

In: 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 de Tópicos Probabilísticos

Mark Steyvers

University of California, Irvine

Universidade

Tom Griffiths

Brown

Enviar 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 render 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 desta 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 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 idéia de que os documentos são misturas de tópicos, onde um tópico é uma distribuição de probabilidade sobre as palavras. Um modelo de tópico é um *modelo generativo* para documentos: ele 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 este 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 conteúdo diferente podem ser gerados escolhendo diferentes distribuições ao longo dos tópicos. Por exemplo, dando igual probabilidade 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 igual probabilidade aos dois últimos tópicos, pode-se construir um documento sobre uma pessoa que experimentou uma perda de memória, o que exigiu uma visita ao médico.

Tópico 247Topic 5Topic 43Topic 56

palavra	prob.	palavra	prob.	palavra	prob.	palavra	prob.
DRUGS	.069	VERMELHO	.202	MENTE	.081	DOCTOR	.074
DRUG	.060	BLUE	.099	THOUGHT	.066	DR.	.063
MEDICINA	.027	VERDE	.096	LEMBRE-SE	.064	PATIENTE	.061
EFEITOS	.026	AMARELO	.073	MEMÓRIA	.037	HOSPITAL	.049
CORPO	.023	BRANCO	.048	PENSAGEM	.030	CARE	.046
MEDICAMENTO	.019	COLOR	.048	PROFESSOR	.028	MEDICAL	.042
S		BRILHO	.030	FELT	.025	NÚMERO	.031
DOR	.016	COLORS	.029	REMETIDO	.022	PATIENTES	.029
PESSOA	.016	ORANGE	.027	THOUGHTS	.020	DOUTORES	.028
MARIJUANA	.014	COROA	.027	FORGOTTEN	.020	SAÚDE	.025
LABEL	.012	PINK	.017	MOMENTO	.020	MEDICINA	.017
ALCOHOL	.012	LOOK	.017	PENSAGEM	.019	NURSING	.017
PERIGOSO	.011	PRETO	.016	OBSERVAÇÃO	.016	DENTAL	.015
ABUSO	.009	PROPRIEDADE	.015	MARAVILHA	.014	NOLSES	.013
EFEITO	.009	CROSS	.011	ESQUEÇA	.012	PHYSICIAN	.012
CONHECIDO	.008	CORADO	.009	LEMBRE-SE	.012	HOSPITAIS	.011
PILLS	.008						

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 quanto os 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 as idéias-chave por trás dos modelos de tópicos com mais detalhes, e delineamos como é possível identificar os tópicos que aparecem em um conjunto de documentos. Em seguida, discutimos 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 de 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 a 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 utilizado para a amostragem da palavra. Da forma como o modelo é definido, não há noção de exclusividade mútua que restrinja as palavras a serem 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 quanto o tópico rio podem dar alta probabilidade à palavra BANCO, o que é sensato dada a natureza polissêmica da palavra.

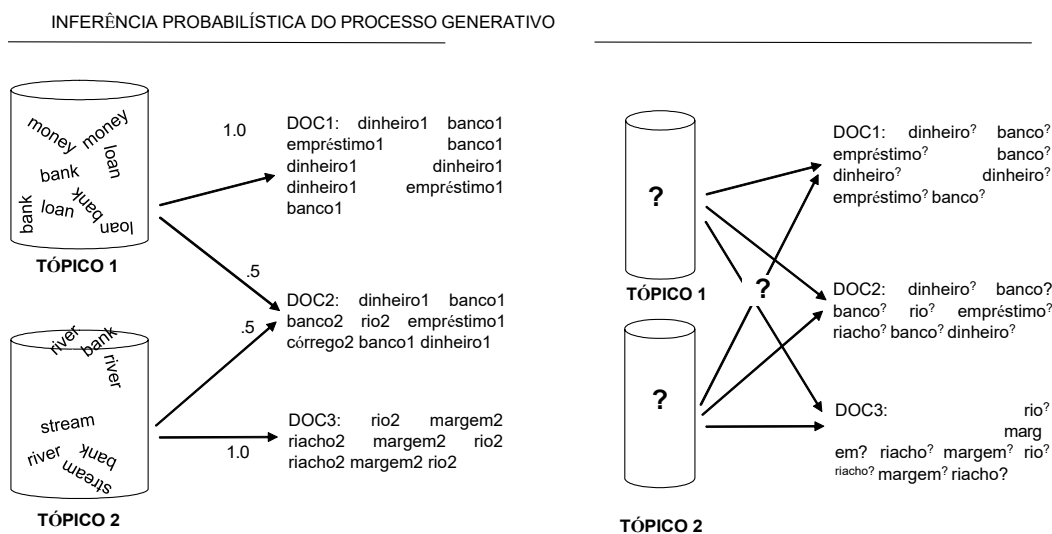


Figura 2. Ilustração do processo generativo e o problema da inferência estatística subjacente aos modelos temáticos

