas diretas de e . Esta abordagem sofre de problemas envolvendo máximos locais da função de probabilidade, o que motivou uma busca por melhores algoritmos de estimação (Blei et al., 2003; Buntine, 2002; Minka, 2002). Ao invés de estimar diretamente as distribuições de palavras-palavras e tópicos para cada documento, outra abordagem é estimar diretamente a distribuição posterior sobre z (a atribuição de fichas de palavras a tópicos), dadas as palavras observadas $\boldsymbol{w}$, ao mesmo tempo em que se marginaliza e . Cada z_i dá um valor inteiro [1... T] para o tópico ao qual a palavra token i é atribuída. Como muitas coleções de texto contêm milhões de palavras token, a estimativa do posterior sobre $\boldsymbol{z}$ requer procedimentos eficientes de estimativa. Descreveremos um algoritmo que usa a amostragem de Gibbs, uma forma de cadeia de Markov Monte Carlo, que é fácil de implementar e fornece um método relativamente eficiente de extrair um conjunto de tópicos de um grande corpus (Griffiths & Steyvers, 2004; ver também Buntine, 2004, Erosheva 2002 e Pritchard et al., 2000). Mais informações sobre outros algoritmos de extração de tópicos de um corpus podem ser obtidas nas referências dadas acima.

A cadeia Markov Monte Carlo (MCMC) refere-se a um conjunto de técnicas iterativas aproximadas concebidas para amostrar valores de distribuições complexas (muitas vezes de alta dimensão) (Gilks, Richardson, & Spiegelhalter, 1996). A amostragem de Gibbs (também conhecida como amostragem condicional alternada), uma forma específica de MCMC, simula uma distribuição altamente dimensional por amostragem em subconjuntos de variáveis de dimensões inferiores, onde cada subconjunto é condicionado no valor de todos os outros. A amostragem é feita sequencialmente e prossegue até que os valores amostrados se aproximem da distribuição alvo. Enquanto o procedimento de Gibbs que descreveremos não fornece estimativas diretas e , mostraremos como e pode ser

aproximado usando estimativas posteriores de $\boldsymbol{z}$.

O algoritmo de amostragem de Gibbs. Representamos a coleção de documentos por um conjunto de índices de palavras w_i e índices de documentos d_i , para cada palavra token *i.* O procedimento de amostragem de Gibbs considera cada palavra token na coleção de texto por sua vez, e estima a probabilidade de atribuir a palavra token atual a cada tópico, condicionada na atribuição do tópico a todos os outros tokens de palavras. A partir desta distribuição condicional, um tópico é amostrado e armazenado como a nova atribuição de tópico para esta palavra token. Escrevemos esta distribuição condicional como $P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i,)$, onde $z_i = j$ representa a atribuição de tópico do token i ao tópico j, $\mathbf{z}_{-i}$ refere-se à atribuição de tópico de todas as outras palavras tokens, e "" refere-se a todas as outras informações conhecidas ou observadas, tais como todos os outros índices de palavras e documentos $\boldsymbol{w}_{-i}$ e $\boldsymbol{d}_{-i}$, e hiperparâmetros , e . Griffiths e Steyvers (2004) mostraram como isto pode ser calculado por:

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \ldots) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \cdot \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{3}$$

onde $\mathbf{C}^{WT}$ e $\mathbf{C}^{DT}$ são matrizes de contagens com dimensões L x T e D x T respectivamente; C^{WT}_{wj} contém o número de vezes a palavra w é atribuída ao tópico j, não incluindo a instância atual i e C^{DT}_{dj} contém o número de vezes O tópico j é atribuído a algum símbolo de palavra no documento d, não incluindo a instância atual i. Note que a Equação 3 dá a probabilidade não normalizada. A probabilidade real de atribuir uma palavra token ao tópico j é calculada dividindo a quantidade na Equação 3 para o tópico t pela soma em todos os tópicos T.

Os factores que afectam a atribuição de tópicos para uma determinada palavra simbólica podem ser compreendidos examinando as duas partes da Equação 3. A parte esquerda é a probabilidade da palavra w sob o tópico j enquanto a parte direita é a probabilidade que o tópico j tem sob a atual distribuição de tópicos para o documento d. Uma vez que muitos tokens de uma palavra tenham sido atribuídos ao tópico j (entre documentos), aumentará a probabilidade de atribuir qualquer token particular dessa palavra ao tópico j. Ao mesmo tempo, se o tópico j *tiver* sido usado várias vezes em um documento, aumentará a probabilidade de que qualquer palavra desse documento seja atribuída ao tópico j. Portanto, as palavras são atribuídas aos tópicos dependendo da probabilidade que a palavra tem para um tópico, assim como o quão dominante um tópico é em um documento.

