BT2101 Week 10
Topic Models

Agenda

Document Generation

- Document, Topic, Word
- p-LSA (Probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

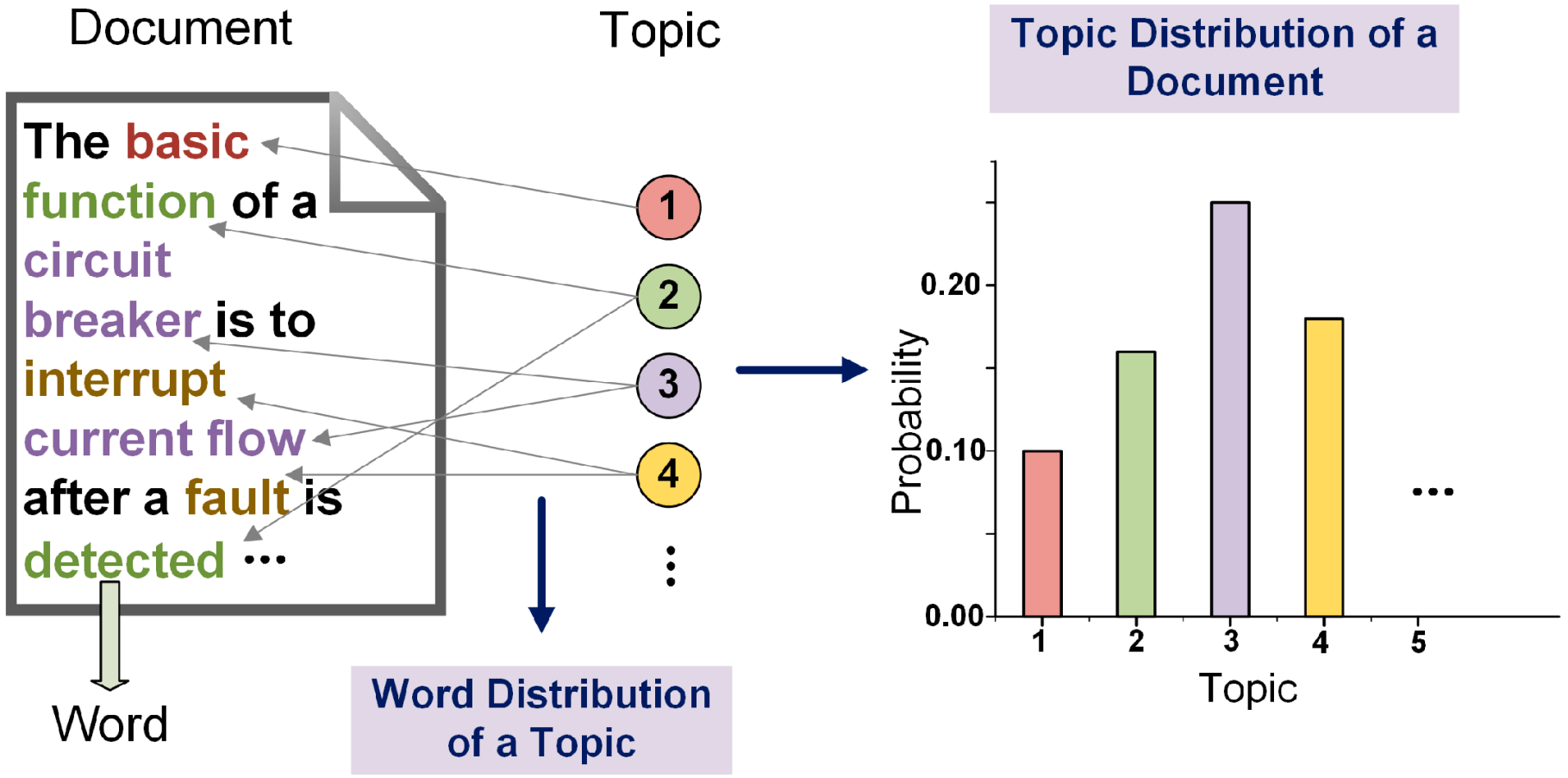
Estimations

- Frequentist Approach: Expectation-Maximization (EM)
- Bayesian Approach: Gibbs Sampling

Overview: Document Generation

Suppose you want to write an 1000-word article/document

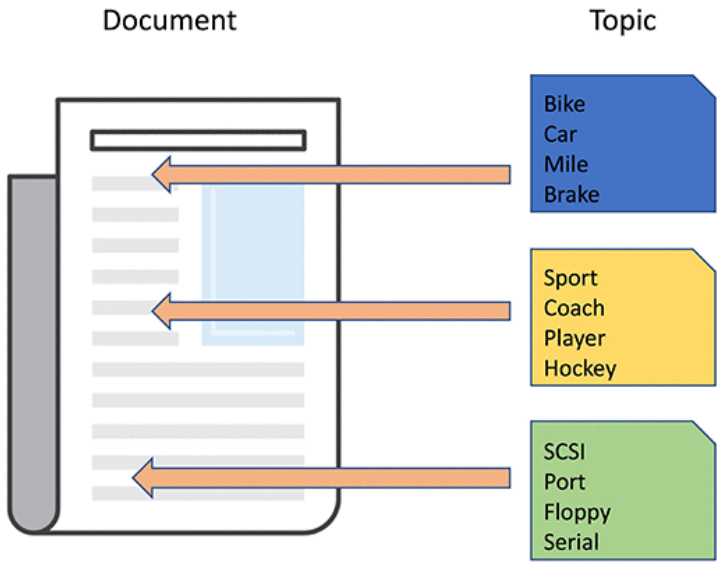
- First, you determine 1000 latent topics (e.g., math, CS, politics, arts, etc.) in each position;
- Second, for each topic, you choose one word that belongs to this topic, and fill in this position;
- Finally, you create an 1000-word article/document



Overview

Every document is a mixture of topics.

- We imagine that each document may contain words from several topics in particular proportions. For example, in a **2-topic model** we could say **"Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B."**

