

Topics: $\beta_t \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_{20000})$, which is a prob over words for topics

eg: $[1/4 \quad 1/4 \quad 1/2]$

$w_1 \quad w_2 \quad w_3 \dots w_{50}$

Documents:

1. $\theta_d \sim \text{Dirichlet}(\gamma_1, \dots, \gamma_{50})$, which is distribution over topics

eg: $[1/4 \quad 1/2 \quad 1/4]$

$t_1 \quad t_2 \quad t_3 \dots t_{50}$

2. *Each word is assigned to topic Z_{id} (topic for word i in doc d)*

$Z_{id} \sim \text{Multi}(\theta_d)$, which means $\text{prob}(\text{word } i = \text{topic } 1) = \theta_{d,1}$

3. Each word is drawn from topic prob vector ω_{id}

$\omega_{id} \sim \text{Multi}(\beta_{Z_{id}})$

Then we can find:

$$\begin{aligned} &\beta_1, \dots, \beta_{50} \\ &\theta_1, \dots, \theta_N \\ &Z_{i,1}, \dots, Z_{i,n_d} \quad (i = 1, \dots, N) \end{aligned}$$

3.11 ColumbiaX: DS101X Statistical Thinking for Data Science and Analytics

word distribution for each topic:

city town Seattle zoology
topic 1 $\beta_1 = [0.1 \ 0.08 \dots 0.02 \dots 10^{-12}]$

cow sheep zoology
topic 50 $\beta_{50} = [0.01 \ 0.009 \dots \dots \dots 10^{-13}]$

topic distribution for each document:

topic 27 topic 43 topic 2
doc 1 : $\theta_1 = [0.35 \ 0.21 \dots \dots 0.01]$

topic 5 topic 6
doc 3721 : $\theta_{3721} = [0.21 \dots \dots \dots 0.01]$