O processo generativo aqui descrito não faz nenhuma suposição sobre a ordem das palavras como elas aparecem nos documentos. A única informação relevante ao 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. É claro que as informações por ordem de palavras podem conter indicações importantes para o conteúdo de um documento e estas informações não são utilizadas 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 fatores sintáticos e semânticos que orientam a escolha das palavras (veja também Dennis, este livro para uma abordagem diferente deste 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 as palavras associadas a cada tópico, a distribuição sobre tópicos para cada documento e, freqüentemente, o tópico responsável pela geração de cada palavra.

3. Modelos de Tópicos Probabilísticos

Uma variedade de modelos tópicos probabilísticos tem sido usada 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 idéia fundamental - que um documento é uma mistura de tópicos - mas fazem suposições estatísticas ligeiramente diferentes. Para introduzir a notação, escreveremos $P(z)$ para a distribuição sobre tópicos z em um determinado documento e $P(w|z)$ para a distribuição de probabilidade sobre palavras w dado tópico z . Várias distribuições de palavras tópico $P(w|z)$ 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 a partir da distribuição de tópicos, depois escolhendo uma palavra a partir da distribuição de palavras do tópico. Escrevemos $P(z_i = j)$ como a probabilidade de que o j th tópico foi amostrado para o i th word token e $P(w_i | z_i = j)$ como a probabilidade de palavra w_i sob o tópico j . O modelo especifica a seguinte distribuição sobre as palavras dentro de um documento:

$$P(w_i) = \sum_{z=1}^T P(z) P(w_i | z) \quad (1)$$

onde T é o número de tópicos. Para simplificar a notação, deixe $\theta_j = P(z=j)$ referir-se à distribuição multinomial sobre palavras para o tópico j e $\phi^{(d)} = P(z)$ 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 de palavras (ou seja, $N = \sum_d N_d$). Os parâmetros e 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 em 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 generalidade do modelo para novos documentos. Blei et al. (2003) estenderam este modelo introduzindo um Dirichlet prévio em , chamando o modelo generativo resultante de Latent Dirichlet Allocation (LDA). Como um anterior conjugado para a multinacional, 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:

$$Dir(p; \alpha) = \frac{\prod_{j=1}^T \alpha_j^{p_j}}{\sum_{p \in \Delta} \prod_{j=1}^T \alpha_j^{p_j}} \quad (2)$$

Os parâmetros desta distribuição são especificados por $\alpha_1 \dots \alpha_T$. 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 quaisquer palavras reais daquele documento. É conveniente usar uma distribuição simétrica de Dirichlet com um único hiperparâmetro α de modo que $\alpha_1 = \alpha_2 = \dots = \alpha_T = \alpha$. Ao colocar previamente um Dirichlet sobre a distribuição temática, o resultado é uma distribuição temática suavizada, com a quantidade de suavização determinada pelo parâmetro α . A Figura 3 ilustra a distribuição do 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, 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 (frequentemente utilizado na prática), há um viés para a esparsidade, e a pressão é para escolher distribuições temáticas favorecendo apenas alguns tópicos.

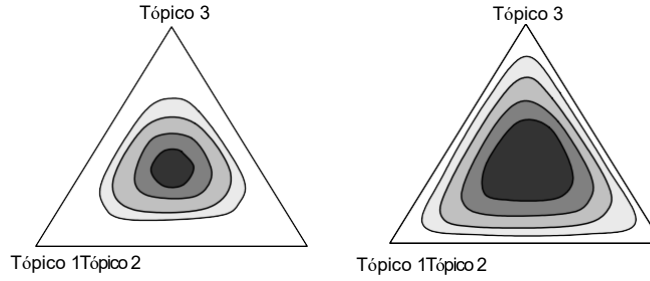


Figura 3. Ilustrando a distribuição simétrica de Dirichlet para três tópicos em um simplex bidimensional. As cores mais escuras indicam maior probabilidade. Esquerda: $\alpha = 4$. Direita: $\alpha = 2$.

Griffiths e Steyvers (2002; 2003; 2004) exploraram uma variante deste modelo, discutida por Blei et al. (2003), colocando um Dirichlet simétrico (anterior também em cima). O hiperparâmetro pode ser interpretado como a observação prévia conta com o número de vezes que as palavras são amostradas de um tópico antes que qualquer palavra do corpus seja 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.

