

Modelagem de tópicos

Prof. Walmes Zeviani

walmes@ufpr.br

Laboratório de Estatística e Geoinformação
Departamento de Estatística
Universidade Federal do Paraná

Justificativa e objetivos

- ▶ Classificação de documentos em grupos.
- ▶ Reconhecimento do assunto principal e secundários em cada grupo.
- ▶ Classificação não booleana mas *fuzzy*.
- ▶ Serve para:
 - ▶ Organização e resumo de coleções.
 - ▶ Sistemas de busca e recomendação.
 - ▶ Detecção de conteúdo duplicado.

Latent Dirichlet Allocation (LDA)

O modelo do LDA

- ▶ *Latent Dirichlet Allocation* (LDA) é o método padrão para modelagem de tópicos.
- ▶ Descrito por Blei et. al (2003): <https://www.seas.harvard.edu/courses/cs281/papers/blei-ng-jordan-2003.pdf>.
- ▶ Assume um modelo generativo:
 - ▶ Cada documento é uma mistura de tópicos.
 - ▶ Cada tópico é uma mistura de termos.
- ▶ *Correlated topic model* (CTM) é uma extensão do LDA.

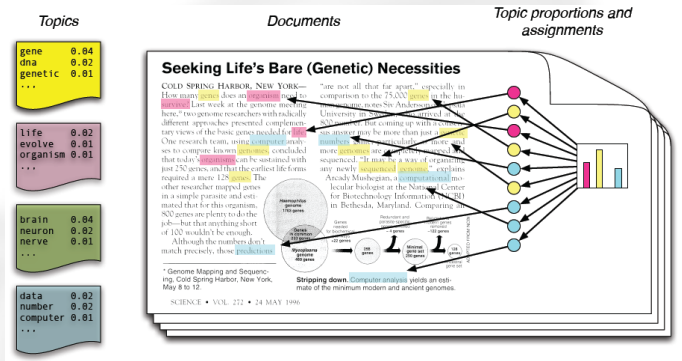


Figura 1. Uma ilustração do modelo generativo da alocação latente de Dirichlet. Fonte: <<http://www.scottbot.net/HIAL/index.html@p=221.html>>.

Funcionamento

- ▶ Segundo o LDA, o corpus é resultado de um **processo generativo**.
- ▶ Cada documento é uma mistura de K assuntos.
- ▶ Cada assunto possui uma distribuição de probabilidade para os V termos do vocabulário.
- ▶ Os tópicos são distribuições de probabilidade sobre o amplo vocabulário hipotético.
- ▶ Com esse modelo, em hipótese, se gera os documentos caso fossem conhecidos os parâmetros.
- ▶ É um modelo baseado em probabilidade condicional.

A distribuição de Dirichlet

- ▶ Dirichlet é a distribuição de probabilidades contínua que generaliza a multinomial do caso discreto.
- ▶ A função densidade de probabilidade é

$$f(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1},$$

em que $x_i \geq 0$, $\sum_{i=1}^K x_i = 1$ e $\alpha_i > 0$. $B(\alpha)$ é a função beta multinomial.

- ▶ X é uma variável aleatória composicional que representa o **teor** de cada tópico em um documento.
- ▶ Considere que, para cada assunto, existe uma distribuição de Dirichlet para os V termos dentro daquele assunto.

Recursos no `topicmodels`.

- ▶ O LDA está implementado no pacote `topicmodels`.
- ▶ O input da função é a matriz de documentos e termos.
- ▶ Retorna:
 - ▶ A proporção de cada tópico em cada documento (γ).
 - ▶ O peso de cada termo em cada tópico (β).
- ▶ A quantidade de tópicos é definida pelo usuário.
- ▶ É um modelo não supervisionado.
- ▶ Pode-se perfilar o K e examinar os resultados para pegar o mais satisfatório.

Outras abordagens

- ▶ Pacotes para fazer modelagem de tópicos:
 - ▶ `topicmodels`.
 - ▶ `lda`.
 - ▶ `LDAvis`.
- ▶ Outras abordagens similares/relacionadas:
 - ▶ Correlated topic model.
 - ▶ Família `word2vec`.