(Supplement)
What is LDA in Topic Modeling?

- Machine Learning is a field of computer science that uses statistical techniques that enables computer systems to learn certain latent patterns within data.
- If the computer systems learn the hidden pattern well, those can make the accurate prediction on data.
- In machine learning, a topic model is a type of statistical model for discovering the hidden "topics" that are described in a set of documents.
- Machine learning tasks are basically divided into two types, supervised and unsupervised learning.
- *LDA used in this paper works in unsupervised learning.*

(Supplement)
Unsupervised Learning in Natural Language Processing

- An example of spam mail clustering using unsupervised learning in natural language processing.

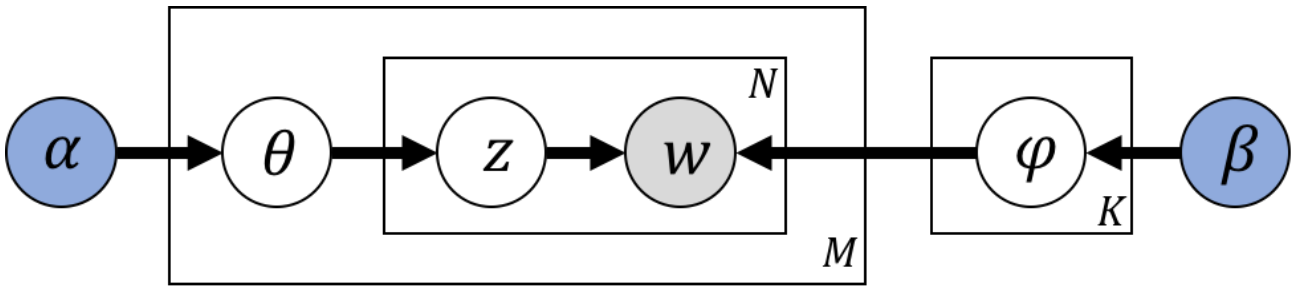
Input Data(X)	Label(Y)	Cluster
Up to 30% off	N/A	A
Coupons & Free Shipping	N/A	A
H Department store - Opening Discounts, Coupons	N/A	A
New arrival 2018	N/A	A
Today's schedule	N/A	B
The progress of the project	N/A	B
The best contract	N/A	B
New meeting	N/A	B
No schedule today	N/A	B
Clearance(Free Shipping)	N/A	?

- If it has learned well, the computer system will divide the emails into two clusters, say A and B.
- Then, by carefully looking the features of elements in each clusters, *Cluster A* can be named as *Spam* and *Cluster B* as *Importance*.
- And, the computer system estimates *"Clearance(Free Shipping)"* as *Spam*.

(Supplement)
Latent Dirichlet Allocation

- LDA working in unsupervised learning regards the writing process as the generative process that a writer
  1. specifies topics to write about, choosing a distribution over topic.
  2. randomly picks a topic over the topic distribution.
  3. randomly chooses a word in relation to the topic from the distribution over the vocabulary.
  4. goes back to step 2 and repeat.

Figure: A Graphical Model of a LDA Bayesian Network



(Supplement)
Latent Dirichlet Allocation

- LDA working in unsupervised learning regards the writing process as the generative process that a writer
  1. specifies topics to write about, choosing a distribution over topic.
  2. randomly picks a topic over the topic distribution.
  3. randomly chooses a word in relation to the topic from the distribution over the vocabulary.
  4. goes back to step 2 and repeat.
- Thus, we can simply say that the writing process is based on the distributions of topic and vocabulary.
- Since we have words used in a document, this will be a bayesian problem to find out posterior distributions of topic and vocabulary given the words.

(Supplement)
Latent Dirichlet Allocation

- Let's say topic distribution as θ, vocabulary distribution as φ, and words as W.
- Since a topic of multiple topics is selected, θ follows multinomial distribution.
- In the same way, since a word of multiple words is selected, φ also follows multinomial distribution.
- We want to know $f(\theta, \varphi | W)$, but it is hard to estimate both θ and φ given only W.
- Thus, LDA introduces topic allocation, Z, meaning that a word within a document is used for describing a topic.
- Then, the bayesian problem will be changed into the following :

$$f(\theta, \varphi | W) \Longrightarrow f(\theta, \varphi | Z, W)$$

(Supplement)
Latent Dirichlet Allocation

- Now, let's describe how to solve the bayesian problem intuitively.
- The bayesian problem, $f(\theta, \varphi | Z, W)$, can be solved with its factorization like the following :

  i) $f(Z|W, \theta, \varphi)$: Given W, θ, and φ, this problem is trivial.
  ii) $f(\theta, \varphi | Z, W)$: Given Z and W, this problem is easier than $f(\theta, \varphi | W)$.
- Now, we want to know θ, φ, and Z.
- Since Z, θ, φ follows multinomial distributions, prior distributions are assumed to follow dirichlet distributions with hyperparameters, α, β.
- This paper introduces collapsed gibb sampling algorithms, integrating $f(Z|W, \theta, \varphi)$ of the equation i) with respect to θ and φ, such that

  i) $f(Z|W, \theta, \varphi) \Longrightarrow f(Z|W)$
  ii) $f(\theta, \varphi | Z, W)$: Given Z and W, this problem is easier than $f(\theta, \varphi | W)$.