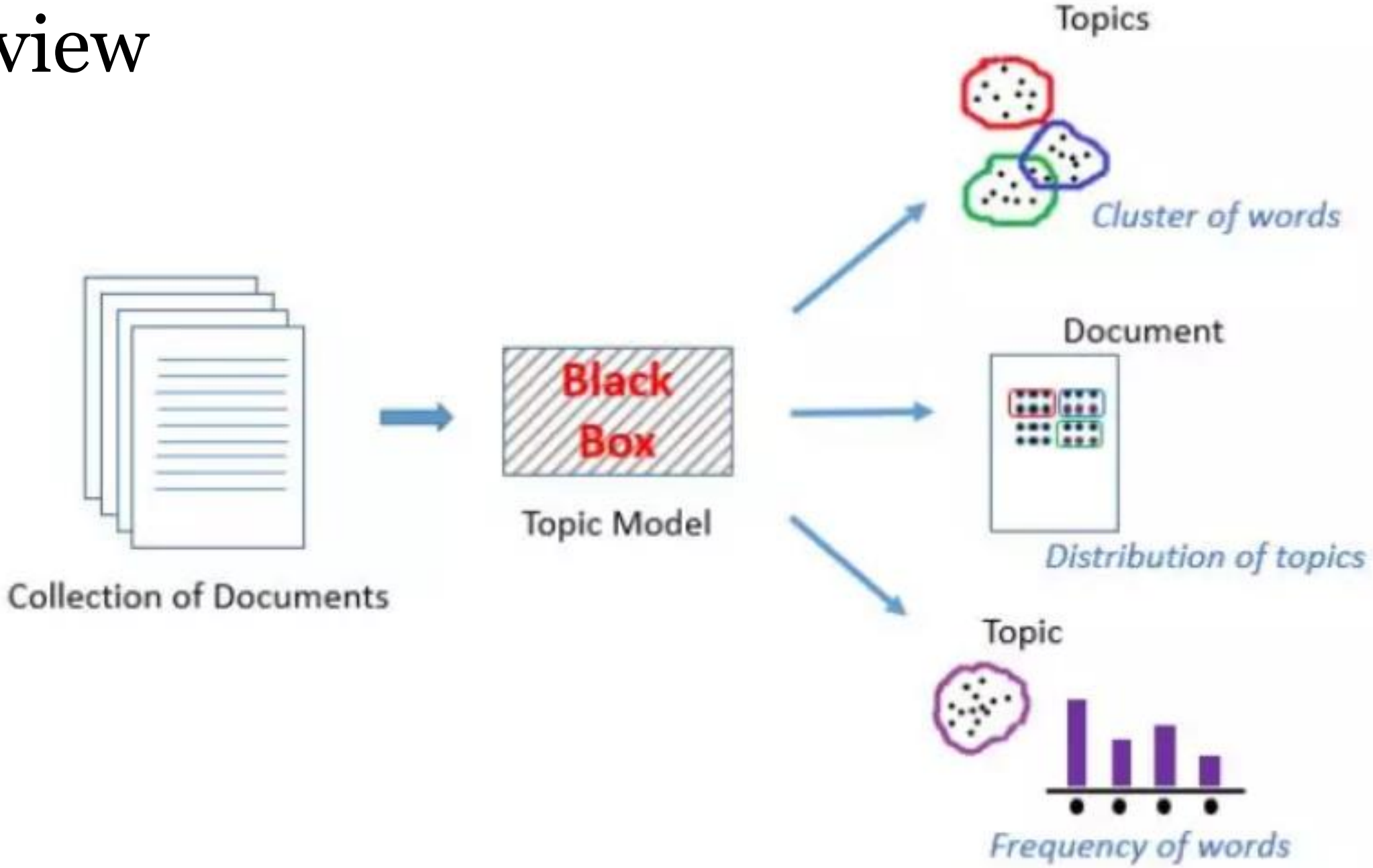


Every topic is a mixture of words.

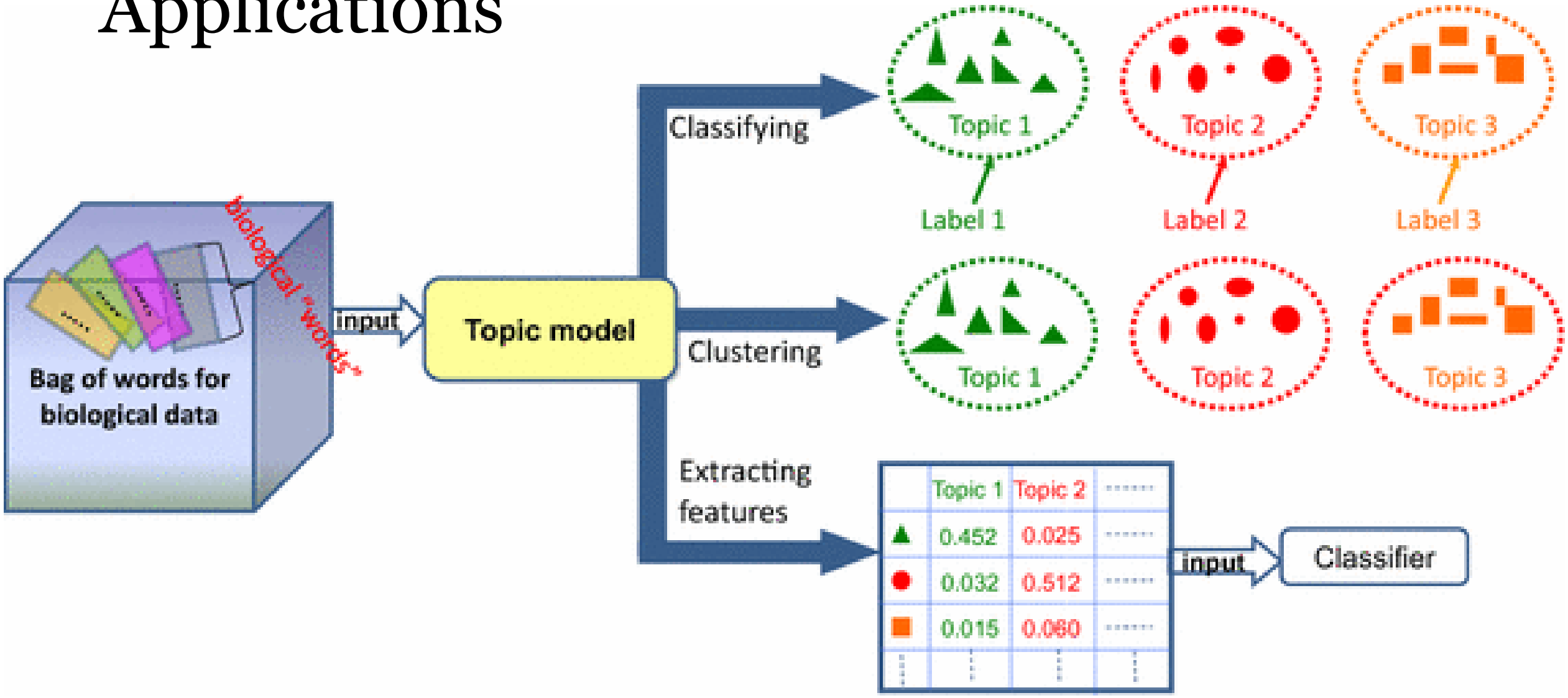
- For example, we could imagine a **2-topic model** of American news, with one topic for "politics" and one for "entertainment."
- The most common words in the **politics topic** might be "President", "Congress", and "government", while the **entertainment topic** may be made up of words such as "movies", "television", and "actor".
- Importantly, words can be shared between topics: A word like "budget" might appear in all topics equally.

topic	proof	induction	object	bouquet	memory
maths	0.45	0.35	0.2	0	0
comp sc.	0.23	0.17	0.35	0	0.25
wine	0	0	0.1	0.75	0.15

Overview

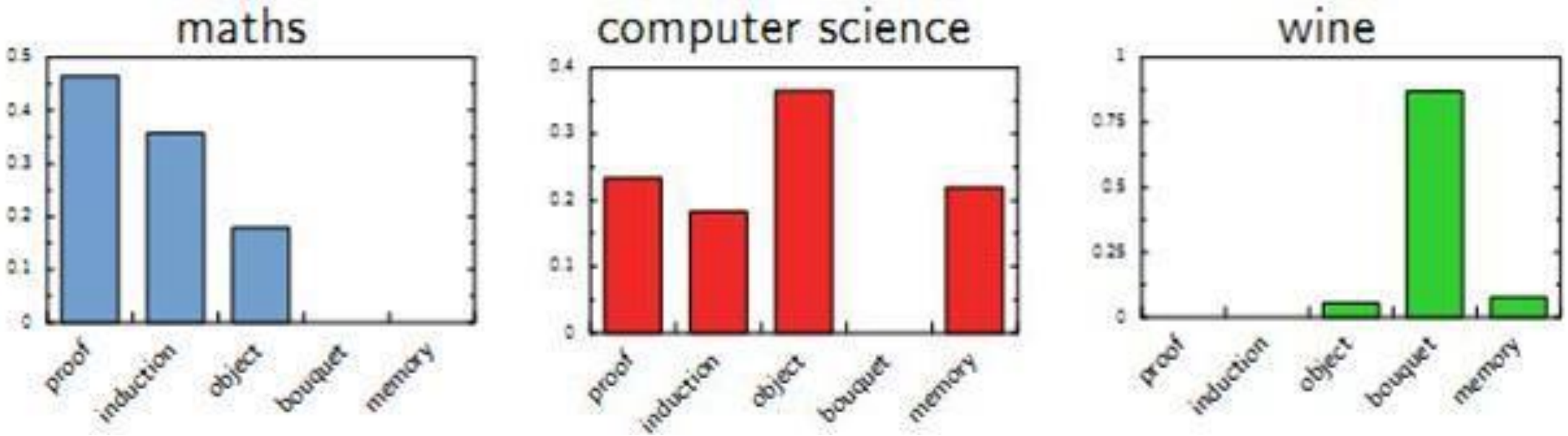

Topics
Cluster of words
Document
Distribution of topics
Topic
Frequency of words
Black Box
Topic Model
Collection of Documents

