

1) Latent Dirichlet Allocation is a probabilistic used to generate the topics. LDA is the iterative model which requires 3 parameters, which are number of topics and deep a prior, knowledge of the dataset.

We evaluate the LDA performance using perplexity to evaluate the LDA model. One document is taken and split in two. The first half is fed in LDA to compute the topics composition, from that composition, then, the word distribution is estimated.

This distribution is then compared with the word distribution is estimated. This distribution is then compared with the measure of distance is extracted. Perplexity is often used to select the best number of topics of the LDA model.

LDA

Input: words in document d

Output: topic assignment Z and counts $n_{d,k}$, $n_{k,w}$ and n_k

begin
randomly initialized Z and increment Countless for each iteration do

for $i = 0 \rightarrow N-1$ do

word $\leftarrow w[i]$

topic $\leftarrow z[i]$

$n_{d, \text{topic}} = 1$; $n_{\text{word}, \text{topic}} = 1$; $n_{\text{topic}} = 1$

for $k = 0 \rightarrow K-1$ do

$$p(Z=k) = \frac{n_{d,k} + \alpha}{n_k + 13 \times w}$$

End

topic \leftarrow Sample from $p(Z)$

$z[i] \leftarrow \text{topic}$

$n_{d, \text{topic}} = 1$; $n_{\text{word}, \text{topic}} = 1$; $n_{\text{topic}} = 1$

End

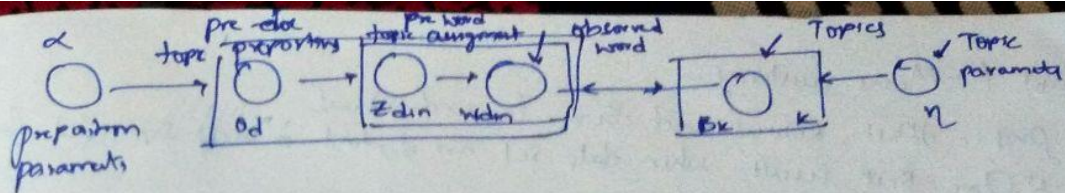
End

return $n_{d,k}$, $n_{k,w}$, n_k

1. decide how many topics you need

2. The algorithm will assign every word to a temporary word

3. will check and update topic

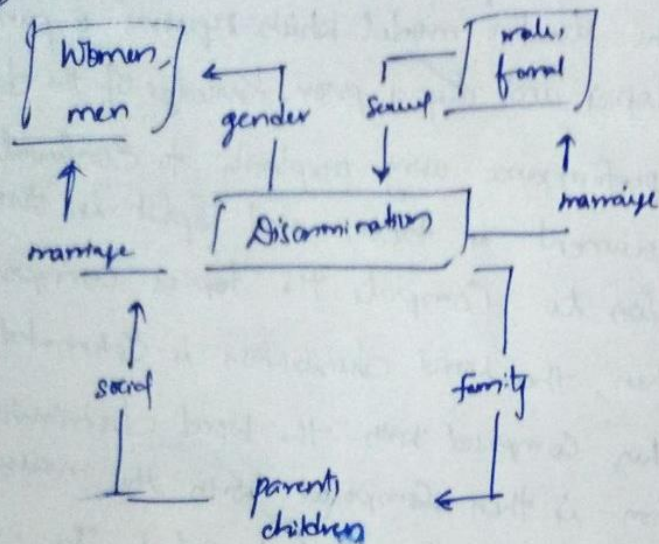


$$P(\beta, \theta, z, w) = \left(\prod_{t=1}^k P(\beta, l_t) \right) \left(\prod_{d=1}^D P(z_{dn} / \theta_d) \right) P(w_{d,n} / \beta; k, z_{d,n})$$

2. a) $K=3$ Centen D_1, D_5, D_7

$$\begin{aligned} D_1 \text{ to } D_2 &= \sqrt{4} = 2 & D_3 \text{ to } D_2 &= \sqrt{6} = 2.4 \\ D_1 \text{ to } D_5 &= \sqrt{7} = 2.6 & D_3 \text{ to } D_5 &= \sqrt{3} = 1.7 \\ D_1 \text{ to } D_7 &= \sqrt{5} = 2.2 & D_3 \text{ to } D_7 &= \sqrt{5} = 2.2 \\ D_2 \text{ to } D_5 &= \sqrt{7} = 2.6 & D_5 \text{ to } D_7 &= \sqrt{2} = 1.4 \\ D_4 \text{ to } D_2 &= \sqrt{8} = 2.8 & D_5 \text{ to } D_2 &= 0 \\ D_4 \text{ to } D_5 &= \sqrt{9} = 3 & D_5 \text{ to } D_7 &= \sqrt{8} = 2.8 \\ D_7 \text{ to } D_7 &= \sqrt{3} = 1.7 & D_7 \text{ to } D_2 &= 2.2 \\ D_7 \text{ to } D_5 &= \sqrt{8} = 2.8 & D_7 \text{ to } D_5 &= 3.4 \\ D_7 \text{ to } D_2 &= \sqrt{7} = 2.6 & D_9 \text{ to } D_7 &= 3.2 \\ D_8 \text{ to } D_5 &= \sqrt{6} = 2.4 & D_{10} \text{ to } D_2 &= 2.4 \\ D_8 \text{ to } D_7 &= \sqrt{5} = 2.2 & D_{10} \text{ to } D_5 &= 2.2 \end{aligned}$$

Doc	D_2	D_5	D_7	Min	Chunks
D_1	2.0	2.6	2.2	2.0	D_2
D_2	0	2.6	2.8	0	D_2
D_3	2.4	3.6	2.2	2.2	D_7
D_4	2.8	3.0	2.6	2.6	D_7
D_5	2.6	0	2.8	0	D_5
D_6	2.4	3.9	2.6	2.4	D_7
D_7	1.7	2.8	0	0	D_5
D_8	2.6	2.0	2.8	2.0	D_2
D_9	2.0	3.0	3.6	2.2	D_2
D_{10}	2.2	3.5	2.4		



(1c) ~~Since~~ Since the words in the document y are assigned to Topic C and Topic P in a 50-50 ratio, the remaining 'fish' word seems equally likely to be about either topic

	Doc X		Doc Y
F	Ram	?	Fish
F	Fish	F	Ram
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetably	P	Kitten

(1d)

- 1) Each topic is distribution over words.
- 2) Each document is a mixture of corpus-wide topics.
- 3) Each word is drawn from one of those topics.
- 4) We only observe the documents.
- 5) The other stuff are hidden variables.
- 6) Our goal is to infer the hidden values i.e. compute their distribution conditional on the document.
- 7) Encode assumption
- 8) Before factorization of the joint distribution

2.b) K-Means clustering:-

- Pros:-
- 1) Fast, Robust and Easy to understand
 - 2) Give Best result when data set are distinct or well separated from each other.
 - 3) It is a great solution from pre-clustering.
 - 4) Works great from spherical clusters.

LDA TOPIC DISCOVERY Model.

Pros: We can infer context spread of each sentence by a word count.

② We can derive the properties that each word constitutes in given topics.

Cons:

- ① We have to specify the number of topics.
- 2) LDA's efficiency is pretty low when compared to m.l. Algo.
 - 3) LDA cannot compare correlation.
 - 4) Unsupervised
- ⑤ uses Bow (word are Exchangeable)