Modelos de Alocação Latente de Dirichlet para Verificação dos Tópicos de Apresentação do SIGKDD 2016

Eduardo Elias Ribeiro Junior - PGEST UFMG
<edujrrib@gmail.com> | <jreduardo.github.io>

Introdução

Comumente eventos da comunidade científica reúnem pesquisadores para exposição e discussão de seus recentes trabalhos. Esses trabalhos, em geral, são classificados em tópicos facilitando **i)** os participantes a localizarem seus interesses e **ii)** consultas ao acervo após finalização. Todavia, há pouco rigor na atribuição dos tópicos e muitas vezes, pela má atribuição, esses acabam sendo dispensáveis. Portanto o presente trabalho visa:

- Apresentar uma abordagem probabilística para atribuição de tópicos;
- Verificar a definição e atribuição de tópicos realizada no SIGKDD 2016.

Toda a análise do trabalho foi realizada com o software R, cujo códigos foram disponibilizados no sítio eletrônico do autor.

Conjunto de dados

Para exemplificação do problema exposto e aplicação da proposta, escolheu-se a *22nd SIGKDD Conference*, maior evento de Knowledge Discovery and Data Mining. Os dados desse evento estão disponíveis em páginas *web*, conforme ilustrado na Figura 1 sendo extraídos as porções detacados em vermelho e azul.

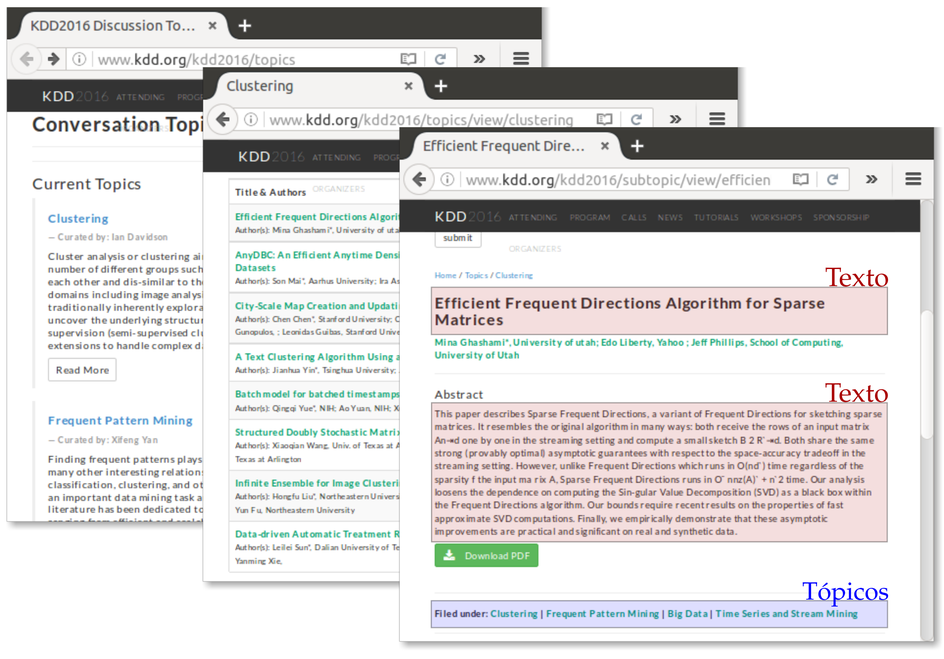

Figura 1: Sítio do SIGKDD2016 de onde foram extraídos os títulos, resumos e tópicos.

Ao todo foram 203 artigos contemplados com 3171 palavras distintas, após mineração (higienização e radicalização). Na Figura 2 têm-se uma descrição dos dados. As palavras mais frequentes correspondem àquelas comuns no ambiente de KDD e a maioria dos artigos está relacionada a 1 ou 2 tópicos (175 artigos). Para a frequência de artigos em cada tópico, nota-se que há tópicos dominantes como big-data e mining-rich-data-types.