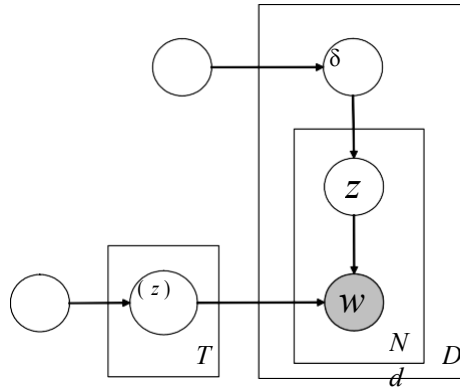


Figura 4. O modelo gráfico para o modelo tópico utilizando notação de placa.

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 (isto é, não observadas) respectivamente. As variáveis δ e z , assim como z (a atribuição de símbolos de palavras a 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 se referindo 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 $w^{(d)}$ ilustra a amostragem de uma distribuição sobre tópicos para cada documento d para um total de documentos D . A placa ao redor de $z^{(z)}$ 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 tipos de palavras W 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 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-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 (isto é, $T \ll W$), os tópicos abrangem um subgrupo de baixa dimensão e a projeção de cada documento no subgrupo de baixa dimensão pode ser pensada como redução de dimensionalidade. Esta formulação do modelo é semelhante à Análise Semântica Latente. Buntine (2002) apontou correspondências formais entre modelos temáticos e análise de componentes principais, um procedimento intimamente relacionado à LSA.

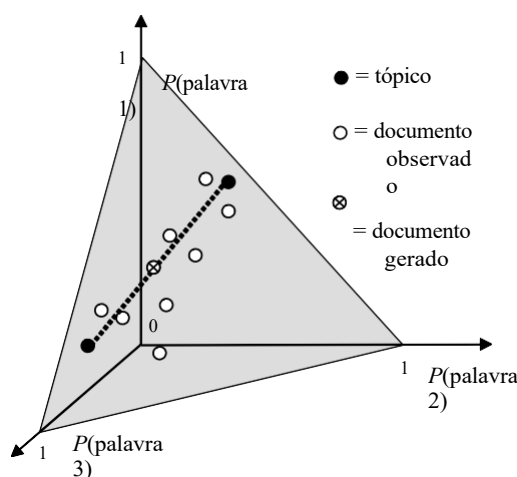


Figura 5. Uma interpretação geométrica do modelo temático.

Interpretação da Factorização Matriz. Na LSA, uma matriz de co-ocorrência de documentos pode ser decomposta por decomposição de valores singulares em três matrizes (ver outro capítulo deste livro de Martin & Berry): uma matriz de vetores de palavras, uma matriz diagonal com valores singulares e uma matriz com vetores 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 de documentos de trabalho é dividida em duas partes: uma matriz de tópicos e um documento

matriz. Note que a matriz diagonal D no LSA pode ser absorvida na matriz U ou V , tornando a semelhança entre as duas representações ainda mais clara.

Esta fatorização destaca uma semelhança conceitual entre o LSA e os modelos temáticos, ambos encontrando uma representação de baixa dimensão para o conteúdo de um conjunto de documentos. No entanto, ela também mostra várias diferenças importantes entre as duas abordagens. Nos modelos tópicos, a palavra e os vetores dos documentos das duas matrizes decompostas são distribuições de probabilidade com a restrição de que os valores das características são não-negativos e totalizam até um. No modelo LDA, restrições adicionais a priori são colocadas sobre as distribuições de palavras e tópicos. Não existe tal restrição sobre vetores LSA, embora existam outras técnicas de fatorização de matrizes que requerem valores de características não negativas (Lee & Seung, 2001). Em segundo lugar, a decomposição do 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 temáticas são independentes, mas não ortogonais; a inferência do modelo precisa ser feita separadamente para cada dimensionalidade.

