

Lecture 12:

Text Analysis:

Latent Dirichlet Allocation


Jaeki Song, Ph.D.

Latent Dirichlet Allocation (LDA)

"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	MANGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Assumption:

- One document is composed of topics
- One topic is composed of terms



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.

LDA Assumption

- “bag-of-words” – exchangeability, not i.i.d
- **Dirichlet** distribution
 - A k -dimensional Dirichlet random variable θ can take values in the $(k-1)$ -simplex, and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$$



Notation and terminology

- A *word* is an item from a vocabulary indexed by $\{1, \dots, V\}$. We represent words using unit-basis vectors. The v^{th} word is represented by a V -vector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$

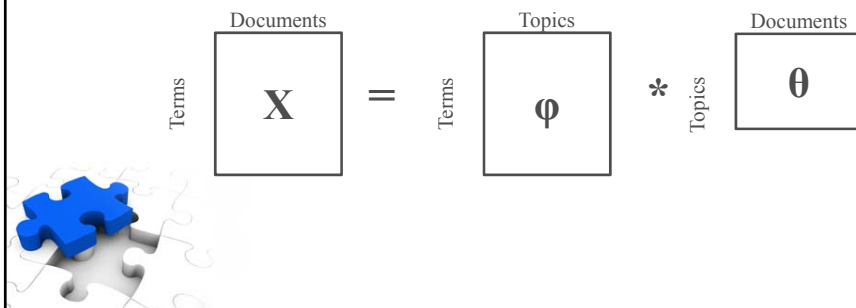
$$w = [0 \dots 0 \underset{\substack{v^{th} \\ V-dim}}{1} 0 \dots 0]^T$$

- A *document* is a sequence of N words denoted by $W = (w_1, w_2, \dots, w_n)$, where w_n is the n^{th} word in the sequence
- A *corpus* is a collection of M documents denoted by $D = [D_1, D_2, \dots, D_m]$

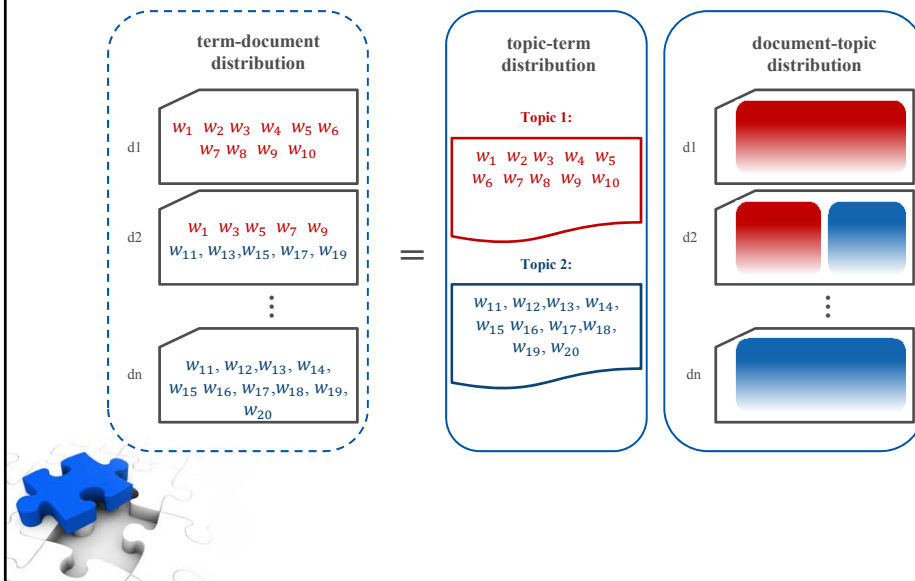


Latent Dirichlet Allocation (LDA)

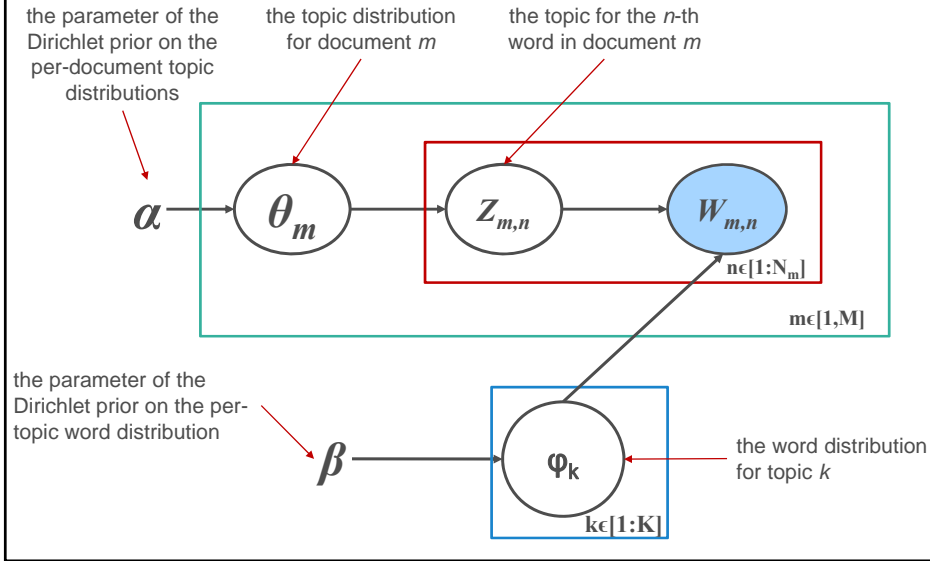
- LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words.