Applications

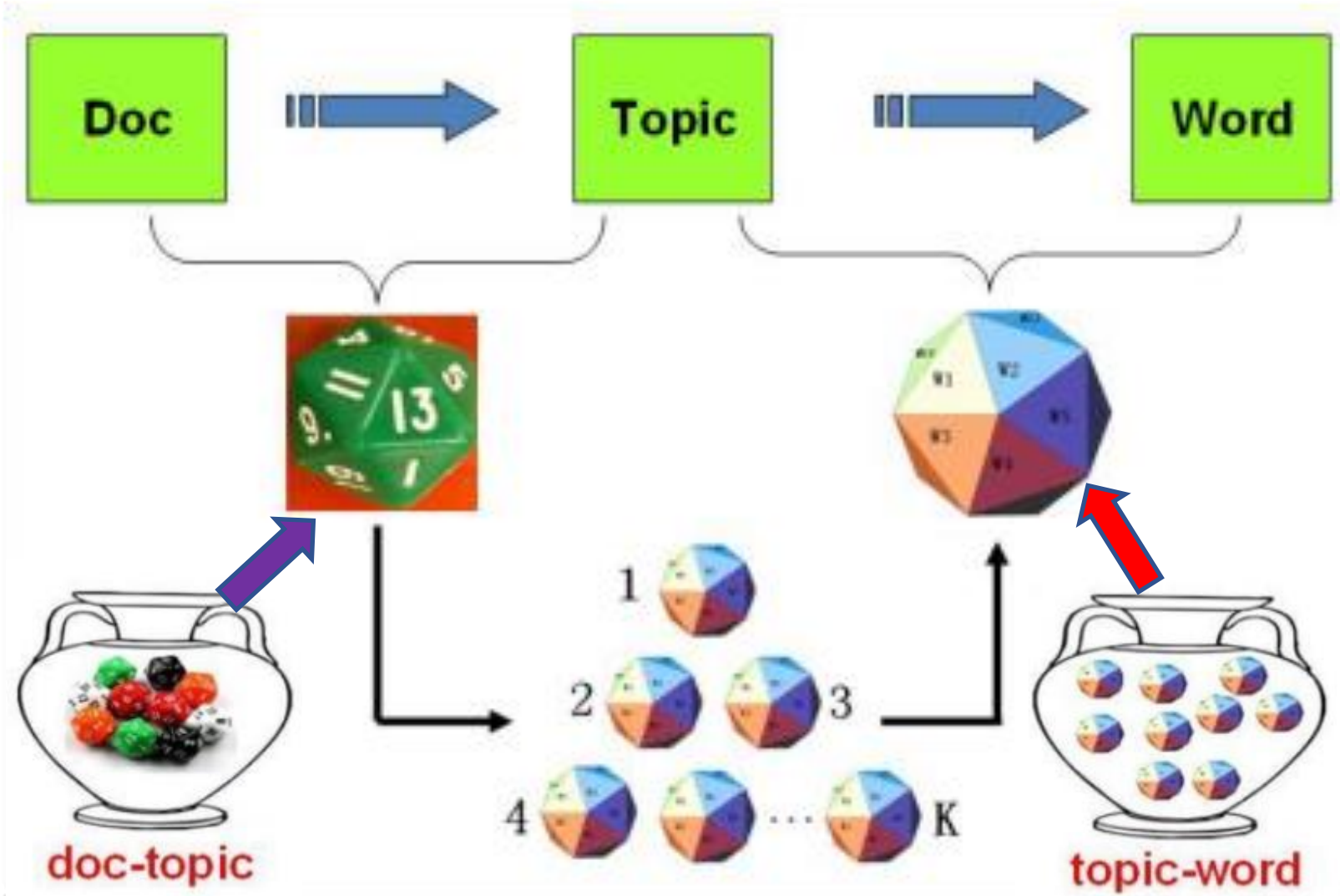


Topic Models: Important Terms

- Documents
  - For example, an article can be treated as a document
  - Usually you will have multiple documents/articles, and each document is a (fixed) probability distribution of multiple topics
- Topics
  - A topic is a latent cluster of multiple words with a (fixed) probability distribution
  - For example: Math, CS, Wine, Politics, Sports, etc.



Topic Models: Document Generation



(1) Generate a latent topic:

- Select the **"document-topic" dice** of this document
- Roll the dice
- Get the topic z from K available topics;

(2) Generate an observed word:

- Select the **"topic-word" dice** of this topic z
- Roll the dice
- Get the word w from V available words

Topic Models: Document Generation

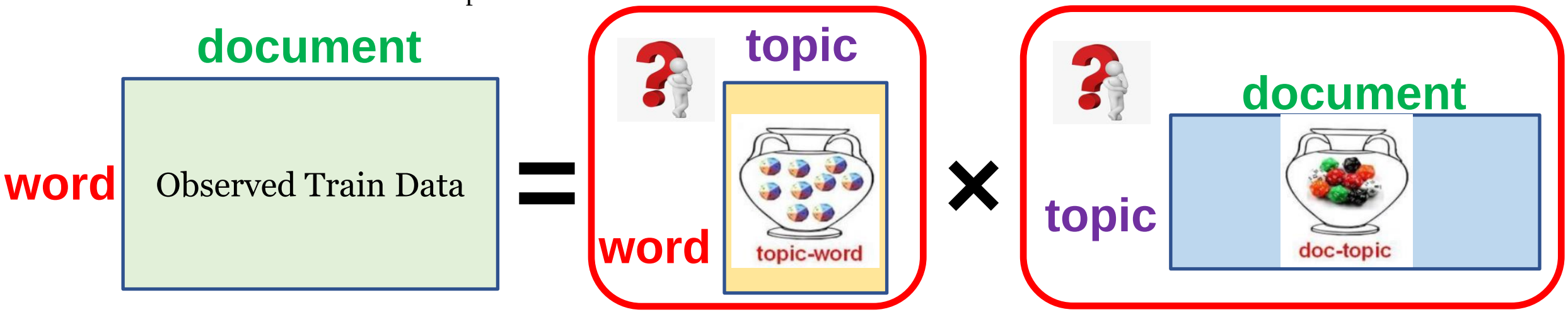
(1) Generate a latent topic:

- Select the **"document-topic"** dice of this document
- Roll the dice
- Get the topic z from K available topics;

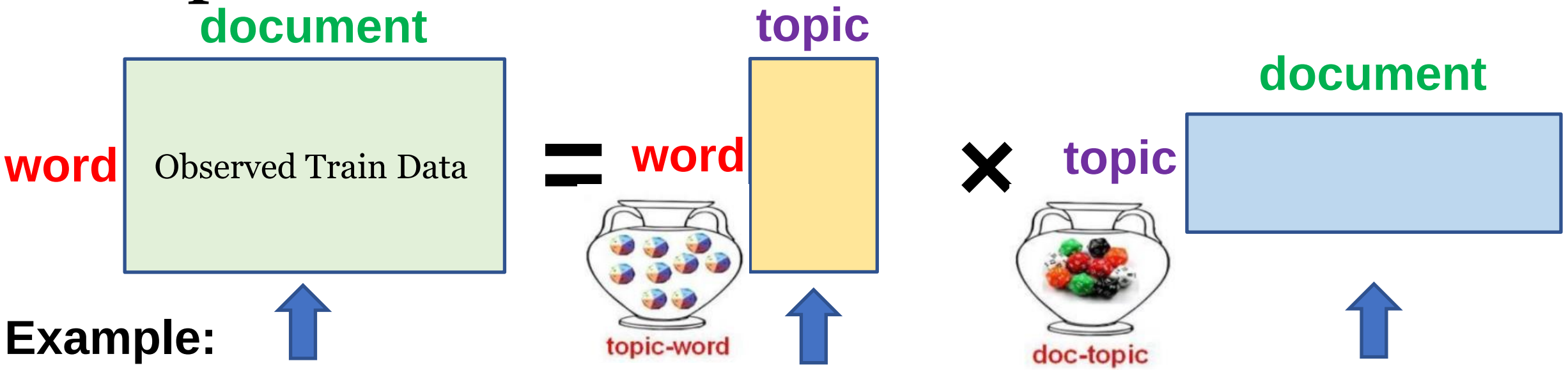
(2) Generate an observed word:

