Machine Learning course,
advanced track

Topic Modeling:
from PLSA to LDA

Anastasia Ianina

MIPT

04.10.2019

Outline

1. Topic modeling
2. Probabilistic latent semantic analysis (PLSA)
3. Latent Dirichlet Allocation (LDA)
4. Q & A

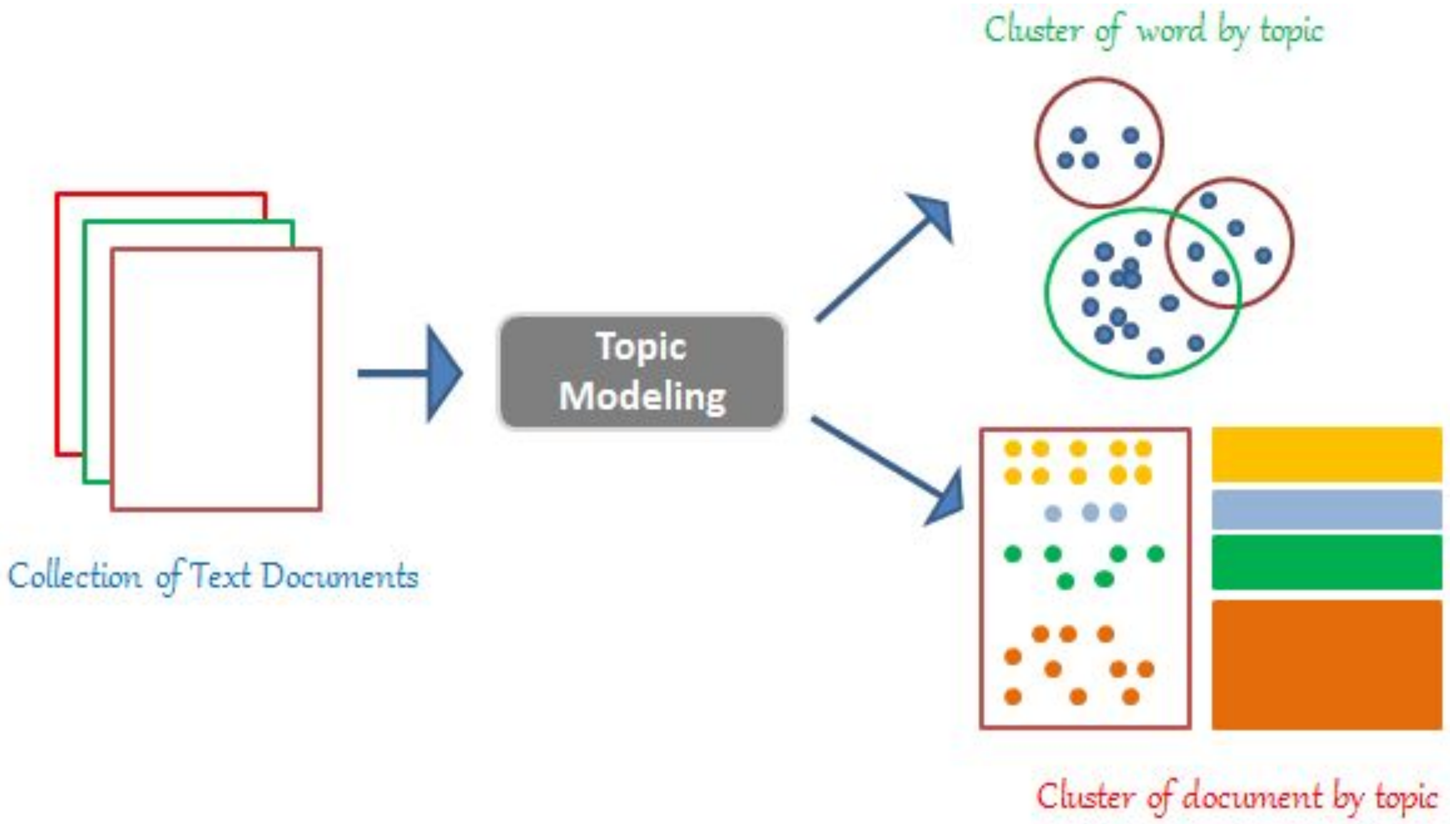
Based on: https://www.coursera.org/learn/language-processing/

Topic Modeling

Topic Modeling: why?

