LDA

① ② Latent Dirichlet Allocation a probistic is used to generate the topics. LDA is the iterative model which requires 3 parameters, which are number of topics and deep a priori knowledge of the dataset.

To evaluate a LDA model, one document is taken and split in two. The first half is fed to LDA to compute the topics Compisitions from that composition then word distribution is estimated. This distribution is then computed with word distribution of 2nd half of document. A measure of distance is extracted.

LDA Algorithm

Input : words $w \in$ documents d

output : topic assignments z and counts $n_{d,k}$ $n_{k,w}$ and n_k

begin.

Randomly Initialize z and Increment counter

for each iteration do

for $i \leftarrow 0$ $N-1$ do

word $\leftarrow$ wo[i]

topic $\leftarrow$ z[i]

nd, topic - =1, nword, topic - =1, ntopic- =1.

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

$$p(z=k \mid \cdot) = (n_{d,k} + \alpha_k) \frac{n_{k,w} + \beta_w}{n_k + \beta \times w}$$

end

topic ← sample from p(z|)

z[i] ← topic

nd, topic += 1 , nword, topic+1 = 1 , ntopic += 1

end

end

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

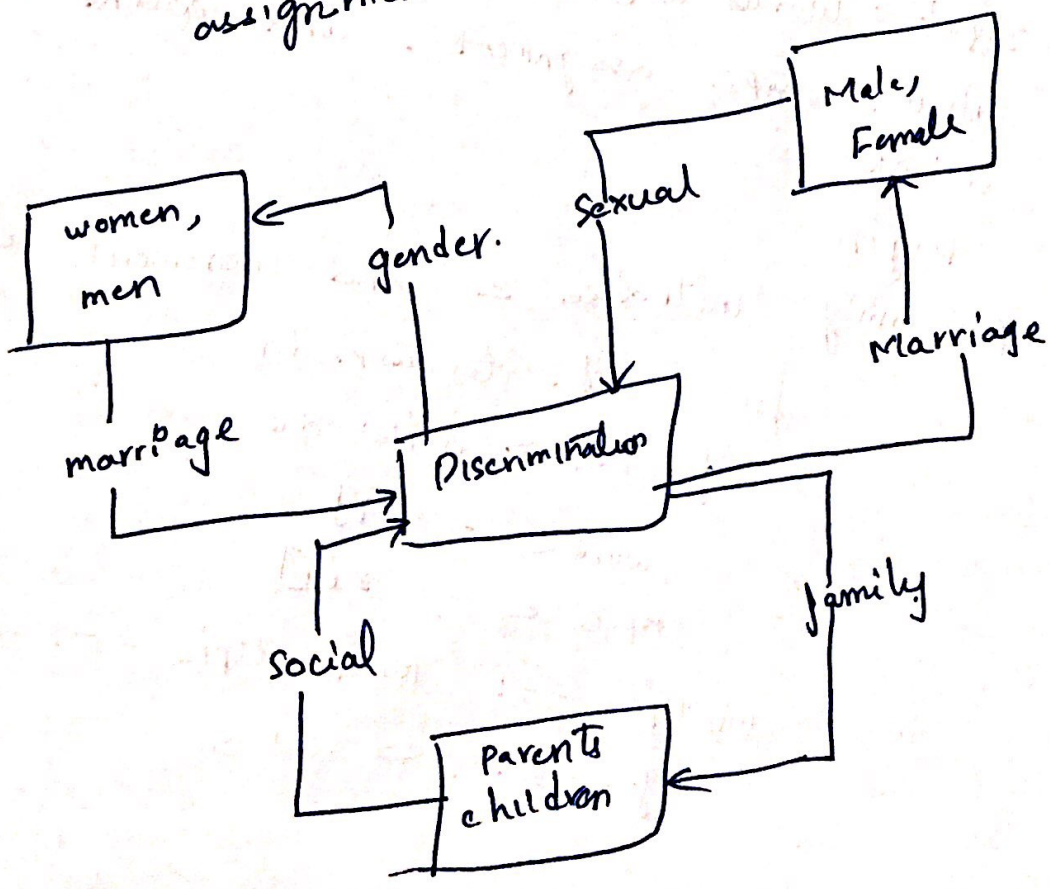
end

step1 : Decide how many topics we need

step2 The algorithm will assign every word to a temporary topic

step3 - The algorithm will check or update topic assignment

(1b)



(1c) How prevalent are topics in the document?

Since the words in Doc Y are assigned to Topic F and Topic P in a 50-50 ratio. the remaining "fish", words seems equally likely to be about either topic.