- Select the **"topic-word"** dice of this topic z
- Roll the dice
- Get the word w from V available words

$$\Pr(\text{word}|\text{document}) = \sum_{\text{z from k topics}} \Pr(\text{word}|\text{topic z}) \times \Pr(\text{topic z}|\text{document})$$



Topic Models: Document Generation



Example:

Word -Doc	Doc 1	Doc 2	…	Doc M
Word 1	65	2	… …	90
Word 2	72	3	… …	12
… …	… …	… …	… …	… …
Word V	50	1	1	5

Topic-Word	Topic 1	Topic 2	…	Topic K
Word 1	13	36	…	1
Word 2	27	33	…	3
… …	… …	… …	…	… …
Word V	20	101	…	9

Doc-Topic	Doc 1	Doc 2	…	Doc M
Topic 1	62	3	…	17
Topic 2	27	47	…	85
… …	… …	… …	…	… …
Topic K	11	9	…	22

Topic Models: Estimations

Frequentist Approach

- pLSA: Probabilistic Latent Semantic Analysis
- https://arxiv.org/ftp/arxiv/papers/1301/1301.6705.pdf
- https://en.wikipedia.org/wiki/Probabilistic_latent_semantic_analysis

Bayesian Approach

- LDA: Latent Dirichlet Allocation
- David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3, p993-1022, 2003
- http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Topic Models: Estimations

- pLSA (Probabilistic Latent Semantic Analysis):
  - $p(d_m)$: Pr. of selecting a document d_m
  - $p(z_k|d_m)$: Pr. of selecting a topic z_k in this document d_m
  - $p(w_n|z_k)$: Pr. of selecting a word w_n from this topic z_k
  - $n(d_m, w_n)$: Number of occurrence of pair (d_m, w_n)

Max. Likelihood Function:

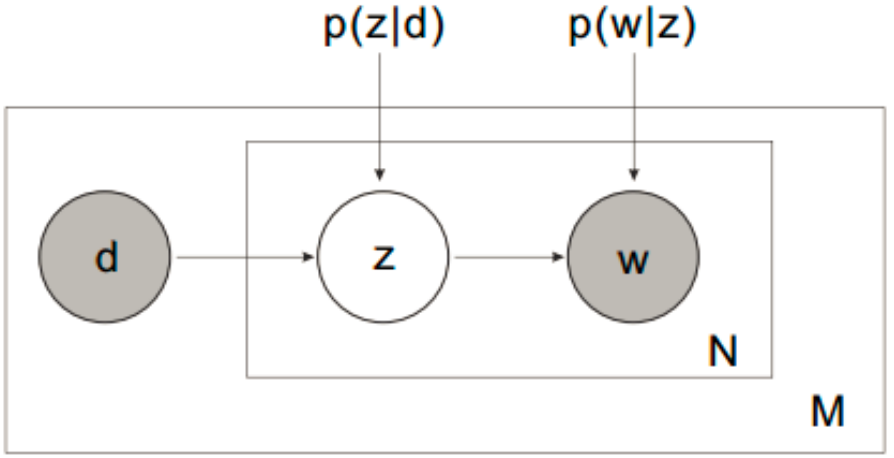
$$L = \prod_m^M \prod_n^N p(d_m, w_n)^{n(d_m, w_n)}$$

$$ll = log(L) = \sum_m^M \sum_n^N n(d_m, w_n) \times \log p(d_m, w_n)$$

$$= \sum_m^M \sum_n^N n(d_m, w_n) \times \log(p(d_m) \times p(w_n | d_m))$$

$$= \sum_m^M \sum_n^N n(d_m, w_n) \times \log p(d_m) \times \sum_k^K p(z_k | d_m) p(w_n | z_k)$$

$$= \sum_m^M \sum_n^N n(d_m, w_n) \times \log\{\sum_k^K p(d_m) p(z_k | d_m) p(w_n | z_k)\}$$

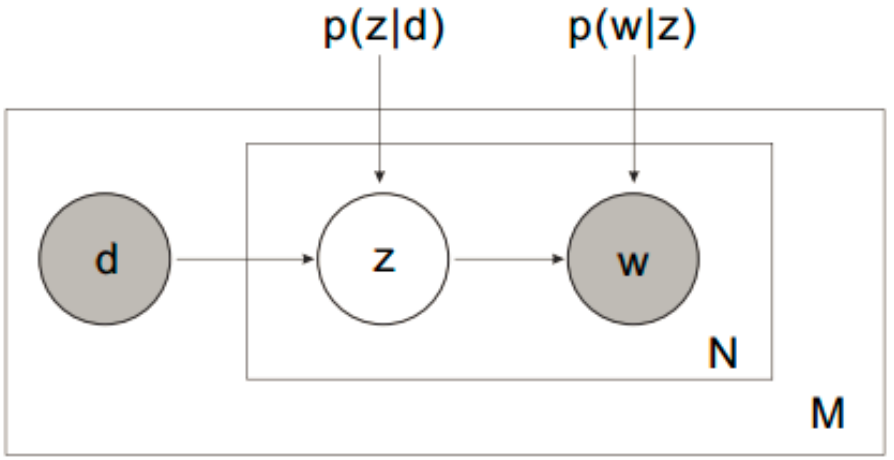


Question: How to get $p(z_k|d_m)$ and $p(w_n|z_k)$ such that ***ll*** is maximized ?

- Expectation-Maximization Algorithm (EM)

https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

Topic Models: Prediction



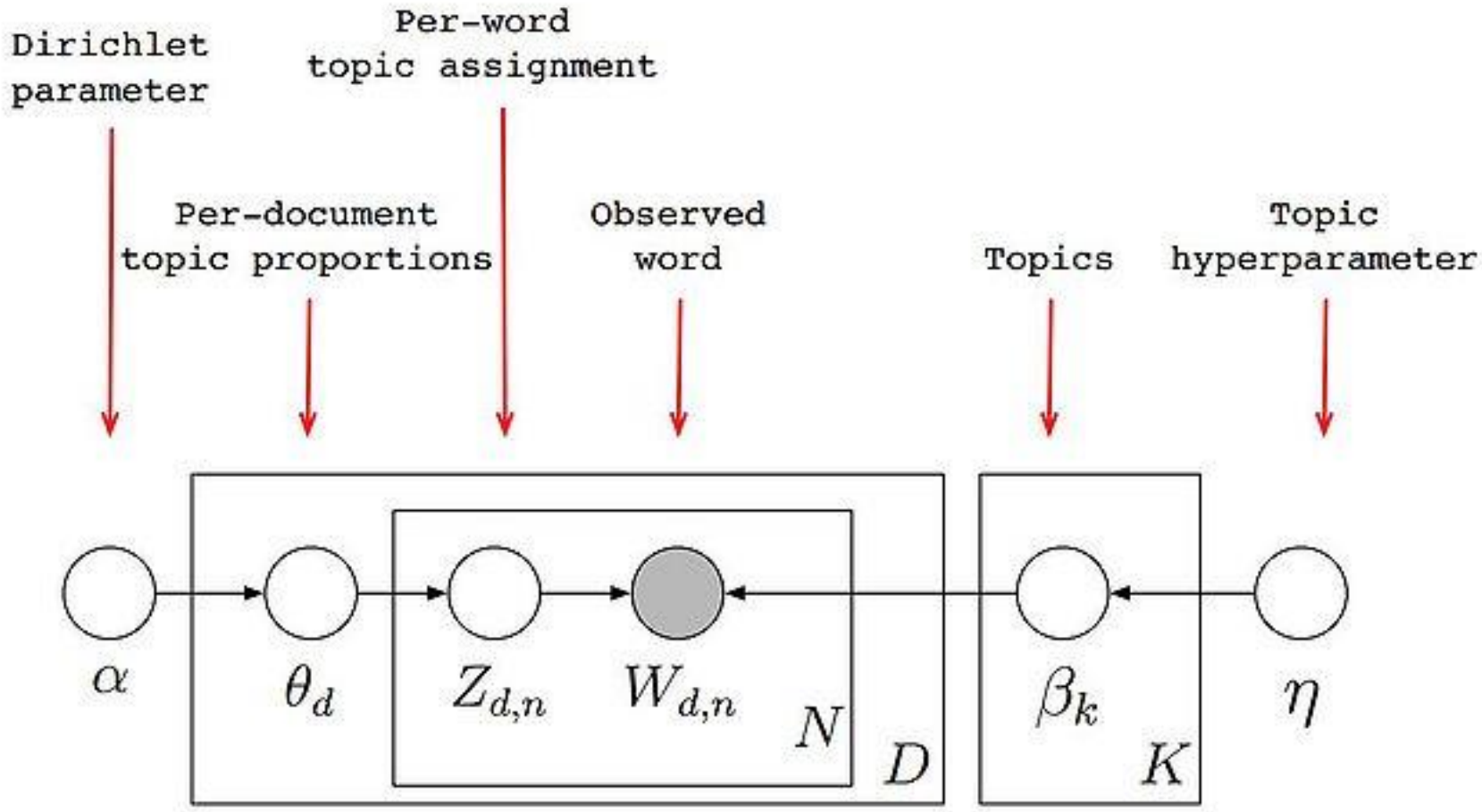
- pLSA (Probabilistic Latent Semantic Analysis):
  - $p(d_m)$: Pr. of selecting a document d_m
  - **$p(z_k | d_m)$:** Pr. of selecting a topic z_k in this document d_m
  - **$p(w_n | z_k)$:** Pr. of selecting a word w_n from this topic z_k
  - $n(d_m, w_n)$: Number of occurrence of pair (d_m, w_n)