- We want to find topics in documents – useful for e.g. search or browsing
- We don’t want to do supervised topic classification
- Need an approach to automatically tease out the topics
- This is essentially a clustering problem - can think of both words and documents as being clustered

Topic Modeling



Topic Modeling

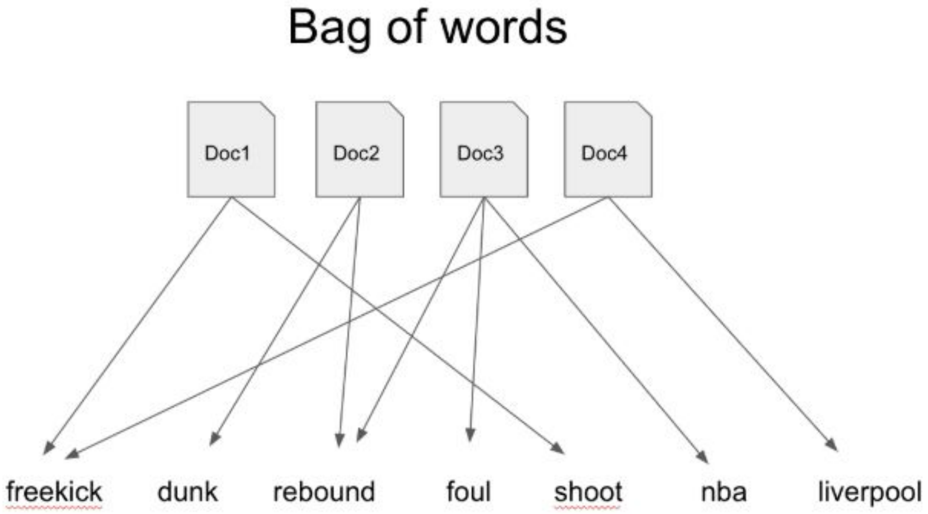
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

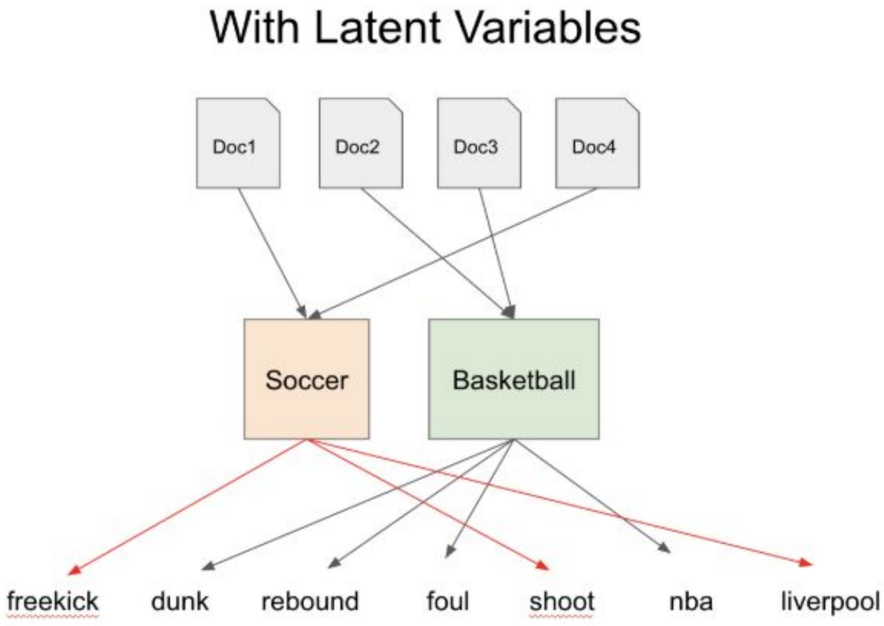
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Two assumptions:

- each **document** consists of a mixture of ***topics***, and
- each ***topic*** consists of a collection of **words**.

Topic Modeling





PLSA

- **Documents:** D={d1,d2,d3,...dN}, N is the number of documents. di denotes ith document in the set D.
- **Words:** W={w1,w2,...wM}, M is the size of our vocabulary. wi denotes ith word in the vocabulary W.
- **Topics:** T={t1,t2,...tk} — Latent or hidden variables. The number k is a parameter specified by us.