O algoritmo de amostragem de Gibbs começa atribuindo cada palavra simbólica a um tópico aleatório em [1... T]. Para cada palavra token, as matrizes de contagem $\mathbf{C}^{WT}$ e $\mathbf{C}^{DT}$ são primeiro decrescidas por uma para as entradas que correspondem à atribuição do tópico atual. Em seguida, um novo tópico é amostrado da distribuição na Equação 3 e as matrizes de contagem $\mathbf{C}^{WT}$ e $\mathbf{C}^{DT}$ são incrementadas com a nova atribuição de tópicos. Cada amostra de Gibbs consiste no conjunto de atribuições de tópicos a todos os N símbolos de palavras do corpus, alcançado por uma única passagem por todos os documentos. Durante a fase inicial do processo de amostragem (também conhecido como período de queima), as amostras de Gibbs têm que ser descartadas porque são estimativas pobres do posterior. Após o período de queima, as sucessivas amostras de Gibbs começam a aproximar-se da distribuição alvo (ou seja, a distribuição posterior sobre as atribuições dos tópicos). Neste ponto, para obter um conjunto representativo de amostras desta distribuição, um número de amostras de Gibbs é guardado em intervalos regularmente espaçados, para evitar correlações entre amostras (ver Gilks et al. 1996).

Estimar e . O algoritmo de amostragem dá estimativas directas de z para cada palavra. Contudo, muitas aplicações do modelo requerem estimativas $'$ e $'$ das distribuições de palavras-tópicos e distribuições de documentos temáticos, respectivamente. Estas podem ser obtidas a partir das matrizes de contagem da seguinte forma:

$$'(j)_i = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \qquad '(d)_j = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha} \tag{4}$$

Estes valores correspondem às distribuições preditivas da amostragem de um novo símbolo da palavra i do tópico j, e da amostragem de um novo símbolo (ainda não observado) no documento d do tópico j, e são também o meio posterior destas quantidades condicionadas a uma determinada amostra $\boldsymbol{z}$.

Um exemplo. O algoritmo de amostragem de Gibbs pode ser ilustrado pela geração de dados artificiais a partir de um modelo temático conhecido e pela aplicação do algoritmo para verificar se ele é capaz de inferir a estrutura generativa original. Nós ilustramos isto expandindo o exemplo que foi dado na Figura 2. Suponha que o tópico 1 dá igual probabilidade às palavras DINHEIRO, EMPRÉSTIMO e BANCO, i.e., $(^{1})$

(1) *EMPRÉSTIMO* (1) *BANCO* 1/ 3, enquanto que o tópico 2 dá igual probabilidade às palavras RIVER, STREAM, e BANK, ou seja, $(^{2)}_{RIVER}$ (2) *STREAM* (2) *BANCO* 1/ 3 . A Figura 7, painel superior, mostra como 16 documentos podem ser

gerada pela mistura arbitrária dos dois tópicos. Cada círculo corresponde a um único símbolo de palavra e cada linha a um documento (por exemplo, o documento 1 contém 4 vezes a palavra BANCO). Na Figura 7, a cor dos círculos indica as atribuições dos tópicos (preto = tópico 1; branco = tópico 2). No início da amostragem (painel superior), as atribuições ainda não mostram nenhuma estrutura; estas apenas refletem as atribuições aleatórias aos tópicos. O painel inferior mostra o estado do coletor de amostras Gibbs após 64 iterações. Com base nestas atribuições, a Equação 4 dá as seguintes estimativas para o
distribuições sobre palavras para tópico 1 e 2: $\hat{\varphi}^{(1)}_{DINHEIRO} .32$, $\hat{\varphi}^{(1)}_{EMPRÉSTIMO} .29$, $\hat{\varphi}^{(1)}_{BANCO} .39$ e
$\hat{\varphi}^{(2)}_{RIVER} .25$, $\hat{\varphi}^{(2)}_{STREAM} .4$, $\hat{\varphi}^{(2)}_{BANCO} .35$. Dado o tamanho do conjunto de dados, estas estimativas são razoáveis.
reconstruções dos parâmetros utilizados para gerar os dados.