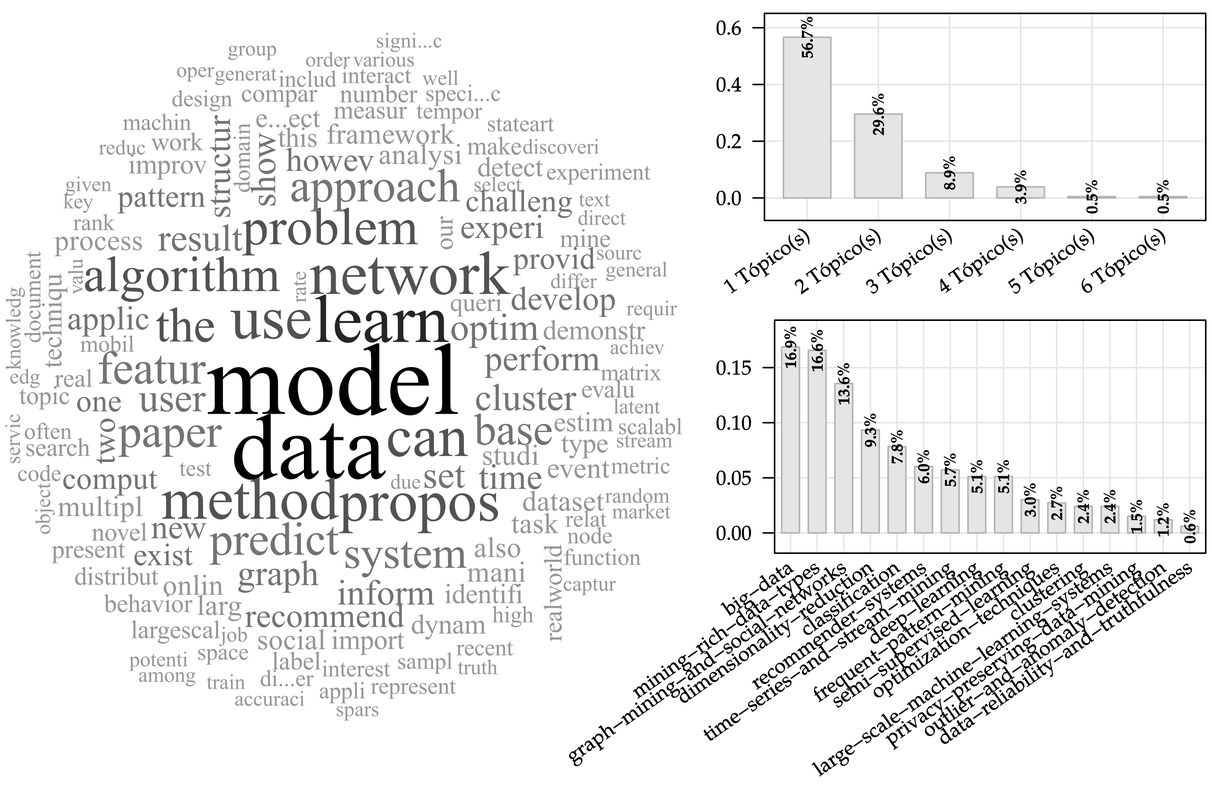

Figura 2: 5% das palavras mais frequentes (esquerda), frênquencias dos papers: sob a quantidade de tópicos a que pertence (superior à direita) e em cada tópico do evento (inferior direita).

Métodos

Nesse trabalho os modelos de alocação latente de Dirichlet (LDA), em que considera-se que documentos exibam múltiplos tópicos (BLEI, 2012), são empregados. A Figura 3 representa o modelo LDA, sendo **K**, **D** e **N** são os conjuntos dos tópicos, documentos e palavras respectivamente.

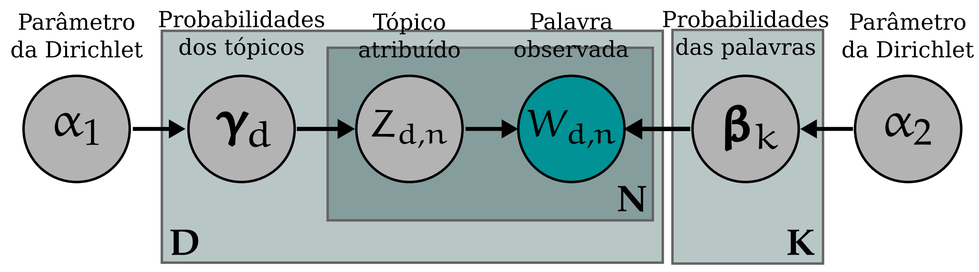

Figura 3: Representação gráfica do modelo LDA.

A partir da estrutura descrita na Figura 3 têm-se a distribuição posteriori descrita na Equação 1. A inferência é realizada via amostrador de Gibbs, utilizando as condicionais completas (GRUN; HORNIK, 2011).

$$f(\beta, \gamma, z \mid w) \propto \prod_{k=1}^{K} [\beta_k] \prod_{d=1}^{D} [\gamma_d] \prod_{n=1}^{N} \left([z_{d,n} \mid \gamma_d] [w_{d,n} \mid \beta_k, z_{d,n}] \right) \quad (1)$$

Resultados

Os resultados exibidos são referentes à média de 1000 iterações do amostrador de Gibbs (descartando os 1000 primeiros estados e armazenando de 10 em 10). São ao todo **53984** parâmetros no modelo (γ_{203x16} e $\beta_{16x3171}$). A matriz de proporções γ é apresentada na Figura 4.