(Supplement)
Latent Dirichlet Allocation

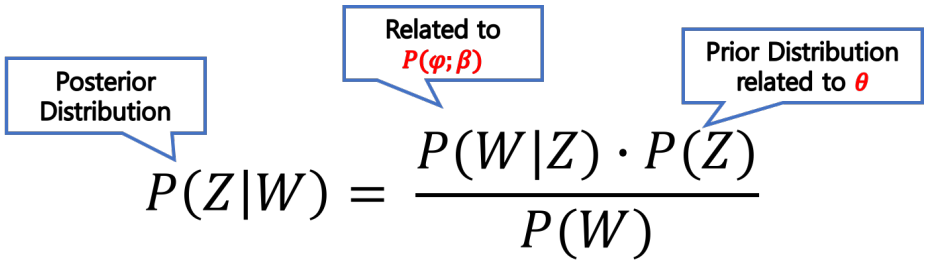


Figure: Basic Bayesian Equation of LDA

- Pseudo code to solve the bayesian problem will the following :

Initialize Z
Count $n^i_{j,r}$, the number of words assigned to topic i in document j from the Z
While Perplexity(performance measure) converges:
  For i in the # of Documents:
    For l in the # of words:
      Sample k from $f(Z_{m,l}|Z\_m, l, W)$
      Update $n^i_{j,r}$ by Assigning $Z_{m,l} = k$ and Update Z
Calculate θ, φ given the result of Z and W

(Supplement)
Latent Dirichlet Allocation

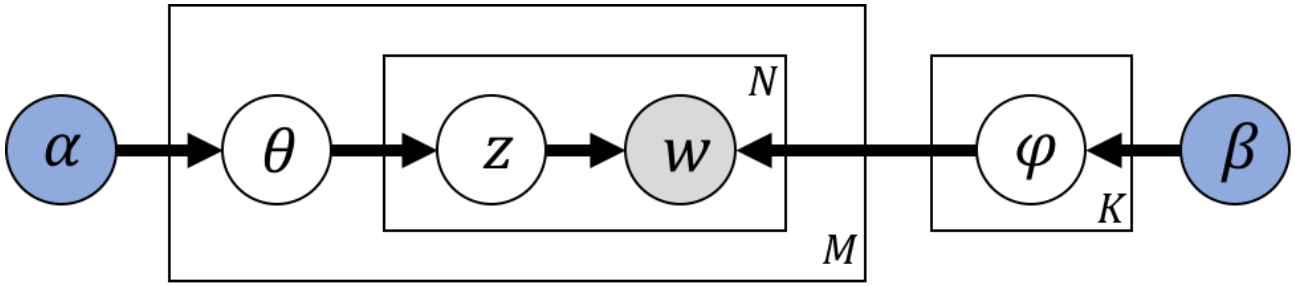


Figure: A Graphical Model of a LDA Bayesian Network

- $\theta_i \sim Dir(\alpha), i \in \{1, \ldots, M\}$: Topic distribution of a document, i.
- $\varphi_k \sim Dir(\beta), k \in \{1, \ldots, K\}$: Word distribution of a topic, k.
- $z_{i,l} \sim Mult(\theta_i), i \in \{1, \ldots, M\}, l \in \{1, \ldots, N\}$: Topic allocation for a word, l, in a document, i.
- $w_{i,l} \sim Mult(\varphi_{z_{i,l}}), i \in \{1, \ldots, M\}, l \in \{1, \ldots, N\}$: a word generated from the distribution of $\varphi_{z_{i,l}}$, word-topic distributions.
- α, β, and K are hyperparameters.

Results of Topic Modeling

Term	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	*Breakfast*	*Metro*	*Check*	One	*Great*
2	*Free*	*Walk*	*Night*	*Like*	*Staff*
3	*Food*	*Locat*	*Time*	*Price*	Servic
4	*Bed*	*Station*	*Day*	Can	*Excel*
5	Floor	Clean	Back	Time	Locat
6	Bar	Minut	Even	Place	Nice
7	*Wifi*	Close	Book	Will	Red
8	Small	Also	Ask	Much	Best
9	Bathroom	Citi	*Front*	Just	*View*
10	Night	*Restaur*	*Desk*	*Better*	*Recommend*
11	Well	Nice	One	Busi	Location
12	Also	Staff	Servic	*Standard*	Well
13	Restaur	Airport	First	Bit	Help
14	Area	Street	*Arriv*	Star	Perfect
15	Larg	English	Recept	*Get*	Visit
16	Water	Train	Got	Mani	*High*
17	Buffet	Can	Get	Realli	Veri
18	Clean	Near	Taxi	Need	Kremlin
19	Shower	Away	Just	Quit	Realli
20	Use	Just	Made	Expect	Square
21	Includ	Center	Will	*Internet*	Breakfast
22	Nice	Busi	Hour	Lobbi	Love
23	Offer	Shop	Make	Want	Friend
24	Etc	Get	Call	Old	Trip
25	Day	Friend	Never	*Look*	Enjoy
26	English	Around	Russian	Thing	Spacious
27	Comfort	Min	Guest	Feel	Service
28	Drink	Speak	Next	Littl	Definit
29	Smoke	Conveni	Way	Big	Beauti
30	Coffe	Red	Took	Lot	Place

- These results of LDA come from *Mankad, Shawn, et al. "Understanding Online Hotel Reviews Through Automated Text Analysis."* Service Science (2016).
- By carefully looking at the words and considering hotel operations, the authors named ;
  - Topic 1 as *Amenities*
  - Topic 2 as *Location*
  - Topic 3 as *Transactions*
  - Topic 4 as *Value*
  - Topic 5 as *Experience*