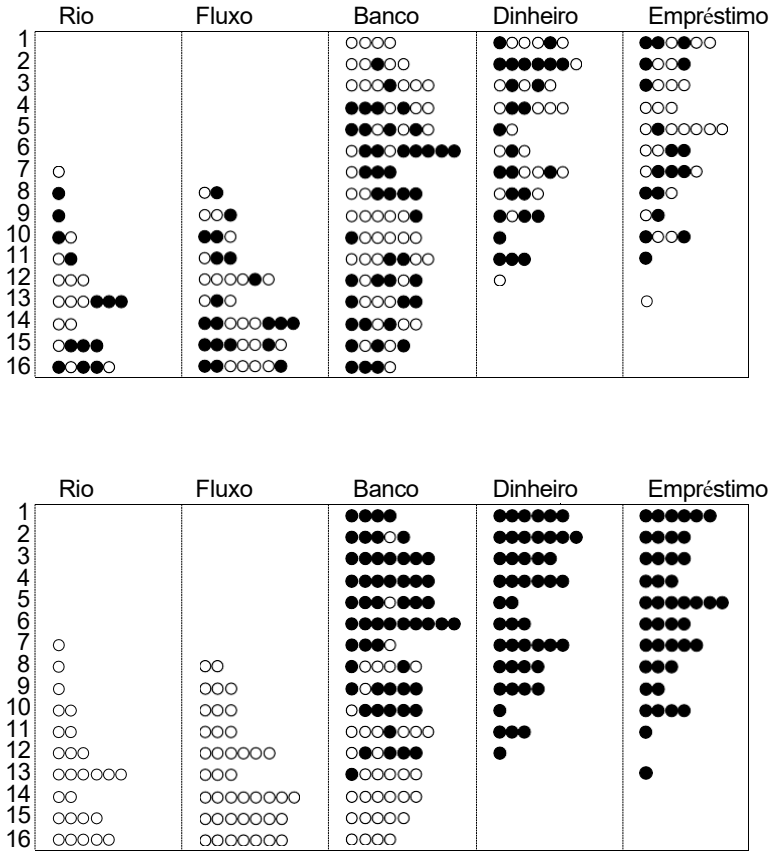


Figura 7. Um exemplo do procedimento de amostragem de Gibbs.

Permutabilidade dos tópicos. Não existe uma ordenação a priori dos tópicos que tornará os tópicos identificáveis entre ou mesmo dentro das execuções do algoritmo. O tópico *j* em uma amostra de Gibbs não está teoricamente limitado a ser semelhante ao tópico *j* em outra amostra, independentemente de as amostras serem provenientes da mesma ou diferente cadeia de Markov (ou seja, amostras espaçadas entre si que começaram com a mesma atribuição aleatória ou amostras de diferentes atribuições aleatórias). Portanto, as diferentes amostras *não podem* ser calculadas como média no nível dos tópicos. No entanto, quando os tópicos são usados para calcular uma estatística invariável à ordenação dos tópicos, torna-se possível e até importante fazer uma média sobre diferentes amostras de Gibbs (ver Griffiths e Steyvers, 2004). É provável que a média do modelo melhore os resultados porque permite a amostragem a partir de múltiplos modos locais do posterior.

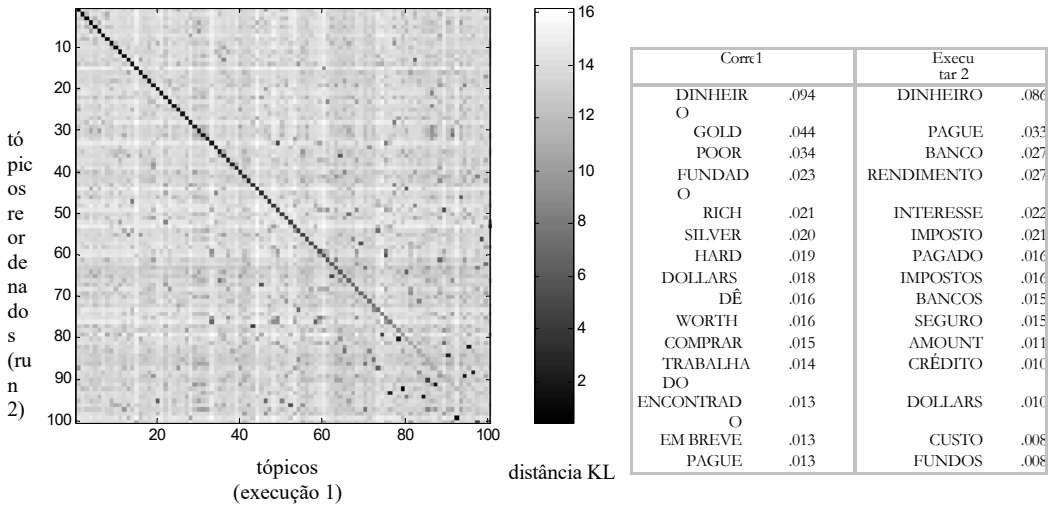
Estabilidade dos Tópicos. Em algumas aplicações, é desejável focar em uma única solução tópica a fim de interpretar cada tópico individual. Nessa situação, é importante saber quais tópicos são estáveis e reaparecerão entre amostras e quais tópicos são idiossincráticos para uma determinada solução. Na Figura 8, é mostrada uma análise do