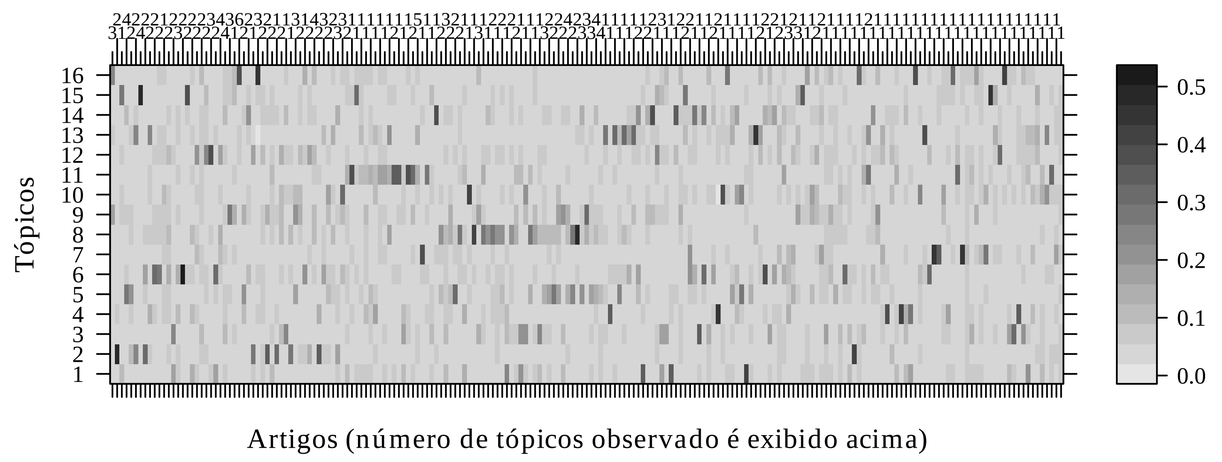

Figura 4: Proporções dos tópicos em cada artigo do SIGKDD 2016.

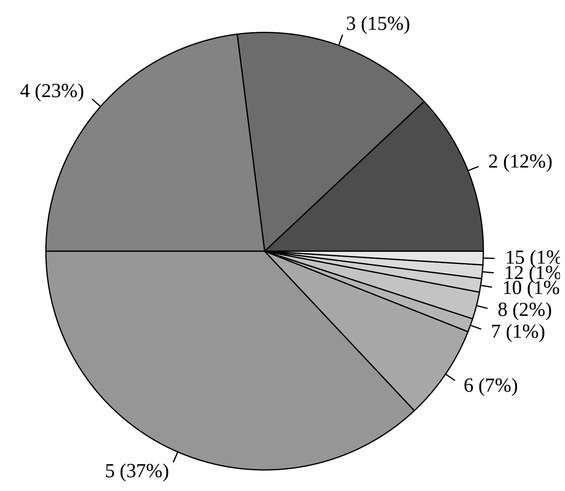

Figura 5: Frequências do número de tópicos remanescentes após reamostragem.

Como a disposição em 16 tópicos não parece adequada, agrupou-se os similares, considerando as distâncias euclidianas da matriz β, que caracteriza os caracteriza. O número de tópicos remanescentes se deu pela maior distância de ligação (via Ward) em um agrupamento hierárquico. O procedimento foi repetido 100 vezes. Os resultados são apresentados na Figura 5. Uma nova amostragem de 1000 estados da cadeia foi tomada considerando 5 tópicos (**16870** parâmetros). Desse modelo as cinco palavras mais prováveis em cada tópico são exibidas na Tabela 1.

Tabela 1: Palavras com maior probabilidade de ocorrência em cada tópico.

Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5
predict (0.026)	model (0.041)	network (0.036)	data (0.036)	learn (0.039)
system (0.020)	user (0.021)	can (0.021)	algorithm (0.022)	use (0.024)
approach (0.017)	experi (0.014)	graph (0.018)	propos (0.020)	data (0.013)
develop (0.015)	onlin (0.013)	structur (0.016)	method (0.019)	queri (0.011)
event (0.013)	social (0.012)	pattern (0.015)	model (0.019)	search (0.010)

Considerações Finais

- Os tópicos do SIGKDD 2016 não agrupam adequadamente os papers;
- A abordagem via LDA apresentou bom resultados, assim como metodologia apresentada para a escolha do número de tópicos.

Referências

BLEI, D. M. Probabilistic topic models. **Commun. ACM**, v. 55, n. 4, p. 77–84, 2012.

GRUN, B.; HORNIK, K. topicmodels: An r package for fitting topic models. **Journal of Statistical Software**, v. 40, n. 1, p. 1–30, 2011.