PLSA

Given:

- n_{wd} - a count of the word w in the document d

Find:

$\phi_{wt} = p(w|t)$ - probabilities of words in topics

$\theta_{td} = p(t|d)$ - probabilities of topics in documents

PLSA

$$p(w|d) = \sum_{t \in T} p(w|t, d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d)$$

PLSA

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d)$$

Law of total probability

$$p(w) = \sum_{t \in T} p(w|t)p(t)$$

PLSA

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d)$$

Law of total probability

$$p(w) = \sum_{t \in T} p(w|t)p(t)$$

Assumption of conditional independence

$$p(w|t,d) = p(w|t)$$

PLSA

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d)$$

Law of total probability

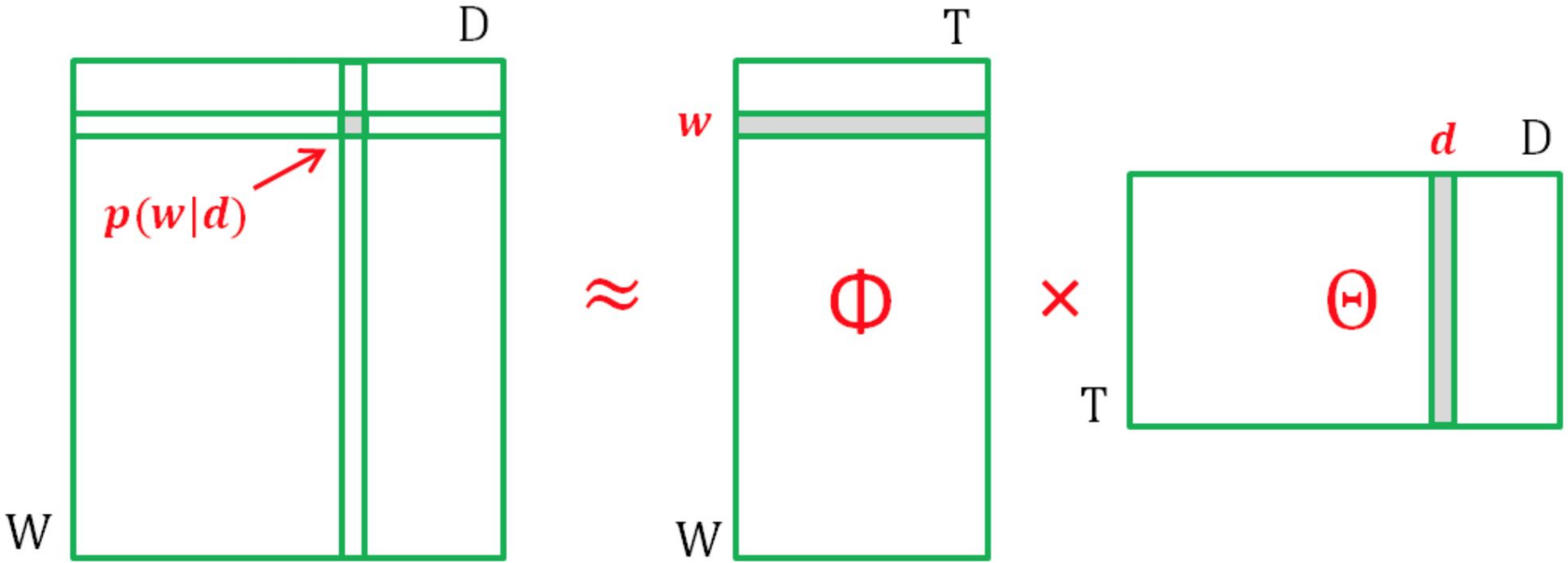
$$p(w) = \sum_{t \in T} p(w|t)p(t)$$

Assumption of conditional independence

$$p(w|t,d) = p(w|t)$$

PLSA

$$p(w|d) = p(w|t)p(t|d) = \phi_{wt}\theta_{td}$$



PLSA: how to train it?

Log-likelihood maximization:

$$log \prod_{d \in D} p(d) \prod_{w \in d} p(w|d)^{n_{dw}} \rightarrow \max_{\Theta,\Phi}$$

$$\sum_{d \in D} \sum_{w \in d} n_{dw} log \sum_{t \in T} \phi_{wt}\theta_{td} \rightarrow \max_{\Phi,\Theta}$$

PLSA: how to train it?