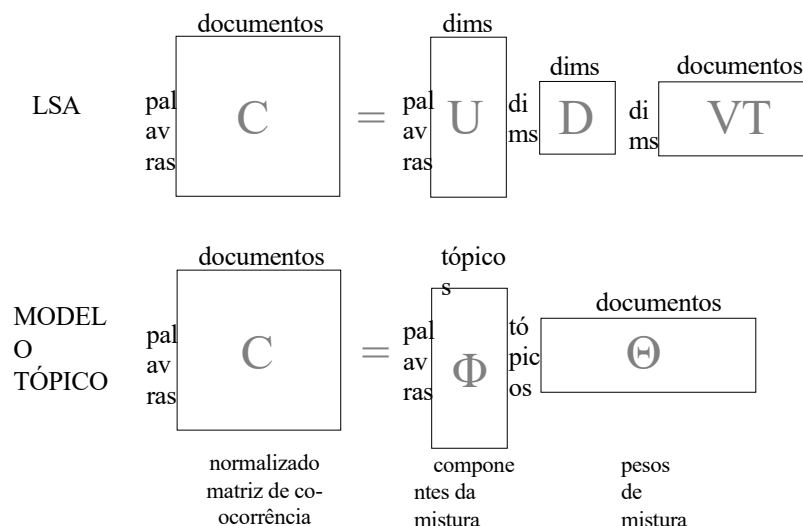


Figura 6. A factorização matricial do modelo LSA em comparação com a factorização matricial do modelo tópico

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 somente 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 informações de autoria com conteúdo. Em vez 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 com 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 além do texto. Os modelos de grade-of-membership (GoM) desenvolvidos por estatísticos nos anos 70 são de forma semelhante (Manton, Woodbury, & Tolley, 1994), e Erosheva (2002) considera um modelo 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 temáticas e as distribuições temáticas para cada documento. Hofmann (1999) utilizou o algoritmo de expectativa-maximização (EM) para obter estimativas diretas de θ e ϕ . Esta abordagem sofre de problemas envolvendo máximos locais da função de probabilidade, o que motivou uma busca por melhores algoritmos de estimativa (Blei et al., 2003; Buntine, 2002; Minka, 2002). Em vez de estimar diretamente as distribuições de palavras-chave e de 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 w , enquanto marginaliza θ e ϕ . Cada z_i dá um valor inteiro $[1 \dots T]$ para o tópico ao qual a palavra token i é atribuída. Como muitas coleções de texto contêm milhões de palavras token, a estimativa da palavra token posterior sobre z exige procedimentos eficientes de estimativa. Descreveremos um algoritmo que usa a amostragem 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 para extrair 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 projetadas para amostrar valores de distribuições complexas (muitas vezes de alta dimensão) (Gilks, Richardson, & Spiegelhalter, 1996). A amostragem 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 ao valor de todos os outros. A amostragem é feita sequencialmente e prossegue até que os valores amostrados se aproximem da distribuição alvo. Embora o procedimento Gibbs que descreveremos não forneça estimativas diretas de θ e ϕ , mostraremos como e pode ser

aproximado usando estimativas posteriores de 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 à 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 $P(z_i = j | z_{-i}, w_i, d_i)$, onde $z_i = j$ representa a atribuição de tópico do token i ao tópico j , 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 w_{-i} e d_{-i} , e hiperparâmetros α e τ . Griffiths e Steyvers (2004) mostraram como isto pode ser calculado por:

$$P(z_i = j | z_{-i}, w_i, d_i) = \frac{C_{wj}^{WT} + \alpha}{\sum_{k=1}^T C_{wk}^{WT} + \alpha} \frac{C_{dj}^{DT} + \tau}{\sum_{k=1}^T C_{dk}^{DT} + \tau} \quad (3)$$

onde C^{WT} e C^{DT} são matrizes de contagens com dimensões $L \times T$ e $D \times T$ respectivamente; C_{wj}^{WT} contém o número de vezes a palavra w é atribuída ao tópico j , não incluindo a instância atual i e C_{dj}^{DT} 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-se a quantidade na Equação 3 para o tópico t pela soma em todos os tópicos T .

Os fatores que afetam 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 distribuição do tópico atual 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 daquele documento seja atribuída ao tópico j . Portanto, as palavras são atribuídas a tópicos dependendo da probabilidade de que a palavra seja para um tópico, bem como do quão dominante um tópico é em um documento.

O algoritmo de amostragem de Gibbs começa atribuindo cada palavra simbólica a um tópico aleatório em $[1 \dots T]$. Para cada palavra token, as matrizes de contagem C^{WT} e C^{DT} são primeiro decretadas por um 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 C^{WT} e C^{DT} são incrementadas com a nova atribuição de tópico. Cada amostra de Gibbs consiste no conjunto de atribuições de tópicos a todos os N símbolos de palavras no corpus, alcançado por uma única passagem por todos os documentos. Durante o estágio 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 se aproximar da distribuição alvo (ou seja, a distribuição posterior sobre as atribuições do tópico). Neste ponto, para obter um conjunto representativo de amostras desta distribuição, várias amostras de Gibbs são salvas em intervalos regularmente espaçados, para evitar correlações entre as amostras (ver Gilks et al. 1996).