grau em que duas soluções tópicas podem ser alinhadas entre amostras de diferentes cadeias de Markov. O corpus da TASA foi tomado como

entrada (*W=26*.414; *D=37*.651; *N=5*.628.867; *T=100*; =50/T=.5; =.01) e uma única amostra de Gibbs foi retirada após 2000 iterações para duas inicializações aleatórias diferentes. O painel esquerdo mostra uma matriz de similaridade entre as duas soluções tópicas. A dissemelhança entre os tópicos j_1 e j_2 foi medida pela distância simetrizada de Kullback Liebler (KL) entre as distribuições dos tópicos:

$$KL\, j_1, j_2 \quad \frac{1}{2} \sum_{k=1}^{W} \log 2 \left(\beta_k^{(j)} \,/\, \gamma_k \right) 2 \quad \frac{1}{2} \sum_{k=1}^{W} \log 2 \left(\beta_k^{(j)} \,/\, \gamma_k \right) \tag{5}$$

onde ' e '" correspondem às distribuições estimadas de palavras-palavras a partir de duas séries diferentes. Os tópicos da segunda execução foram reordenados para corresponder o melhor possível (usando um algoritmo ganancioso) com os tópicos da primeira execução. A correspondência foi medida pela soma (inversa) das distâncias KL na diagonal. A matriz de similaridade na Figura 8 sugere que uma grande percentagem de tópicos contém distribuições semelhantes sobre as palavras. O painel direito mostra o *pior* par de tópicos alinhados com uma distância *KL* de 9,4. Ambos os tópicos parecem estar relacionados com dinheiro, mas enfatizam temas diferentes. No geral, estes resultados sugerem que, na prática, as soluções de diferentes amostras darão resultados diferentes, mas que muitos tópicos são estáveis ao longo das corridas.



Corrc1		Execu tar 2	
DINHEIRO	.094	DINHEIRO	.080
GOLD	.044	PAGUE	.033
POOR	.034	BANCO	.027
FUNDADO	.023	RENDIMENTO	.027
RICH	.021	INTERESSE	.022
SILVER	.020	IMPOSTO	.021
HARD	.019	PAGADO	.016
DOLLARS	.018	IMPOSTOS	.016
DÊ	.016	BANCOS	.015
WORTH	.016	SEGURO	.015
COMPRAR	.015	AMOUNT	.011
TRABALHA	.014	CRÉDITO	.010
DO ENCONTRADO	.013	DOLLARS	.010
EM BREVE	.013	CUSTO	.008
PAGUE	.013	FUNDOS	.008

Figura 8. Estabilidade de tópicos entre diferentes séries.

Determinando o número de tópicos. A escolha do número de tópicos pode afectar a interpretabilidade dos resultados. Uma solução com muito poucos tópicos geralmente resultará em tópicos muito amplos, enquanto uma solução com muitos tópicos resultará em tópicos não interpretáveis que escolhem combinações idiossincráticas de palavras. Há uma série de métodos objetivos para escolher o número de tópicos. Griffiths e Steyvers (2004) discutiram uma abordagem de seleção de modelos Bayesianos. A idéia é estimar a probabilidade posterior do modelo e, ao mesmo tempo, integrar todas as configurações de parâmetros possíveis (ou seja, todas as formas de atribuir palavras aos tópicos). O número de tópicos é então baseado no modelo que leva à maior probabilidade posterior. Outra abordagem é escolher o número de tópicos que levam ao melhor desempenho de generalização para novas tarefas. Por exemplo, um modelo de tópico estimado em um subconjunto de documentos deve ser capaz de prever a escolha de palavras no conjunto restante de documentos. Na lingüística computacional, a medida da perplexidade foi proposta para avaliar a generalização de modelos de texto entre subconjuntos de documentos (por exemplo, ver Blei et al. 2003; Rosen-Zvi et al., 2004). Recentemente, pesquisadores têm usado métodos de estatísticas Bayesianas não paramétricas para definir modelos que automaticamente selecionam o número apropriado de tópicos (Blei, Griffiths, Jordan, & Tenenbaum, 2004; Teh, Jordan, Beal, & Blei, 2004).

5. Polissemia com Tópicos

Muitas palavras em linguagem natural são polissêmicas, tendo múltiplos sentidos; sua ambigüidade semântica só pode ser resolvida por outras palavras no contexto. Modelos temáticos probabilísticos representam a ambiguidade semântica através da incerteza sobre os tópicos. Por exemplo, a Figura 9 mostra 3 tópicos selecionados a partir de

uma solução de 300 tópicos para o corpus TASA (Figura 1

mostrou quatro outros tópicos a partir desta solução). Em cada um destes tópicos, a palavra JOGUE é dada uma probabilidade relativamente alta relacionada com os diferentes sentidos da palavra (*tocar* música, *jogar* teatro, *jogar* jogos).

