Caracterização de documentos por meio de modelagem de tópicos

L. Rossi, R. J. P. Damaceno, J. P. Mena-Chalco

Rede Nacional de Ciência para a Educação (CpE)
http://plataforma-cpe.org/

1. MÉTODO

Este material tem por objetivo descrever o processo utilizado para identificar um conjunto de tópicos, com base nos títulos das publicações de autores destacados pela plataforma CpE. A identificação é feita por meio da modelagem de tópicos denominada Latent Dirichlet Allocation (LDA) [Blei et al. 2003]. Nesse contexto, o conjunto de títulos das publicações (documentos) é considerado para a identificação dos tópicos.

A técnica denominada Topic Modeling é um tipo de modelagem estatística considerada para a identificação de tópicos abstratos em um conjunto de documentos, cuja aplicação, para o agrupamento e para a caracterização de documentos, tem crescido ao longo dos últimos anos, apresentando-se como alternativa a utilização de outras tarefas de aprendizagem de máquina, cujos objetivos são similares.

O método LDA é um tipo de modelagem que classifica os textos de um conjunto de documentos em tópicos específicos, sendo uma especialização de Topic Modeling. O método admite, como requisito, um número arbitrário de k tópicos, para os quais cada um dos documentos será associado. Inicialmente, cada palavra é associada, aleatoriamente, a um dos tópicos. Os documentos são caracterizados pela proporção de palavras em cada tópico e, similarmente, os tópicos são caracterizados pela proporção de palavras daquele tópico específico.

Os documentos são compostos por palavras e um tópico é um conjunto composto por todas as palavras pertinentes aos documentos [Rossi et al. 2020]. A diferença entre os tópicos está na frequência das palavras que o compõem. Veja na Figura 1 uma representação dos tópicos e da associação aos documentos.

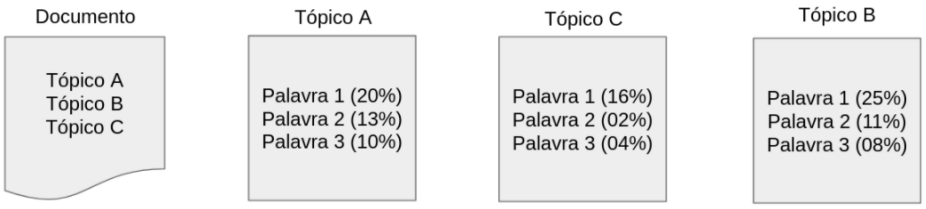

Fig. 1. Exemplo de documento e seus respectivos tópicos associados. Para cada tópico há um score que indica a probabilidade de associação. Os tópicos são arranjos sobre um conjunto de palavras, considerando a frequência observada.

O primeiro passo do processo de modelagem consiste em atribuir, aleatoriamente, um tópico a cada palavra que compõe o documento. Em seguida são calculadas as proporções de palavras em cada tópico por documentos. Veja que, na Figura 2 no documento (a), a proporção de palavras associadas aos tópicos vermelho, azul e verde são, respectivamente 0,333 (4/12), 0,167 (2/12) e 0,500 (6/12).

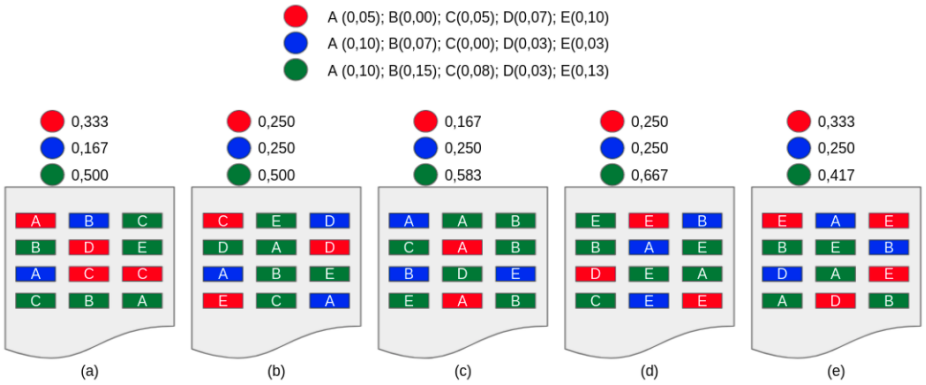

Fig. 2. Exemplo do cálculo das proporções de tópicos nos documentos e de palavras em cada tópico.