Log-likelihood maximization:

$$log \prod_{d \in D} p(d) \prod_{w \in d} p(w|d)^{n_{dw}} \rightarrow \max_{\Theta,\Phi}$$

$$\sum_{d \in D} \sum_{w \in d} n_{dw} log \sum_{t \in T} \phi_{wt}\theta_{td} \rightarrow \max_{\Phi,\Theta}$$

$$\phi_{wt} \geq 0 \qquad \theta_{td} \geq 0 \qquad \sum_{w \in W} \phi_{wt} = 1 \qquad \sum_{t \in T} \theta_{td} = 1$$

EM-algorithm

E-step:

$$p(t|d,w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}$$

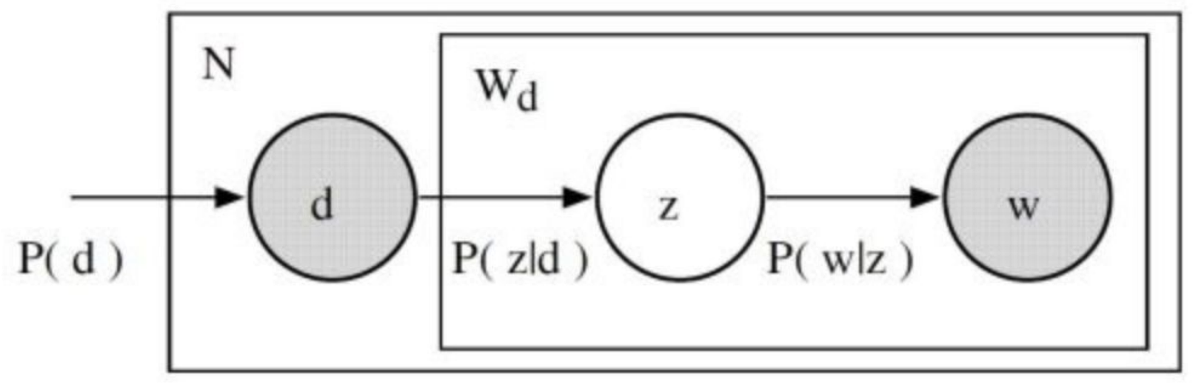
M-step:

$$\phi_{wt} = \frac{n_{wt}}{\sum_w n_{wt}} \qquad n_{wt} = \sum_d n_{dw} p(t|d,w)$$

$$\theta_{td} = \frac{n_{td}}{\sum_t n_{td}} \qquad n_{td} = \sum_w n_{dw} p(t|d,w)$$

PLSA: how to generate a document

- given a document d, topic z is present in that document with probability P(z|d)
- given a topic z, word w is drawn from z with probability P(w|z)



What’s wrong with PLSA?

LDA

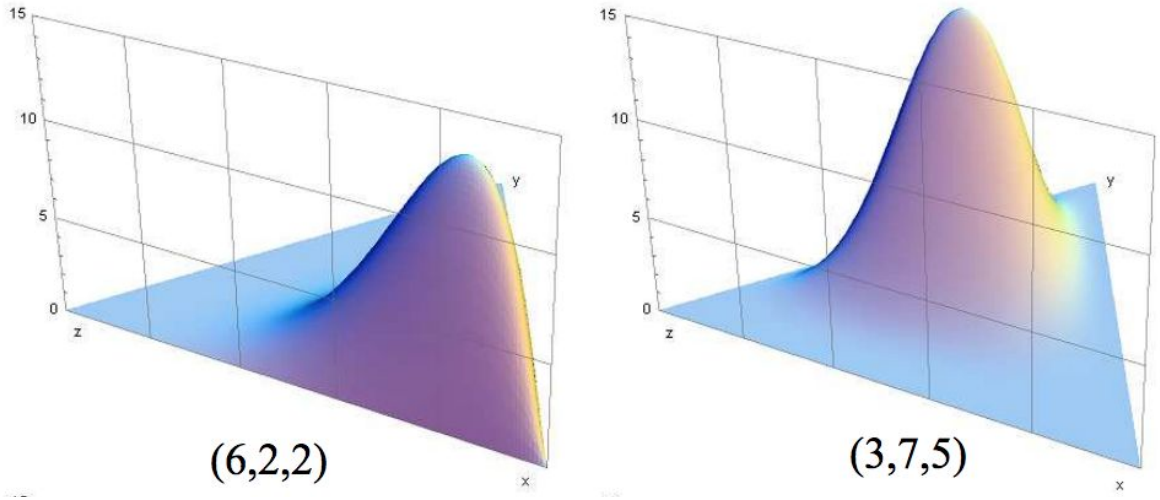
- LDA is a Bayesian version of PLSA

$$\mathrm{Dir}(\theta_d;\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \qquad \alpha_t > 0, \quad \alpha_0 = \sum_t \alpha_t, \qquad \theta_{td} > 0, \quad \sum_t \theta_{td} = 1;$$

$$\mathrm{Dir}(\varphi_t;\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \varphi_{wt}^{\beta_w - 1}, \quad \beta_w > 0, \quad \beta_0 = \sum_w \beta_w, \quad \varphi_{wt} > 0, \quad \sum_w \varphi_{wt} = 1.$$