Tópico 77

palavra	sonda.
MUSICA	.090
DANÇA	.034
SONG	.033
JOGUE	.030
CANTAR	.026
CANTANDO	.026
BANDA	.026
JOGADO	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
JOGANDO	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Tópico 82

palavra	sonda.
LITERATURA	.031
POEM	.028
POETRY	.027
POET	.020
JOGADORES	.019
POEMS	.019
JOGUE	.015
LITERÁRIO	.013
ESCRITORES	.013
DRAMA	.012
WROTE	.012
POETS	.011
ESCRITÓRIO	.011
SHAKESPEARE	.010
ESCREVIDO	.009
ETAPA	.009

Tópico 166

palavra	sonda.
JOGUE	.136
BOLA	.129
JOGO	.065
JOGANDO	.042
HIT	.032
JOGADO	.031
BASEBALL	.027
GAMES	.025
BAT	.019
EXECUÇÃO	.019
ALARGAR	.016
BOLAS	.015
TENNIS	.011
HOME	.010
CATCH	.010
DOMÍNIO	.010

Figura 9. Três tópicos relacionados com a palavra JOGUE.

Documento #29795

Bix beiderbecke, na idade[060] quinze[207] , sentou[174] na encosta[071] de um bluff[055] com vista para[027] o mississippi[137] rio[137] . Ele estava ouvindo[077] a música[077] vinda[009] de um barco de passagem[043] rio. A música[077] já tinha capturado[006] o seu coração[157] assim como o seu ouvido[119] . Era jazz[077] . Bix beiderbecke já tinha tido aulas de música [077077] . Ele mostrou[002] promessa[134] no piano[077] , e seus pais[035] esperava que[268] ele poderia considerar[118] tornar-se um concerto[077] pianista[077] . Mas Bix estava interessado em[268] em outro tipo de música [050077 peça[077]] peça[077]
Ele queria[268] o corneto. E ele queria[268] para jazz[077] ...
para

Documento #1883

Há uma simples razão[050 106] porque há tão poucos períodos[078] de teatro realmente grande[082] em todo o nosso mundo ocidental[046] . Demasiadas coisas[300] têm de vir ao mesmo tempo. Os dramaturgos têm de ter os actores certos[082] , os actores[082] têm de ter as casas de teatro certas, as casas de teatro têm de ter os públicos certos[082] . Devemos lembrar[288] que as peças[082] existem[143] para serem representadas[077] , e não [peça[082]]
apenas[050] para ser lido[254] . (mesmo quando você [peça[082]] para si próprio, tente[288] para o apresentar[062] , para o pôr[174] num lê[254] a junto.) como . em breve[028] assim que palco[078] , como você vai tem para ser uma espécie de[126] do teatral[082] ... realizado[082] , então cerca de

Documento #21359

Jim[296] tem um jogo[166] livro[254] . Jim[296] lê[254] o livro[254] . Jim[296] vê[081] um jogo[166] para um. Jim[296] joga[166] o jogo[166] . Jim[296] gosta de[081] o jogo[166] para um. O jogo[166] livro[254] ajuda[081] jim[296] . Don[180] vem[040] dentro de casa[038] . Don[180] e jim[296] lê[254] o jogo[166] livro[254] . Os rapazes[020] vêem um jogo[166] para dois. Os dois [peça[166]] o jogo[166] . Os meninos[020] [peça[166]] o jogo[166] para dois. meninos[020]
Os rapazes[020] gostam do jogo[166] . Meg[282] vem[040] para dentro de casa[282] . Meg[282] e don[180] e jim[296] lêem[254] o livro[254] . Eles vêem um jogo[166] por três. Meg[282] e don[180] e jim[296] [peça[166]] o jogo[166] . Eles jogam[166] ..

Figura 10. Três documentos da TASA com o *jogo de* palavras.

Em um novo contexto, tendo observado apenas uma única palavra JOGO, haveria incerteza sobre qual destes tópicos

poderia ter gerado esta palavra. Essa incerteza pode ser reduzida pela observação de outras palavras menos ambíguas no contexto. O processo de desambiguação pode ser descrito pelo processo de amostragem iterativa como descrito na seção anterior (Equação 4), onde a atribuição de cada palavra token a um tópico depende das atribuições das outras palavras no contexto. Na Figura 10, são mostrados fragmentos de três documentos da TASA que utilizam o PLAY em

três sentidos diferentes. Os números sobrescritos mostram as atribuições dos tópicos para cada símbolo de palavra. As palavras cinzentas são palavras de paragem ou palavras de muito baixa frequência que não foram utilizadas na análise. O processo de amostragem atribui a palavra PLAY aos tópicos 77, 82, e 166 nos três contextos do documento. A presença de outras palavras menos ambíguas (por exemplo, MÚSICA no primeiro documento) constrói evidências para um tópico em particular no documento. Quando uma palavra tem incerteza sobre tópicos, a distribuição de tópicos desenvolvida para o contexto do documento é o principal fator para desambiguar a palavra.