Ainda na Figura 2, os tópicos são representados como arranjos sobre o conjunto de palavras. Assim, cada palavra é associada à respectiva proporção de ocorrência, considerando o conjunto de documentos. Por exemplo, a palavra "A" ocorre três vezes no tópico vermelho, assim a proporção associada a essa palavra é 0,05 (3/60). Para as demais palavras: "B", "C", "D" e "E", que compõem o tópico vermelho, temos as seguintes proporções: 0,00 (0/60), 0,05 (3/60), 0,07 (4/60) e 0,10 (6/60), respectivamente. As proporções calculadas são utilizadas para a ordenação das palavras em cada tópico.

Em um processo iterativo, cada palavra em cada documento tem seu tópico atualizado por meio da multiplicação das proporções do tópico no documento e da palavra no tópico. Por exemplo, a palavra A, que é associada ao tópico vermelho no documento (a), terá seu tópico atualizado por aquele que apresenta o maior produto entre as proporções de A nos tópicos e as proporções dos tópicos no documento:

- "A" vermelho: 0,05 x 0,333 = 0,017;
- "A" azul: 0,10 x 0,167 = 0,017;
- "A" verde: 0,10 x 0,500 = 0,05.

Assim, o novo tópico da palavra "A" será verde, que é o maior dos três produtos.

As proporções são atualizadas considerando o novo tópico da palavra "A", conforme apresentado na Figura 3. Esse processo é repetido até que não haja alteração na composição dos tópicos ou na associação destes com as palavras.

Para o contexto da Plataforma CpE, os documentos representam os títulos das publicações. Desse modo, cada publicação estará associada a um tópico e cada acadêmico terá o seu respectivo número de publicações distribuído entre os tópicos. Como resultado, teremos um segundo nível de áreas que serão definidas pelos tópicos identificados e uma caracterização da atuação dos acadêmicos.

Importante notar que cada palavra, pertinente ao conjunto de todos os documentos, tem sua frequência computada. Assim, a frequência de cada palavra é utilizada para a correspondente ordenação. O método prevê que as palavras mais e menos frequentes são desconsideradas na análise. As palavras descartadas não contribuem, efetivamente, para a caracterização dos documentos visto que ou elas são frequentes em um grande número de documentos ou são observadas em um grupo muito pequeno. Desse modo, é possível que alguns documentos, cujo título é composto por palavras que foram desconsideradas no processo, não tenham sido associados a algum tópico descoberto.

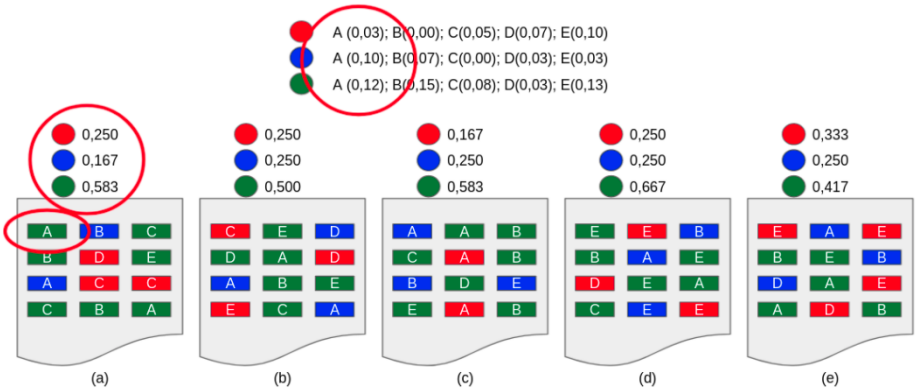

Fig. 3. Exemplo de atualização das proporções em função da alteração do tópico da palavra "A", no documento (a).

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* 3 (Jan): 993–1022, 2003.

Rossi, L., Gouveia, F. C., Mena-Chalco, J. P., et al. A evolução da pesquisa em ciência da informação: Uma análise de três revistas internacionais por topic modeling usando lda, 2020.