	Doc X		Doc Y
F	Fish	2	Fish.
F	Fish	F	fish.
F	cat	F	Milk.
F	cat	P	Kitten
F	vegetable	P	Kitten

(1d). 1) each topic is a distribution over words

2) each document is a mixture of corpus wide topics

3) each word is drawn from one of these topics

4) we only observe the documents

5) The other structure are hidden variable

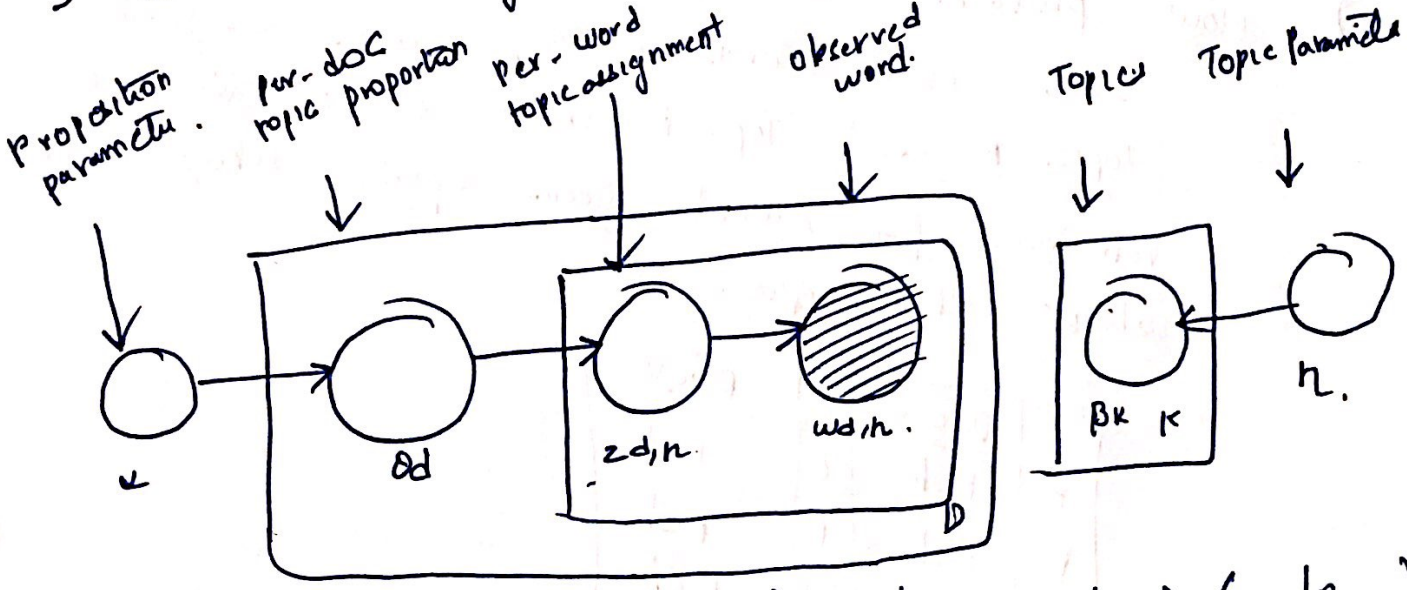
6) our goal is to infer the hidden. Variable ie. compute their distribution. Conditioned on documents.

P(topics, proportions, assignments | documents)

7) Encode assumption

8) Define a factorization of the joint distribution

a) connect to algorithm to compute with date



$$P(\beta, \theta, z, w) = \left(\prod_{i=1}^{K} P(\beta_i | \eta)\right)\left(\prod_{d=1}^{N} P(z_{d,n}|\theta_d) P(w_{d,n}|\beta_k, z_{d,n})\right)$$

(2)(a) we have to create K = 3 clusters

Lets choose D2, D5 & D7 as initial three seeds

Now we have to calculate euclidean distance from other documents D2, D5 & D7

O → online, F → Festival, B → Book, F → Flight, D → Delhi

$$D_1 \text{ to } P_2 = \sqrt{(O_1-O_2)^2 + (F_1-F_2)^2 + (B_1-B_2)^2 + (T_1-T_2)^2 + (P_1-P_2)^2}$$

$$= \sqrt{(1-2)^2 + (0-1)^2 + (1-2)^2 + (1-1)^2}$$

$$= \sqrt{4} = 2.$$

$$D_1 \text{ to } P_5 = \sqrt{(1-3)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2}$$

$$= \sqrt{7} = 2.6.$$

D_1 to D_7 $= \sqrt{(1-2)^2 + (0-0)^2 + (1-1)^2 + (0-2)^2 + (1-1)^2}$

$= \sqrt{5} = 2.2$

D_2 to P_2 $= 0$

D_2 to D_5 $= \sqrt{(2-3)^2 + (1-1)^2 + (2-0)^2 + (1-0)^2 + (1-0)^2} = \sqrt{7}, 2.6$