6. Similitudes Informáticas

O conjunto de tópicos derivados de um corpus pode ser usado para responder a perguntas sobre a similaridade de palavras e documentos: duas palavras são semelhantes na medida em que aparecem nos mesmos tópicos, e dois documentos são semelhantes na medida em que os mesmos tópicos aparecem nesses documentos.

Similaridade entre documentos. A semelhança entre documentos d_1 e d_2 pode ser medida pela semelhança entre suas distribuições temáticas correspondentes $^{(d1)}$e$^{(d_2)}$. Há muitas opções para funções de similaridade

entre as distribuições de probabilidade (Lin, 1991). Uma função padrão para medir a diferença ou *divergência* entre duas distribuições p e q é a divergência de Kullback Leibler (KL),

$$D\,p,\theta\,\varphi 1^{T} p_{j}\log 2\frac{p_{j}}{q_{j}} \tag{6}$$

Esta função não negativa é igual a zero quando para todos j, $p_j = q_j$. A divergência KL é assimétrica e em muitas aplicações, é conveniente aplicar uma medida simétrica baseada na divergência KL:

$$KL\,p,\theta\,\frac{1}{2}D\,p,\theta\,D\,q,\pi \tag{7}$$

Outra opção é aplicar a divergência simetrizada de Jensen-Shannon (JS):

$$JS\,p,\theta\,\frac{1}{2}D\,p,p\,\theta\,\theta\,/\,2\;D\,q,p\,\theta\,/_{2} \tag{8}$$

que mede a semelhança entre p e q através da média de p e q -- duas distribuições p e q serão semelhantes se forem semelhantes à sua média (p+q)/2. Ambas as funções de divergência simetrizadas KL e JS parecem funcionar bem na prática. Além disso, também é possível considerar as distribuições temáticas como vetores e aplicar funções com motivação geométrica como distância Euclidiana, produto ponto ou cosseno.

Para aplicações de recuperação de informação, a comparação de documentos é necessária para recuperar os documentos mais relevantes para uma consulta. A consulta pode ser um (novo) conjunto de palavras produzido por um usuário ou pode ser um documento existente da coleção. Neste último caso, a tarefa é encontrar documentos semelhantes ao documento em questão. Uma abordagem para encontrar documentos relevantes é avaliar a similaridade entre as distribuições temáticas correspondentes à consulta e cada documento candidato d_i , usando uma das funções de similaridade distributiva como discutido anteriormente. Outra abordagem (por exemplo, Buntine et al., 2004) é modelar a recuperação de informação como uma consulta probabilística ao modelo do tópico - os documentos mais relevantes são os que maximizam a probabilidade condicional da consulta, dado o documento candidato. Escrevemos isto como $P(\ q \mid d_i\ )$ onde q é o conjunto de palavras contidas na consulta. Usando as suposições do modelo de tópicos, isto pode ser calculado por:

$$P(q\mid_{di})\;{}_{w\,q}P(wk_{\;\mid\;di})\;{}_{k}{}^{T}\;{}_{w\,q}P\;w_{k}\mid z\,\varphi\;P\,z\,j\mid d_{\,i}\;{}_{k}\;{}_{\varphi 1} \tag{9}$$

Note que esta abordagem também enfatiza a semelhança através de tópicos, com documentos relevantes tendo distribuições de tópicos que provavelmente geraram o conjunto de palavras associadas à consulta.

Qualquer que seja a semelhança ou função de relevância utilizada, é importante obter estimativas estáveis para as distribuições temáticas. Isto é especialmente importante para documentos curtos. Com uma única amostra de Gibbs, a distribuição tópica pode ser

influenciado por atribuições idiossincráticas de tópicos para as poucas fichas de palavras disponíveis. Nesse caso, torna-se importante calcular a média da função de similaridade em várias amostras de Gibbs.

Semelhança entre duas palavras. A similaridade entre duas palavras w_1 e w_2 pode ser medida pela medida em que partilham os mesmos tópicos. Usando uma abordagem probabilística, a similaridade entre duas palavras pode ser calculada com base na similaridade entre $\theta^{(1)}$ e $\theta^{(2)}$, as distribuições de tópicos condicionais para palavras w_1 e w_2 onde

$^{(1)} P\, z \mid w_{\ \ i}\, w_1$ e $^{(2)} P\, z \mid w^{\ \ i}\, w_2$. Ou a divergência simetrizada KL ou JS seria apropriado para medir a semelhança de distribuição entre estas distribuições.