Estimativa e . O algoritmo de amostragem fornece estimativas diretas de z para cada palavra. Entretanto, muitas aplicações do modelo requerem estimativas θ e ϕ das distribuições de palavras-tópicos e distribuições de documentos-tópicos, respectivamente. Estas podem ser obtidas a partir das matrizes de contagem da seguinte forma:

$$\theta_j = \frac{C_{.j}^{WT}}{\sum_{k=1}^T C_{.k}^{WT} + \alpha} \quad \phi_j = \frac{C_{dj}^{DT}}{\sum_{k=1}^T C_{dk}^{DT} + \tau} \quad (4)$$

Estes valores correspondem às distribuições preditivas de amostragem de um novo símbolo da palavra i do tópico j , e 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 z .

Um exemplo. O algoritmo de amostragem 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. Ilustramos isto expandindo sobre o exemplo que foi dado na Figura 2. Suponha que o tópico 1 dá igual probabilidade às palavras

DINHEIRO, EMPRÉSTIMO e BANCO _{DINHEIRO} i.e., $\theta^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ EMPRÉSTIMO

(1)
BANCO

1/ 3,
enqu
anto
que o
tópic
o 2
dá
igual
prob
abili
dade
às
palav
ras
RIVE
R,

STREAM, e BANK, ou seja, (2)
RIVER

(2)
STREAM

(2)
BANC
O

1/ 3 . A figura 7, painel superior, mostra como 16 documentos
podem ser

gerados 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 apresenta as seguintes estimativas para o

distribuições sobre palavras para tópico 1 e 2: $\theta^{(1)}_{DINHEIRO} = .32$, $\theta^{(1)}_{EMPRÉSTIMO} = .29$, $\theta^{(1)}_{BANCO} = .39$ e $\theta^{(2)}_{RIVER} = .25$, $\theta^{(2)}_{STREAM} = .4$, $\theta^{(2)}_{BANCO} = .35$. Dado o tamanho do conjunto de dados, estas estimativas são razoáveis.

reconstruções dos parâmetros usados para gerar os dados.

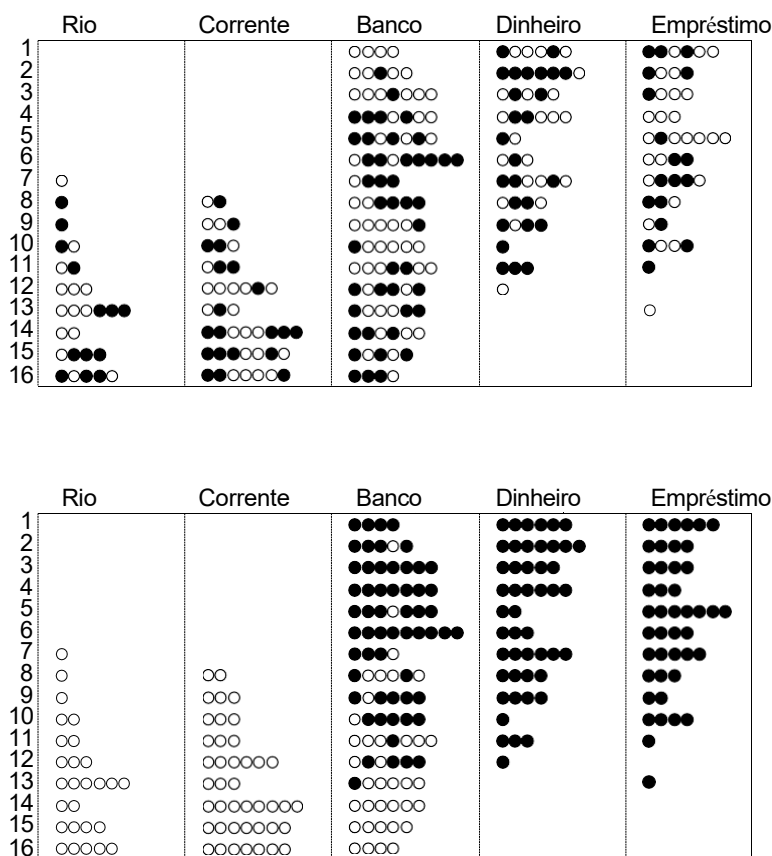


Figura 7. Um exemplo do procedimento de amostragem de Gibbs.