D_2 to D_7 $= \sqrt{(2-2)^2 + (1-0)^2 + (2-1)^2 + (1-2)^2 + (1-0)^2} = \sqrt{3} = 1.7$

D_3 to D_2 $= \sqrt{6} = 2.4$

D_3 to D_5 $= \sqrt{13} = 3.6$

D_3 to D_7 $= \sqrt{5} = 2.2$

D_5 to D_2 $= \sqrt{7} = 2.6$

D_5 to D_5 $= 0$

D_5 to D_7 $= \sqrt{8} = 2.8$

D_8 to D_2 $= \sqrt{4} = 2$

D_8 to D_5 $= \sqrt{9} = 3$

D_8 to D_7 $= \sqrt{7} = 2.6$

D_4 to D_2 $= \sqrt{8} = 2.8$

D_4 to D_5 $= \sqrt{9} = 3$

D_4 to D_7 $= \sqrt{7} = 2.6$

D_6 to D_2 $= \sqrt{6} = 2.4$

D_6 to D_5 $= \sqrt{15} = 3.8$

D_6 to D_7 $= \sqrt{7} = 2.6$

D_4 to $D_4 = 0$

D_7 to D_2 $\sqrt{3} = 1.7$

D_7 to D_5 $= \sqrt{8} = 2.8$

D_1 to D_2 $= \sqrt{6} = 2.4$

D_8 to D_5 $= \sqrt{5} = 2.2$

D_8 to D_7 $= \sqrt{5} = 2.2$

D_{10} to D_2 $= \sqrt{5} = 2.2$

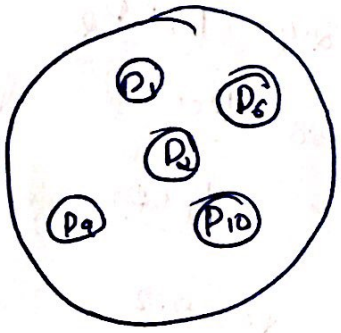
D_{10} to D_5 $= \sqrt{12} = 3.4$

D_{10} to D_7 $= \sqrt{6} = 2.4$

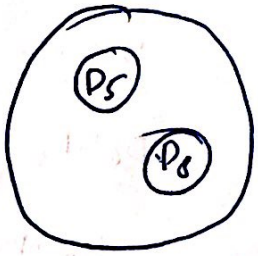
Documents	D_2	D_5	D_7	Minds	cluster
D_1	2.0	2.6	2.2	2.0	D_2
D_2	0.0	2.6	1.7	0.0	D_2
P_3	2.4	3.6	2.2	2.2	D_7
P_4	2.8	3.0	2.6	2.6	D_2

P_6	2.6	0.0	2.8	0.0	P_5
P_4	2.4	3.9	2.6	2.4	P_2
P_7	1.7	2.8	0.0	0.0	P_7
P_8	2.6	2.0	2.8	2.0	P_5
P_9	2.0	3.0	3.6	2.0	P_2
P_{10}	2.2	3.5	2.4	2.2	P_2

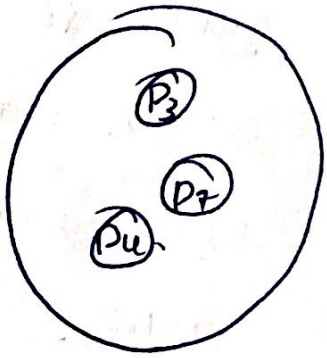
P_2 cluster



P_5 cluster



P_7 cluster



(2)(b) K Means clustering

Pros

(1) Fast, robus and easier to understand → Computational cost - O(K*n*d)

(2) gives Best result when data set are distinct or well seperated from each other.

(3) It is a great solution for pre-clustering

(4) works great for spherical clusters

Cons

① K-value is not known and is difficult to predict

② There is no unique solution for a certain value since initial partitions can be different

③ Does not work well with clusters of different size and different density

LDA Topic Discovery model

Pros

① We can refer the content spread of each sentence by a word count

② We can derive the proportions that each word constitute in given topic

Cons

① We have to specify number of topics

② LDA's efficiency is pretty low when compared to machine learning to algorithm

③ LDA cannot capture correlations

④ unsupervised (sometimes we need supervision. Sentiment analysis)

⑤ Uses BOW (assumes words are exchangable)