Existe uma abordagem alternativa para expressar a semelhança entre duas palavras, enfatizando as relações associativas entre as palavras. A associação entre duas palavras pode ser expressa como uma distribuição condicional sobre potenciais palavras de resposta w_2 para cue word w_1 , i.e. $P(w2 \mid w1)$ -- quais são as prováveis palavras que são geradas como uma resposta associativa a outra palavra? Muitos dados têm sido coletados sobre associação de palavras humanas. Tipicamente, uma palavra taco é apresentada e o sujeito escreve a primeira palavra que vem à mente. Nelson, McEvoy e Schreiber (1998) desenvolveram normas de associação de palavras para mais de 5000 palavras usando centenas de assuntos por palavra taco. Na Figura 9, painel esquerdo, a distribuição das respostas humanas é mostrada para a palavra taco PLAY. As respostas revelam que diferentes sujeitos associam com a taco nos diferentes sentidos da palavra (por exemplo, PLAYBALL e PLAYACTOR). No modelo tópico, a associação de palavras corresponde a ter observado uma única palavra em um novo contexto, e tentar prever novas palavras que possam aparecer no mesmo contexto, com base na interpretação do tópico para a palavra observada. Para um determinado assunto que ativa um único tópico j, a distribuição prevista para w_2 é apenas $P(w2 \mid z\, j)$. Se for assumido que cada assunto ativa apenas um único tópico amostrado da distribuição $P(z\, j \mid w1)$, as distribuições condicionais preditivas podem ser calculadas por:

$$P(w2 \mid w1) \prod_{\varphi 1}^{T} \left(w2 \mid z\, \varphi \right) P(z\, j \mid w1) \tag{10}$$

Na associação humana de palavras, as palavras de alta freqüência são mais propensas a serem usadas como palavras de resposta do que as palavras de baixa freqüência. O modelo captura este padrão porque o termo esquerdo $P(w2 \mid z\, j)$ será influenciado pela palavra freqüência de w_2 - palavras de alta freqüência (em média) têm alta probabilidade condicionada em um tópico.

O painel direito da Figura 9 mostra as previsões do modelo de tópicos para a palavra-chave PLAY usando uma solução de 300 tópicos do corpus da TASA. Griffiths e Steyvers (2002; 2003) compararam o modelo tópico com o LSA na previsão da associação de palavras, descobrindo que o equilíbrio entre a influência da frequência de palavras e a relação semântica encontrada pelo modelo tópico pode resultar em um melhor desempenho do que o LSA nesta tarefa.

HUMANOS		**TÓPICOS**	
DIVERTIDO	.141	BOLA	.036
BOLA	.134	JOGO	.024
JOGO	.074	CRIANÇAS	.016
TRABALHO	.067	TEAM	.011
TERRENO	.060	DESEJA	.010
ESCOLHA	.027	MÚSICA	.010
CRIANÇA	.020	PROJETE	.009
DESFRUTAR	.020	HIT	.009
GANHO	.020	CRIANÇA	.008
ACTOR	.013	BASEBALL	.008
LUTA	.013	GAMES	.007
CAVALO	.013	DIVERTIDO	.007
KID	.013	ETAPA	.007
MÚSICA	.013	DOMÍNIO	.006

Figura 9. Distribuições de respostas observadas e previstas para a palavra JOGAR.

7. Conclusão

Modelos geradores de texto, como o modelo temático, têm o potencial de dar importantes contribuições para a

análise estatística de grandes coleções de documentos e para o desenvolvimento de uma compreensão mais profunda da linguagem humana

aprendizagem e processamento. Estes modelos fazem pressupostos explícitos sobre o processo causal responsável pela geração de um documento e permitem a utilização de métodos estatísticos sofisticados para identificar a estrutura latente que está subjacente a um conjunto de palavras. Consequentemente, é fácil explorar diferentes representações de palavras e documentos, e desenvolver modelos mais ricos capazes de capturar mais do conteúdo da linguagem. Os modelos temáticos ilustram como a utilização de uma representação diferente pode proporcionar novas perspectivas na modelação estatística da linguagem, incorporando muitos dos pressupostos-chave por detrás do LSA, mas tornando possível identificar um conjunto de tópicos probabilísticos interpretáveis em vez de um espaço semântico. Os modelos temáticos também foram alargados para capturar algumas propriedades interessantes da linguagem, tais como as relações semânticas hierárquicas entre palavras (Blei et al., 2004), e a interacção entre sintaxe e semântica (Griffiths et al., 2004). A grande maioria dos modelos generativos está ainda por definir, e a investigação destes modelos proporciona a oportunidade de expandir tanto os benefícios práticos como a compreensão teórica da aprendizagem estatística da língua.