Permutabilidade dos tópicos. Não há nenhuma ordem a priori sobre os tópicos que tornarão 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 atribuições aleatórias diferentes). Portanto, as diferentes amostras *não podem* ser calculadas como média no nível dos tópicos. Entretanto, quando são usados tópicos para calcular uma estatística que é invariante para a 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 dos modelos melhore os resultados porque permite a amostragem a partir de múltiplos modos locais do posterior.

Estabilidade dos Tópicos. Em algumas aplicações, é desejável concentrar-se 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 solução particular. 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 temáticas. A dissimilaridade entre os tópicos j_1 e j_2 foi medida pela distância simetrizada Kullback Liebler (KL) entre as distribuições dos tópicos:

$$KL(j_1, j_2) = \frac{1}{2} \sum_{k=1}^K \left(\frac{w_{k1}}{w_{k1} + w_{k2}} \log \frac{w_{k1}}{w_{k1} + w_{k2}} + \frac{w_{k2}}{w_{k1} + w_{k2}} \log \frac{w_{k2}}{w_{k1} + w_{k2}} \right) \quad (5)$$

em que 'e' e " correspondem às distribuições de palavras-temas estimadas a partir de duas séries diferentes. Os tópicos da segunda série foram reordenados para corresponder o melhor possível (usando um algoritmo ganancioso) com os tópicos da primeira série. A correspondência foi medida pela soma (inversa) das distâncias KL na diagonal. A matriz de similaridade na Figura 8 sugere que uma grande porcentagem 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 a 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 em todas as séries.

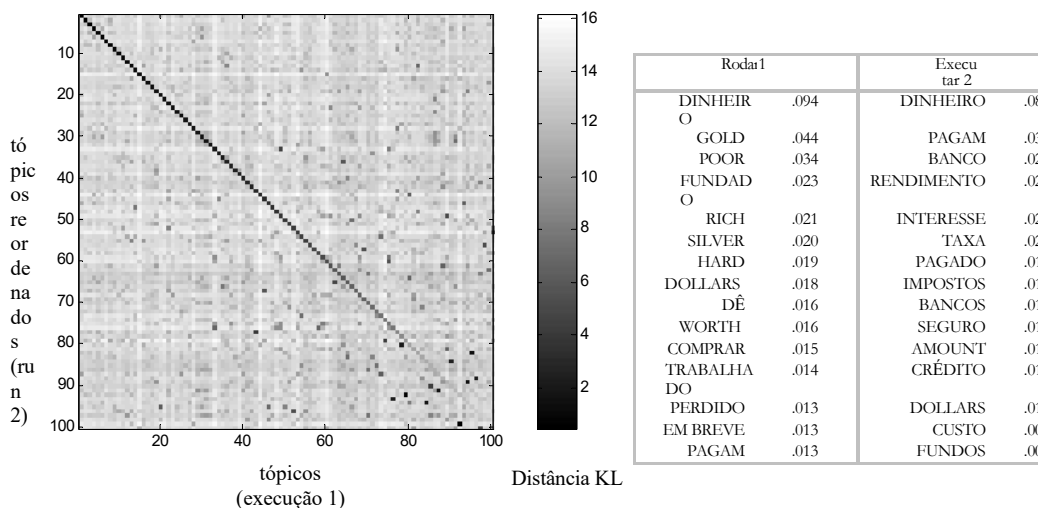


Figura 8. Estabilidade de tópicos entre diferentes séries.

Determinando o número de tópicos. A escolha do número de tópicos pode afetar 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 restante do conjunto de documentos. Na linguística 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 ambigüidade 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 aos diferentes sentidos da palavra (*tocar música, teatro, jogos*).

Tópico 77		Tópico 82		Tópico 166	
palavra	prob.	palavra	prob.	palavra	prob.
MÚSICA	.090	LITERATURA	.031	JOGUE	.136
DANÇA	.034	POEM	.028	BOLA	.129
SONG	.033	POETRY	.027	JOGO	.065
JOGUE	.030	POET	.020	JOGANDO	.042
CANTAR	.026	JOGADORES	.019	HIT	.032
CANTANDO	.026	POEMS	.019	JOGADO	.031
BANDA	.026	JOGUE	.015	BASEBALL	.027
JOGADO	.023	LITERÁRIO	.013	GAMES	.025
SANG	.022	ESCRITORES	.013	BAT	.019
SONGS	.021	DRAMA	.012	EXECUÇÃO	.019
DINHEIRO	.020	WROTE	.012	ALARGAR	.016
PIANO	.017	POETS	.011	BOLAS	.015
JOGANDO	.016	ESCRITÓRIO	.011	TENNIS	.011
RHYTHM	.015	SHAKESPEARE	.010	HOME	.010
ALBERT	.013	ESCREVIDO	.009	CATCH	.010
MUSICAL	.013	ESTÁGIO	.009	FIELD	.010

Figura 9. Três tópicos relacionados com a palavra JOGUE.