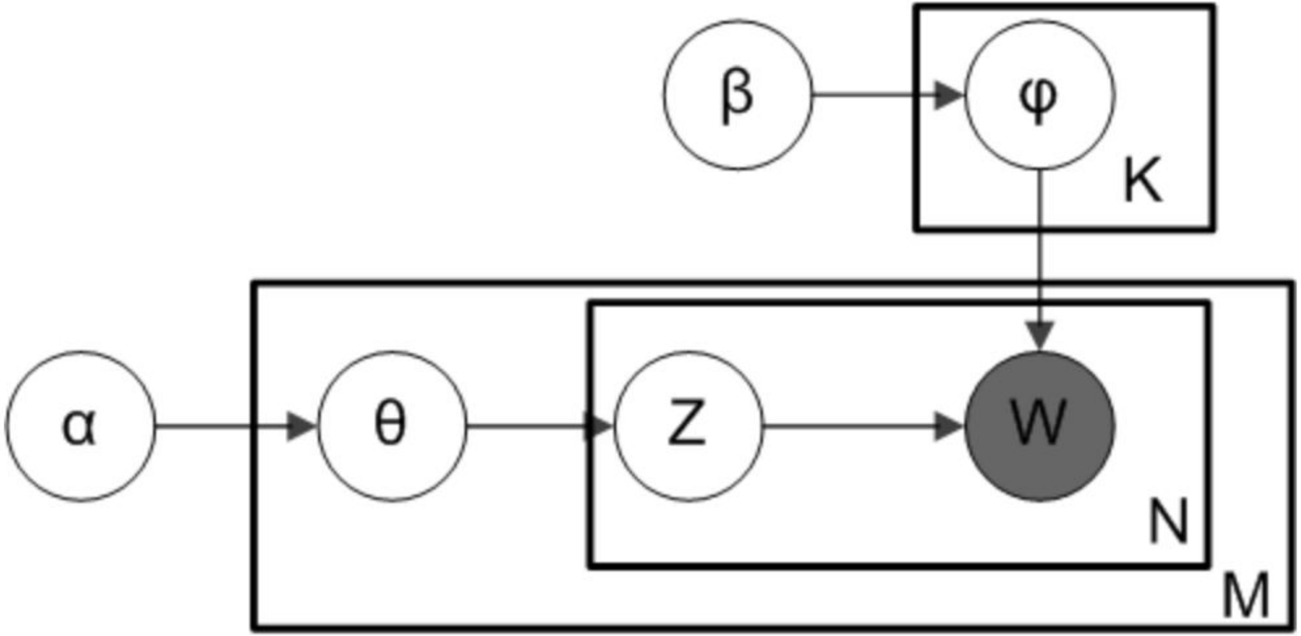
Dirichlet Distribution

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} x_i^{\alpha_i - 1}; \quad \text{for “observations”}: \sum_{i=1}^{d} x_i = 1, \quad x_i \geq 0$$



LDA

- LDA is a Bayesian version of pLSA



LDA

1. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\mathrm{Dir}(\alpha)$ is a Dirichlet distribution

2. Choose $\varphi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$ and β typically is sparse

3. For each of the word positions i, j, where $i \in \{1, \ldots, M\}$, and $j \in \{1, \ldots, N_i\}$

   (a) Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.

   (b) Choose a word $w_{i,j} \sim \mathrm{Multinomial}(\varphi_{z_{i,j}})$.

LDA

1. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\mathrm{Dir}(\alpha)$ is a Dirichlet distribution

2. Choose $\varphi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$ and β typically is sparse

3. For each of the word positions i, j, where $i \in \{1, \ldots, M\}$, and $j \in \{1, \ldots, N_i\}$

   (a) Choose a topic $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.

   (b) Choose a word $w_{i,j} \sim \mathrm{Multinomial}(\varphi_{z_{i,j}})$.

$$
\begin{aligned}
\boldsymbol{\varphi}_{k=1\ldots K} &\sim \mathrm{Dirichlet}_V(\boldsymbol{\beta}) \\
\boldsymbol{\theta}_{d=1\ldots M} &\sim \mathrm{Dirichlet}_K(\boldsymbol{\alpha}) \\
z_{d=1\ldots M, w=1\ldots N_d} &\sim \mathrm{Categorical}_K(\boldsymbol{\theta}_d) \\
w_{d=1\ldots M, w=1\ldots N_d} &\sim \mathrm{Categorical}_V(\boldsymbol{\varphi}_{z_{dw}})
\end{aligned}
$$