Question: Given a new document d_{new}, what is the topic distribution of this new document ?

- Goal: What are $\Pr(z_k | d_{new})$
- Include this document in the **EM algorithm** again, and obtain $\Pr(z_k | d_{new})$

Topic Models: Estimations

- LDA (Latent Dirichlet Allocation):
  - Bag-Of-Words Model (i.e., do not care about word order)
  - Bayesian estimation: Gibbs Sampling



Topic Models: Background Knowledge

- Bayesian Statistics
  - Difference between Bayesian Statistical Inference and Classifical/Frequentist Statistical Inference
  - Bayes Law: Prior, Data Likelihood, Posterior Distribution, Conjugate
  - MCMC (Markov Chain Monte Carlo): Gibbs Sampling

- Probability
  - Beta Function, Gamma Function
  - Binomial/Multinomial Distribution: Data Likelihood
  - Dirichlet Distribution: Prior and Posterior Distribution

Topic Models: Bayes Law

- Some Terms
  - Prior Distribution: $P(\theta)$
  - Likelihood of Data X with f features: $Likelihood = P(X = (x_1, \ldots, x_f) \mid \theta)$
  - Posterior Distribution: $P(\theta \mid X = (x_1, \ldots, x_f))$
  - Bayes Formula:

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)} \propto P(X \mid \theta)P(\theta)$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence} \propto Likelihood \times Prior$$

- Conjugate Prior
  - If Prior and Posterior probability distribution are from the same family, then we call this Prior is a **conjugate prior** to Data Likelihood
  - Example 1: Prior is Normal, Data Likelihood is Normal, then Posterior is still Normal
  - Example 2: Prior is Inverse-Gamma, Data Likelihood is Normal, then Posterior is still Inverse-Gamma

Topic Models: Bayesian Inference

- Frequentist Statistics

$$\max_{\theta_{MLE}} P(\text{Data } X \mid \theta)$$

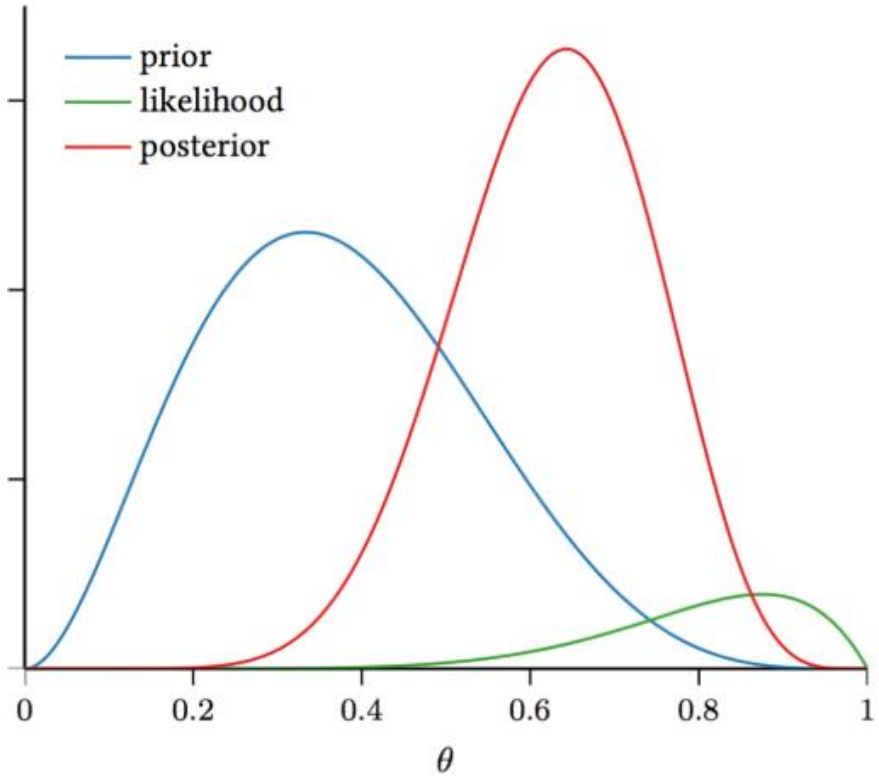
- Bayesian Statistics

$$P(\theta \mid \text{Data } X) \propto P(\text{Data } X \mid \theta) P(\theta)$$

$$E(\theta) \text{ from } P(\theta \mid \text{Data } X)$$

Topic Models: Bayesian Inference

- Frequentist Statistics
  - Parameter value is unknown but fixed
  - Example: Flipping a coin and estimate **θ=Pr(Head)**
  - Scenario 1: Repeat 10 times, 5 Heads and 5 Tails → $\theta_{MLE}=0.5$
  - Scenario 2: Repeat another 10 times, 9 Heads and 1 Tail → $\theta_{MLE}=0.9$ **(Problematic ?)**
- Bayesian Statistics **(Prior serves as regularization)**
  - Parameter value is random but follows a fixed probability distribution (Mean value is meaningful)
  - Example: Flipping a coin and estimate **θ=Pr(Head)**
  - **Prior Knowledge:** This coin may not be too biased → $\theta_0=0.4$
  - Scenario 1: Repeat 10 times, 5 Heads and 5 Tails → $E[\theta|\text{Data}, \theta_0]\approx 0.5$
  - Scenario 2: Repeat 10 times, 9 Heads and 1 Tail → $E[\theta|\text{Data}, \theta_0]\approx 0.5$ **(Great !)**



Topic Models: Probability Distributions

- Beta Function
  - The Euler integral of the 1st kind:

$$B(\alpha_1,\ldots,\alpha_j)=\frac{\prod_{i=1}^{j}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{j}\alpha_i)}$$

- Gamma Function
  - An extension of factorial function:

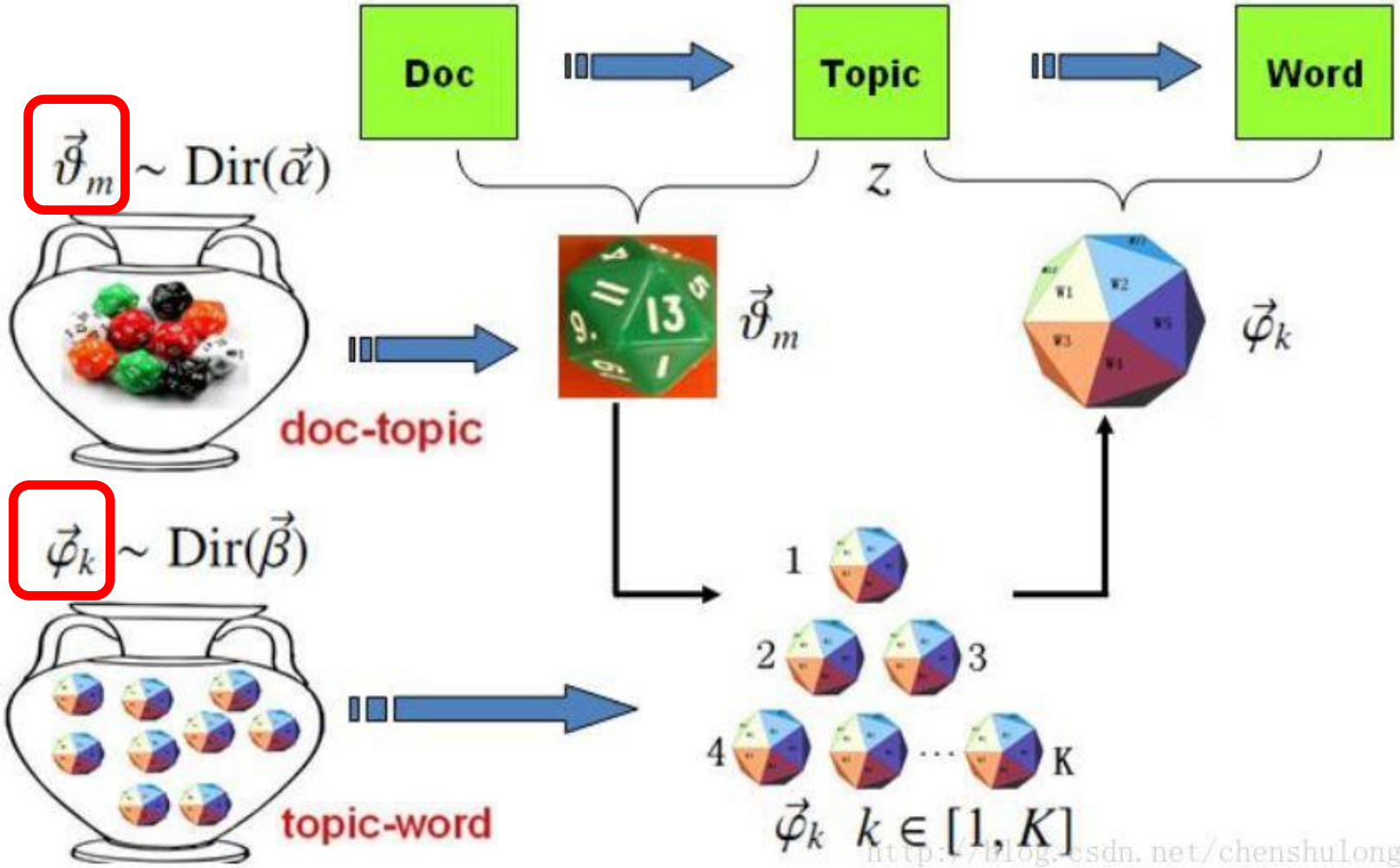
$$\Gamma(\alpha_i)=\int_0^{\infty}x^{\alpha_i-1}e^{-x}dx\approx(\alpha_i-1)!$$

Topic Models: Gibbs Sampling

- Randomly select a sample of random variables (z, θ, φ) from a **complex joint distribution** $p(z, \theta, \varphi)$:
  - 1. At round t=0: Start with initial values $(z^{(t=0)}, \theta^{(t=0)}, \varphi^{(t=0)})$;
  - 2. In current round t, Alternating the following in order:
  - 2.1. Select a sample of $z^{(t)}$ from its **marginal distribution**: $z^{(t)} \sim p(z^{(t)} | \theta^{(t-1)}, \varphi^{(t-1)})$;
  - 2.2. Select a sample of $\theta^{(t)}$ from its **marginal distribution**: $\theta^{(t)} \sim p(\theta^{(t)} | z^{(t)}, \varphi^{(t-1)})$;
  - 2.3. Select a sample of $\varphi^{(t)}$ from its **marginal distribution**: $\varphi^{(t)} \sim p(\varphi^{(t)} | z^{(t)}, \theta^{(t)})$;
  - 2.4. Round $t \leftarrow t+1$
  - 3. Iterating Step 2, until the distribution of these random variables are converging (e.g., after round t=1000, their mean values become stable)

https://en.wikipedia.org/wiki/Gibbs_sampling

Topic Models: Multinomial Distribution



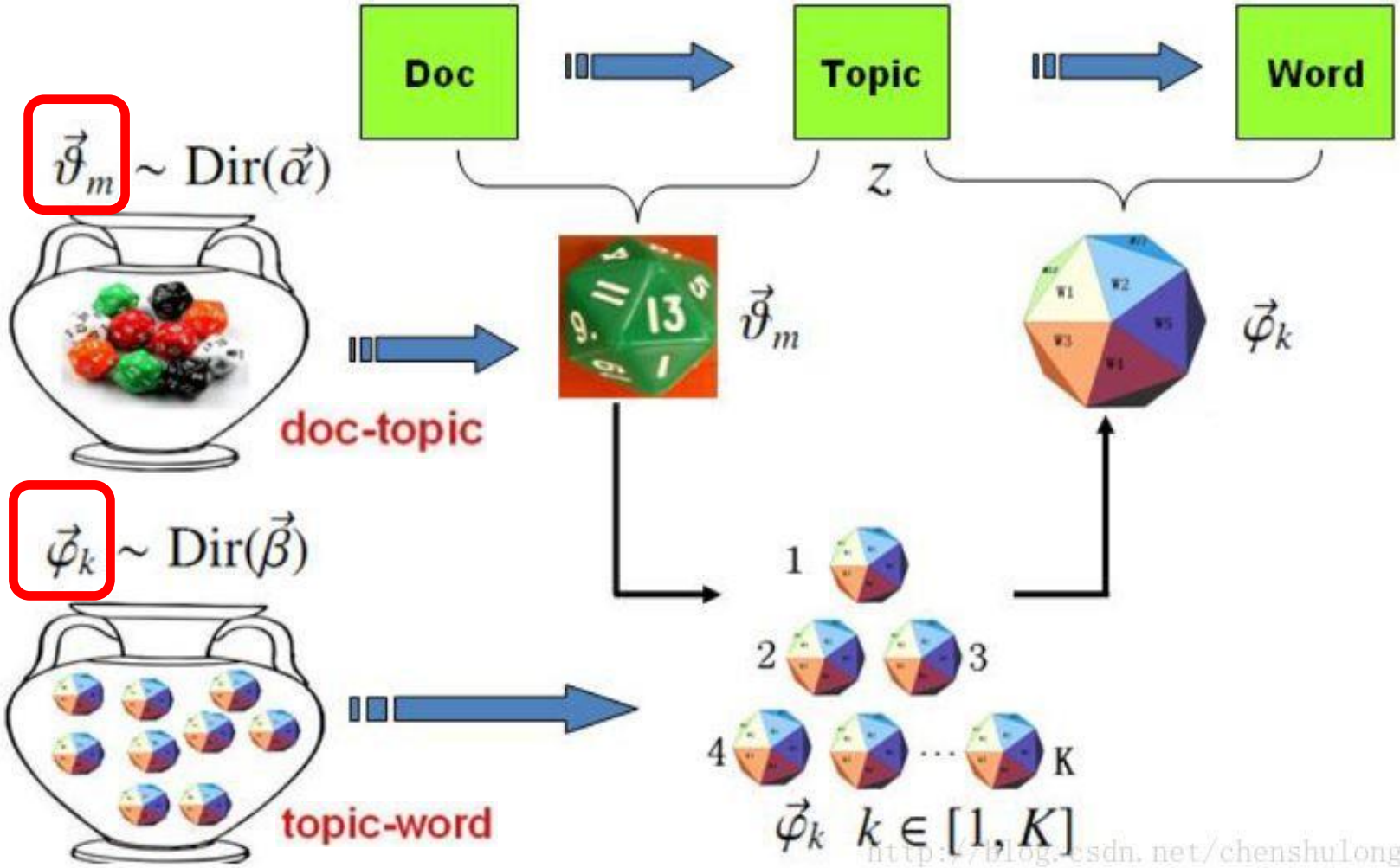
(1) **Document-Topic distribution** and **Topic-Word distribution** are both **Multinomial Distribution**