Documento #29795

Bix beiderbecke, na idade⁰⁶⁰ quinze²⁰⁷, sentou¹⁷⁴ na encosta⁰⁷¹ de um bluff⁰⁵⁵ com vista para⁰²⁷ o mississippi¹³⁷ rio¹³⁷. Ele estava escutando⁰⁷⁷ a música⁰⁷⁷ vinda⁰⁰⁹ de um barco de passagem⁰⁴³ do rio. A música⁰⁷⁷ já havia capturado⁰⁰⁶ seu coração¹⁵⁷ assim como seu ouvido¹¹⁹. Era jazz⁰⁷⁷. Bix beiderbecke já tinha tido aulas de música⁰⁷⁷⁰⁷⁷. Ele mostrou⁰⁰² promessa¹³⁴ no piano⁰⁷⁷, e seus pais⁰³⁵ esperavam que²⁶⁸ ele pudesse considerar¹¹⁸ tornar-se um concerto⁰⁷⁷ pianista⁰⁷⁷. Mas Bix estava interessado²⁶⁸ em outro tipo de música⁰⁷⁷. Ele queria²⁶⁸ play⁰⁷⁷ a corneta. E ele queria²⁶⁸ para play⁰⁷⁷ jazz⁰⁷⁷ ... para

Documento #1883

Há uma simples razão^{050 106} porque há tão poucos períodos⁰⁷⁸ de teatro realmente grande⁰⁸² em todo o nosso mundo ocidental¹⁰⁴⁶. Demasiadas coisas³⁰⁰ têm que vir ao mesmo tempo. Os dramaturgos devem ter os atores certos⁰⁸², os atores⁰⁸² devem ter as casas de teatro certas, as casas de teatro devem ter o público certo⁰⁸². Devemos lembrar²⁸⁸ que as peças⁰⁸² existem¹⁴³ para serem representadas⁰⁷⁷, e não apenas⁰⁵⁰ para ser lido²⁵⁴. (mesmo quando você lê²⁵⁴ a junto.) como . logo⁰²⁸ como uma play⁰⁸² para você mesmo, tente²⁸⁸ para apresentar⁰⁶², para colocar¹⁷⁴ em um palco⁰⁷⁸, como você vai tem play⁰⁸² para ser realizado⁰⁸², então alguns espécie¹²⁶ de teatral⁰⁸² ...

Documento #21359

Jim²⁹⁶ tem um jogo¹⁶⁶ livro²⁵⁴. Jim²⁹⁶ lê²⁵⁴ o livro²⁵⁴. Jim²⁹⁶ vê⁰⁸¹ um jogo¹⁶⁶ para um. Jim²⁹⁶ joga¹⁶⁶ o jogo¹⁶⁶. Jim²⁹⁶ gosta de⁰⁸¹ o jogo¹⁶⁶ por um. O jogo¹⁶⁶ livro²⁵⁴ ajuda⁰⁸¹ jim²⁹⁶. Don¹⁸⁰ vem⁰⁴⁰ para dentro de casa⁰³⁸. Don¹⁸⁰ e jim²⁹⁶ lê²⁵⁴ o jogo¹⁶⁶ livro²⁵⁴. Os meninos⁰²⁰ vêem um jogo¹⁶⁶ para dois. Os dois play¹⁶⁶ o jogo¹⁶⁶. Os meninos⁰²⁰ play¹⁶⁶ o jogo¹⁶⁶ para dois. Os meninos⁰²⁰ gostam do jogo¹⁶⁶. Meg²⁸² vem⁰⁴⁰ para dentro de casa²⁸². Meg²⁸² e don¹⁸⁰ e jim²⁹⁶ lêem²⁵⁴ o livro²⁵⁴. Eles vêem um jogo¹⁶⁶ por três. Meg²⁸² e don¹⁸⁰ e jim²⁹⁶ play¹⁶⁶ o jogo¹⁶⁶. Eles jogam¹⁶⁶..