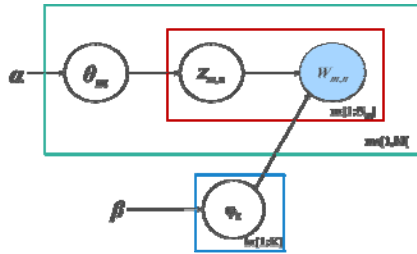
Latent Dirichlet Allocation (LDA)



The Graphical Model



Generative Process



M, N, V, k	fixed known parameters
α, β	fixed unknown parameters
θ, z, w	random variables (w are observable)

Generative process for each document W in a corpus D :

1. Choose $\theta_m \sim \text{Dirichlet}(\alpha)$, $m \in \{1, \dots, M\}$
2. Choose $\phi_k \sim \text{Dirichlet}(\beta)$, $k \in \{1, \dots, K\}$; ϕ_k are V -dimensional vectors
3. For each of the word positions m, n , $n \in \{1, \dots, N_m\}$ and $m \in \{1, \dots, M\}$
 - (a) Choose a topic $z_{m,n} \sim \text{Multinomial}(\theta_m)$
 - (b) Choose a word $w_{m,n} \sim \text{Multinomial}(\phi_{z_{m,n}})$

θ is a document-level variable, z and w are word-level variables.

Document Modeling

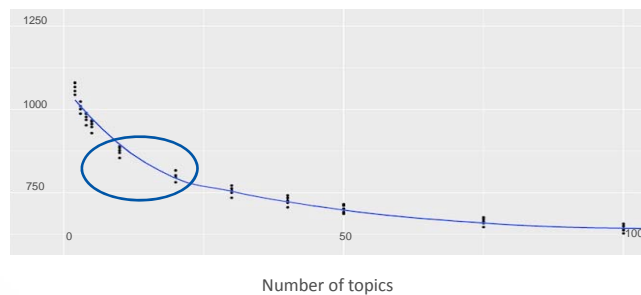
- Unlabeled data – our goal is density estimation.
- Compute the *perplexity* of a held-out test to evaluate the models – lower perplexity score indicates better generalization.

$$\text{perplexity}(D_{\text{test}}) = \exp \left\{ - \frac{\sum_{d=1}^M \log p(\mathbf{w}_d)}{\sum_{d=1}^M N_d} \right\}$$



Document Modeling - Example

- Based on the perplexity graph, the optimal number of topics appears in the range from 5 to 15.



Example

- Suppose you have the following set of sentences:
 - I **eat fish** and **vegetables**.
 - **Fish** are **pets**.
 - My **kitten** **eats fish**.
- LDA might classify the **red** words under a topic F (food) and **blue** words under a topic P (pets).



Example

- There are 2 benefits from LDA defining topics on a word-level:
 - 1) Infer the content spread of each sentence by a word count:
 - ✓ Sentence 1: 100% Topic F
 - ✓ Sentence 2: 100% Topic P
 - ✓ Sentence 3: 33% Topic P and 67% Topic F
 - 2) Derive the proportions that each word constitutes in given topics:
 - ✓ Topic F might comprise words in the following proportions: 40% eat, 40% fish, 20% vegetables.



Example

- LDA achieves the above results in 3 steps. Imagine you have 2 documents with the following words:

Document X		Document Y	
	Fish		Fish
	Fish		Fish
	Eat		Milk
	Eat		Kitten
	Vegetables		Kitten

1. You tell the algorithm how many topics you think there are
2. The algorithm will assign every word to a temporary topic



Example

3. The algorithm will check and update topic assignments
 - 1) How prevalent is that word across topics?

Document X		Document Y	
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten



Example

2) How prevalent are topics in the document?

	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten



LSA vs. LDA

Latent Semantic Indexing (LSA)	Latent Dirichlet Allocation (LDA)
Term-Document Frequency	Bayesian Analysis
Singular Value Decomposition (SVD)	Generative Probabilistic Model
Dimension Reduction	Likelihood Principle
Unsatisfactory Statistical Foundation	Strong Statistical Foundation