(2) Multinomial Distribution: Assume a dice has **K unique faces/values**, each face has a probability **p_k to be chosen**, repeat **n times**, then the probability distribution of each value k (k=1,…,K) appears **x_k times** is:

$$P(x_1,...,x_K;n;p_1,...,p_K)$$

$$=\frac{n!}{x_1!...x_K!}p_1^{x_1}...p_K^{x_K}$$

Topic Models: Dirichlet Distribution

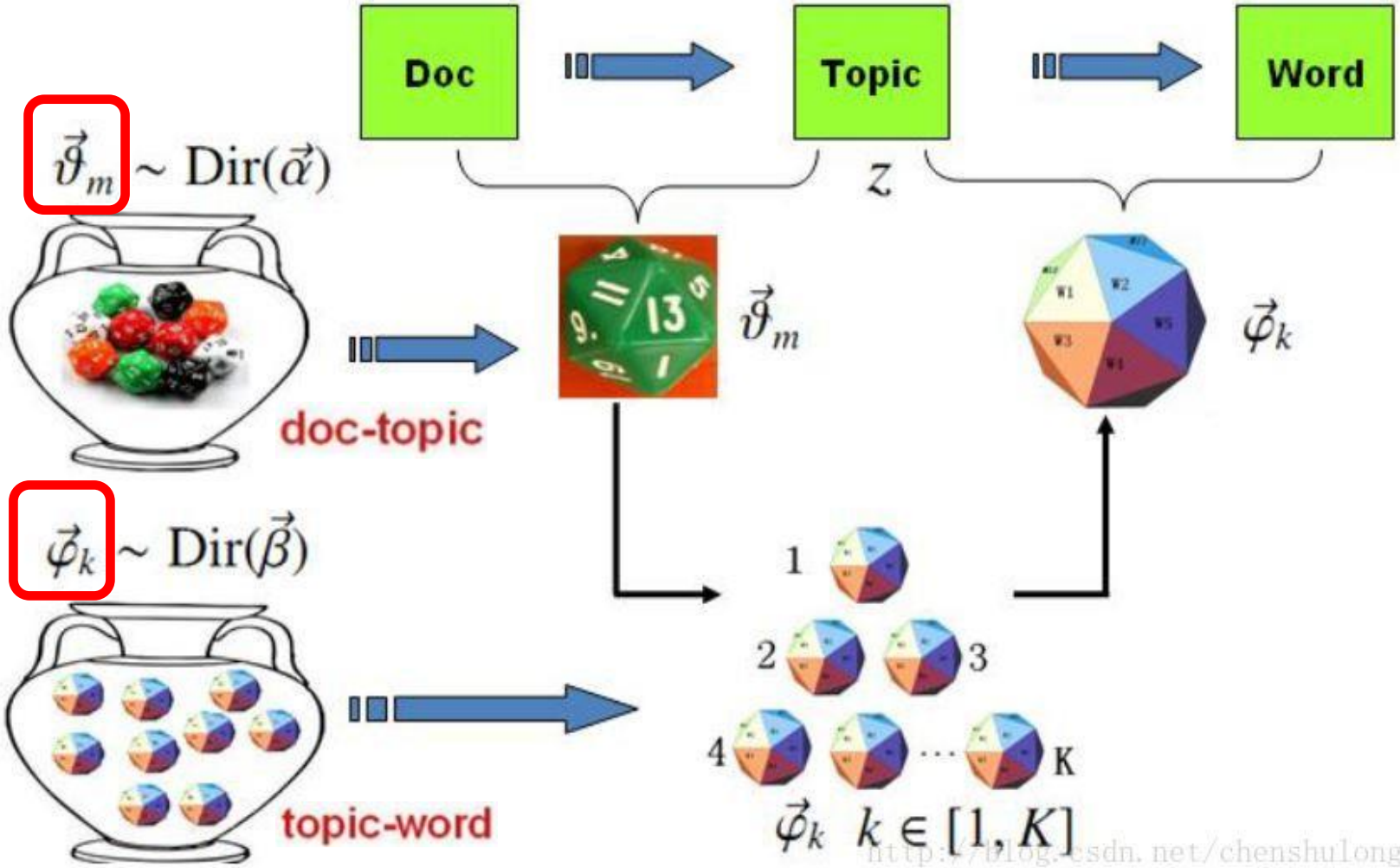


(1) $\mathbf{p_k}$ (k=1,…,K) in the **Document-Topic distribution** and **Topic-Word distribution** are not fixed, but random values drawn from prior probability distribution

(2) Dirichlet Distribution: Assume a dice has **K unique faces/values**, each face has a probability $\mathbf{p_k}$ **to be chosen**, then the probability density of $\mathbf{p_k}$ (k=1,…,K) is:

$$f(p_1,\ldots,p_K;\alpha_1,\ldots,\alpha_K) = \frac{1}{B(\alpha_1,\ldots,\alpha_K)}\prod_{k=1}^{K} p_k^{\alpha_k-1}$$

Topic Models: Document Generation



(1) Generate k topics: You randomly select k **topic-word dices** from **Dirichlet(β);**

(2) Generate a document m with N_m words:

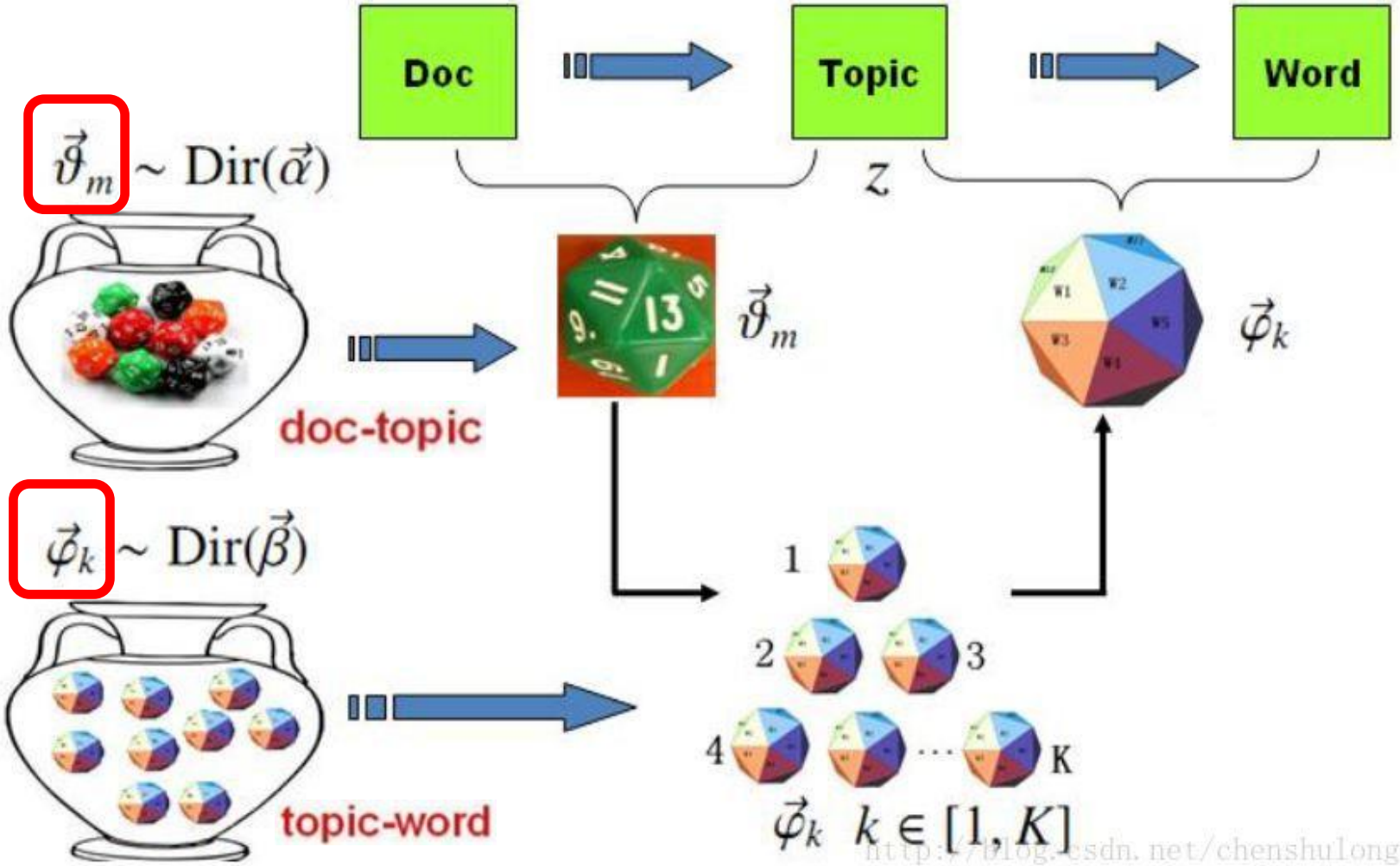
A. You randomly select a **document-topic dice** from **Dirichlet(α)**;

B. Roll this dice and get a topic z_{mn} for n^{th} position in this document;

C. Choose the **topic-word dice** whose topic is z_{mn} from step (1), roll this dice and get a word w_{mn}, to fill in this n^{th} position;

D. Repeat b and c, until get N_m words for this document

Topic Models: Bayesian Updating



(1) Bayesian updating on the posterior distribution of latent variable:

$$p(z, \theta, \varphi \mid \alpha, \beta, w)$$

(2) Using Gibbs sampling, we will get posterior distribution of these latent variables, and their **posterior distributions**

(3) We use **posterior mean values** as the **estimated values** of these latent variables

Topic Models: Bayesian Updating

- Dirichlet Prior is a conjugate prior to Multinomial Data likelihood
  - Bayesian Updating on Latent variables: topics z, topic-word distribution φ, and document-topic distribution θ; **Gibbs Sampling** on z, φ and θ;
  - Obtain **posterior distribution** of each latent variable, and then obtain **posterior means values** as **estimators** of these latent variables;
  - Detailed math proof and equations are provided in references;

For i (document m and position n): $p(z_i = k \mid z_{-i}, w) \propto \hat{\theta}_{mk} \times \hat{\varphi}_{kt}$

Document m and Topic k: $\hat{\theta}_{mk} = \frac{\alpha_k + n_{m,-i}^{(k)}}{\sum_{k=1}^{K} (\alpha_k + n_{m,-i}^{(k)})}$

Topic k and Word t: $\hat{\varphi}_{kt} = \frac{\beta_t + n_{k,-i}^{(t)}}{\sum_{t=1}^{V} (\beta_t + n_{k,-i}^{(t)})}$

Repeat updating alternately and iteratively, Until the posterior distributions of them are converging

$n_{m,-i}^{(k)}$: Number of words in doc m that belongs to topic k but different from word in position n

$n_{k,-i}^{(t)}$: Number of word t in topic k that is different from word in position n

Topic Models: Bayesian Updating

Word-Doc	Doc 1	…	Doc M
Position 1	Trump	… …	CS
Position 1	Kobe	… …	Java
… …	… …	… …	… …
Position N_m	Musician	1	integral

Word-Doc	Doc 1	…	Doc M
Position 1	Topic 1	… …	Topic 4
Position 1	Topic 2	… …	Topic 2
… …	… …	… …	… …
Position N_m	Topic 3	1	Topic 5

At round t,

For i (document m and position n): $p(z_i = k \mid z_{-i}, w) \propto \hat{\theta}_{mk} \times \hat{\varphi}_{kt}$

Topic Models: Bayesian Updating

Word-Doc	Doc 1	…	Doc M
Position 1	Topic 1	… …	Topic 4
Position 1	Topic 2	… …	Topic 2
… …	… …	… …	… …
Position N_m	Topic 3	1	Topic 5

Doc-Topic	Doc 1	…	Doc M
Topic 1	62	…	17
Topic 2	27	…	85
… …	… …	…	… …
Topic K	11	…	22

Topic-Word	Topic 1	…	Topic K
Word 1	13	…	1
Word 2	27	…	3
… …	… …	…	… …
Word V	20	…	9

At round t,

Document m and Topic k: $\hat{\theta}_{mk} = \frac{\alpha_k + n_{m,-i}^{(k)}}{\sum_{k=1}^{K}(\alpha_k + n_{m,-i}^{(k)})}$

Topic k and Word t: $\hat{\varphi}_{kt} = \frac{\beta_t + n_{k,-i}^{(t)}}{\sum_{t=1}^{V}(\beta_t + n_{k,-i}^{(t)})}$

Topic Models: Bayesian Updating

Word-Doc	Doc 1	…	Doc M
Position 1	Topic 1	… …	Topic 4
Position 1	Topic 1	… …	Topic 4
… …	… …	… …	… …
Position N_m	Topic 3	1	Topic 5

Doc-Topic	Doc 1	…	Doc M
Topic 1	62	…	17
Topic 2	27	…	85
… …	… …	…	… …
Topic K	11	…	22

Topic-Word	Topic 1	…	Topic K
Word 1	13	…	1
Word 2	27	…	3
… …	… …	…	… …
Word V	20	…	9

At round t+1,

For i (document m and position n): $p(z_i = k \mid z_{-i}, w) \propto \hat{\theta}_{mk} \times \hat{\varphi}_{kt}$

Alternating and Iterating these procedure, Finally will converge

Topic Models: Prediction

- Predict $\theta_{new,k}$=p(topic k|d_{new}):
  - Remember our posterior distribution: $p(z, \theta, \varphi | \alpha, \beta, w)$
  - Within this new document, Bayesian updating (Gibbs Sampling) on only θ and z_{new}
  - Other latent variables values φ are not changed (they have been obtained in training process)
  - Obtain the new posterior mean values of θ_{mk}, check the values of this new document (i.e., m=new)

Topic Models: Limitations

- Bag-of-words Model:
  - Focus on the frequency of words
  - Ignore the order of words

- Example:
  - Document 1: I like her
  - Document 2: She likes me
  - Bag-of-words: {I, like, She}
  - Do you think these 2 documents have the same meaning ?

Topic Models in R

- Online References:
  - https://www.tidytextmining.com/topicmodeling.html

6.1 Latent Dirichlet allocation

Latent Dirichlet allocation is one of the most common algorithms for topic modeling. Without diving into the math behind the model, we can understand it as being guided by two principles.

- **Every document is a mixture of topics.** We imagine that each document may contain words from several topics in particular proportions. For example, in a two-topic model we could say "Document 1 is 90% topic A and 10% topic B, while Document 2 is 30% topic A and 70% topic B."
- **Every topic is a mixture of words.** For example, we could imagine a two-topic model of American news, with one topic for "politics" and one for "entertainment." The most common words in the politics topic might be "President", "Congress", and "government", while the entertainment topic may be made up of words such as "movies", "television", and "actor". Importantly, words can be shared between topics; a word like "budget" might appear in both equally.

Thank You !