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. Esta 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 simbólica 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 cinzas são palavras de parada ou palavras de frequência muito baixa 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) cria 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 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 os documentos d_1 e d_2 pode ser medida pela semelhança entre suas distribuições temáticas correspondentes $\theta^{(d_1)} e^{(d_2)}$. 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 Kullback Leibler (KL),

$$D_{KL}(p, \theta) = \sum_{j=1}^T p_j \log 2 \frac{p_j}{q_j} \quad (6)$$

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:

$$J_{KL}(p, \theta) = \frac{1}{2} D_{KL}(p, \theta) + \frac{1}{2} D_{KL}(q, \pi) \quad (7)$$

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

$$J_{JS}(p, \theta) = \frac{1}{2} D_{KL}(p, \theta) + \frac{1}{2} D_{KL}(q, \pi) \quad (8)$$

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 $(p+q)/2$. Tanto a função de divergência simetrizada KL como a função de divergência JS parecem funcionar bem na prática. Além disso, também é possível considerar as distribuições temáticas como vetores e aplicar funções de motivação geométrica como distância Euclidiana, produto ponto ou coseno.

Para aplicações de recuperação de informações, 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 similares 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 informações 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 $P(q | d_i)$ onde q é o conjunto de palavras contidas na consulta. Usando as suposições do modelo do tópico, isto pode ser calculado por:

$$P(q | d_i) = \prod_{k=1}^K \frac{P(w_k | d_i)}{\sum_{j=1}^T P(w_k | d_j)} \quad (9)$$

Observe que esta abordagem também enfatiza a semelhança através de tópicos, com documentos relevantes tendo distribuições temáticas 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

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

Semelhança entre duas palavras. A semelhança entre duas palavras w_1 e w_2 pode ser medida pela medida em que elas compartilham 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

$$^{(1)} P(z | w_1) \text{ e } ^{(2)} P(z | w_2). \text{ Ou a divergência simetrizada da KL ou do JS seria}$$

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 , ou seja, $P(w_2 | w_1)$ -- quais são as prováveis palavras que são geradas como 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 assuntos se associam com a taco nos diferentes sentidos da palavra (por exemplo, PLAYBALL e PLAYACTOR). No modelo de 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 $P(w_2 | z_j)$. Se for assumido que cada assunto ativa apenas um único tópico amostrado da distribuição $P(z_j | w_1)$, as distribuições condicionais preditivas podem ser calculadas por:

$$P(w_2 | w_1) = \sum_{\phi=1}^T P(w_2 | z_{\phi}) P(z_{\phi} | w_1) \quad (10)$$

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

O painel direito da Figura 9 mostra as previsões do modelo de tópico 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	
FUN	.141	BOLA	.036
BOLA	.134	JOGO	.024
JOGO	.074	CRANÇAS	.016
TRABALHO	.067	TEAM	.011
TERRITÓRIO	.060	DESEJA	.010
MATE	.027	MÚSICA	.010
CRANÇA	.020	SHOW	.009
DESEFRUTAR	.020	HIT	.009
GANHO	.020	CRANÇA	.008
ATOR	.013	BASEBALL	.008
LUTA	.013	GAMES	.007
CAVALO	.013	FUN	.007
KID	.013	ESTÁGIO	.007
MÚSICA	.013	FIELD	.006

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

7. Conclusão

Modelos geradores de texto, como o modelo temático, têm o potencial de fazer contribuições importantes 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 suposições explícitas sobre o processo causal responsável pela geração de um documento e permitem o uso 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. Modelos temáticos ilustram como a utilização de uma representação diferente pode proporcionar novas perspectivas na modelagem estatística da linguagem, incorporando muitas das principais suposições por trás da LSA, mas tornando possível identificar um conjunto de tópicos probabilísticos interpretáveis, em vez de um espaço semântico. Os modelos temáticos também foram ampliados para capturar algumas propriedades interessantes da linguagem, tais como as relações semânticas hierárquicas entre palavras (Blei et al., 2004), e a interação entre sintaxe e semântica (Griffiths et al., 2004). A grande maioria dos modelos generativos ainda não foi definida, e a investigação desses modelos proporciona a oportunidade de expandir tanto os benefícios práticos quanto a compreensão teórica do aprendizado estatístico 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 de tópicos 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 mecanismo de busca de código aberto com base em tópicos escaláveis. 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. *Sistemas de Processamento de Informações Neurais* 13, 430-436.

Erosheva, E. A. (2002). Modelos de grau de afiliação e estrutura latente com aplicações a dados de pesquisa de 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 informações neurais 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). 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 Matriz Não-Negativa. 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 de genótipo multilocus. *Genetics*, 155, 945-955.
- Rosen-Zvi, M., Griffiths T., Steyvers, M., & Smyth, P. (2004). 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). Processos Hierárquicos de Dirichlet. Relatório Técnico 653, UC Berkeley Statistics, 2004.
- Ueda, N., & Saito, K. (2003). Modelos de misturas paramétricas para textos com várias legendas. In *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press.