Nota do Autor

Gostaríamos de agradecer a Simon Dennis e Sue Dumais pelos comentários atenciosos que melhoraram este capítulo. As implementações Matlab de uma variedade de modelos temáticos probabilísticos estão disponíveis em: http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

8. Referências

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Alocação de Dirichlet Latente. *Journal of Machine Learning Research, 3,* 993-1022.

Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2004). Modelos hierárquicos de tópicos e o processo de restaurante chinês aninhado. In *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT Press.

Buntine, w. (2002). Extensões Variacionais para EM e PCA Multinomial. In: T. Elomaa et al. (Eds.): ECML, LNAI 2430, 23-34. Springer-Verlag, Berlim.

Buntine, W.L. (1994). Operations for learning with graphical models, Journal of Artificial Intelligence Research 2, 159-225.

Buntine, W., Löfström, J., Perkiö, J., Perttu, S., Poroshin, V., Silander, T., Tirri, H., Tuominen, A., & Tuulos, V. (2004). Um Motor de Busca de Código Aberto Baseado em Tópicos Escalonáveis. In: Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence, 228-234.

Cohn, D. & Hofmann, T. (2001). O elo que falta: Um modelo probabilístico de conteúdo de documentos e conectividade de hipertexto. Neural Information Processing Systems 13, 430-436.

Erosheva, E. A. (2002). Modelos de grau de afiliação e de estrutura latente com aplicações aos dados da pesquisa sobre deficiência.
Tese de doutorado inédita, Departamento de Estatística, Universidade Carnegie Mellon.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov cadeia Monte Carlo na prática*. Londres: Chapman & Hall.

Griffiths, T. L., & Steyvers, M. (2002). Uma abordagem probabilística da representação semântica. Em *Proceedings of the 24th Annual Conference of the Cognitive Science Society.*

Griffiths, T. L., & Steyvers, M. (2003). Predição e associação semântica. Em *Sistemas de processamento de informação neural 15.* Cambridge, MA: MIT Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101,* 5228-5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005) Integrando tópicos e sintaxe. In *Advances in Neural Information Processing 17*. Cambridge, MA: MIT Press.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis (Análise Semântica Latente Probabilística). Em *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.*

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis (Aprendizagem sem Supervisão por Análise Semântica Latente Probabilística). *Machine Learning Journal, 42(1)*, 177-196.

Landauer, T. K., & Dumais, S. T. (1997). Uma solução para o problema de Platão: a teoria da Análise Semântica Latente de aquisição, indução e representação do conhecimento. *Revisão Psicológica, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introdução à análise semântica latente. *Processos Discursivos, 25,* 259-284.

Lee, D.D., & Seung, H.S. (2001). Algoritmos para a Factorização de Matrizes Não Negativas. In: *Sistemas de processamento de informação neural 13.* Cambridge, MA: MIT Press.

Lin, J. (1991). Medidas de divergência baseadas na entropia de Shannon. *IEEE Transactions on Information Theory, 37*(14), 145-51.

Manton, K.G., Woodbury, M.A., & Tolley, H.D. (1994). *Aplicações Estatísticas Utilizando Conjuntos Fuzzy*. Wiley, Nova York.

Minka, T. & Lafferty, J. (2002). Expectativa-propagação para o modelo do aspecto generativo. In: *Anais da 18ª Conferência sobre Incerteza em Inteligência Artificial*. Elsevier, Nova York.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. (http://www.usf.edu/FreeAssociation/).

Pritchard, J . K . , Stephens, M . , & Donnelly, P . (2000). Inferência da estrutura da população usando dados do genótipo multilocus. *Genetics*, *155*, 945-955.

Rosen-Zvi, M., Griffiths T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents (O Modelo Autor-Tópico para Autores e Documentos). Na *20ª Conferência sobre Incerteza em Inteligência Artificial.* Banff, Canadá.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic Author-Topic Models for Information Discovery. A *Décima Conferência Internacional ACM SIGKDD sobre Descoberta do Conhecimento e Mineração de Dados*. Seattle, Washington.

Teh, Y. W., Jordan, M. I., Beal, M.J. & Blei, D. M. (2004). Hierarchical Dirichlet Processes. Relatório Técnico 653, UC Berkeley Statistics, 2004.

Ueda, N., & Saito, K. (2003). Modelos de misturas paramétricas para textos com várias etiquetas. In *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press.