1.a. Latent Dirichlet Allocation a probabilistic 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.

We evaluate performance of the LDA using perplexity. To evaluate the LDA model, one document is taken and split in two. The first half is fed into LDA to compute the topics composition, from that composition, then, the word distribution is estimated. This distribution is then compared with the word distribution of the 2nd half of the document. A measure of distance is extracted. Perplexity is often used to select the best number of topics of the LDA model.

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 counters

foreach iteration do

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

word $\leftarrow w[i]$

topic $\leftarrow z[i]$

$n_{d,topic} -= 1$; $n_{word,topic} -= 1$; $n_{topic} = 1$

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

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

end

topic $\leftarrow$ sample from $p(z | \cdot)$

$z[i] \leftarrow$ topic.

$n_{d,topic} += 1$; $n_{word,topic} += 1$; $n_{topic} += 1$

end

end

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

end.

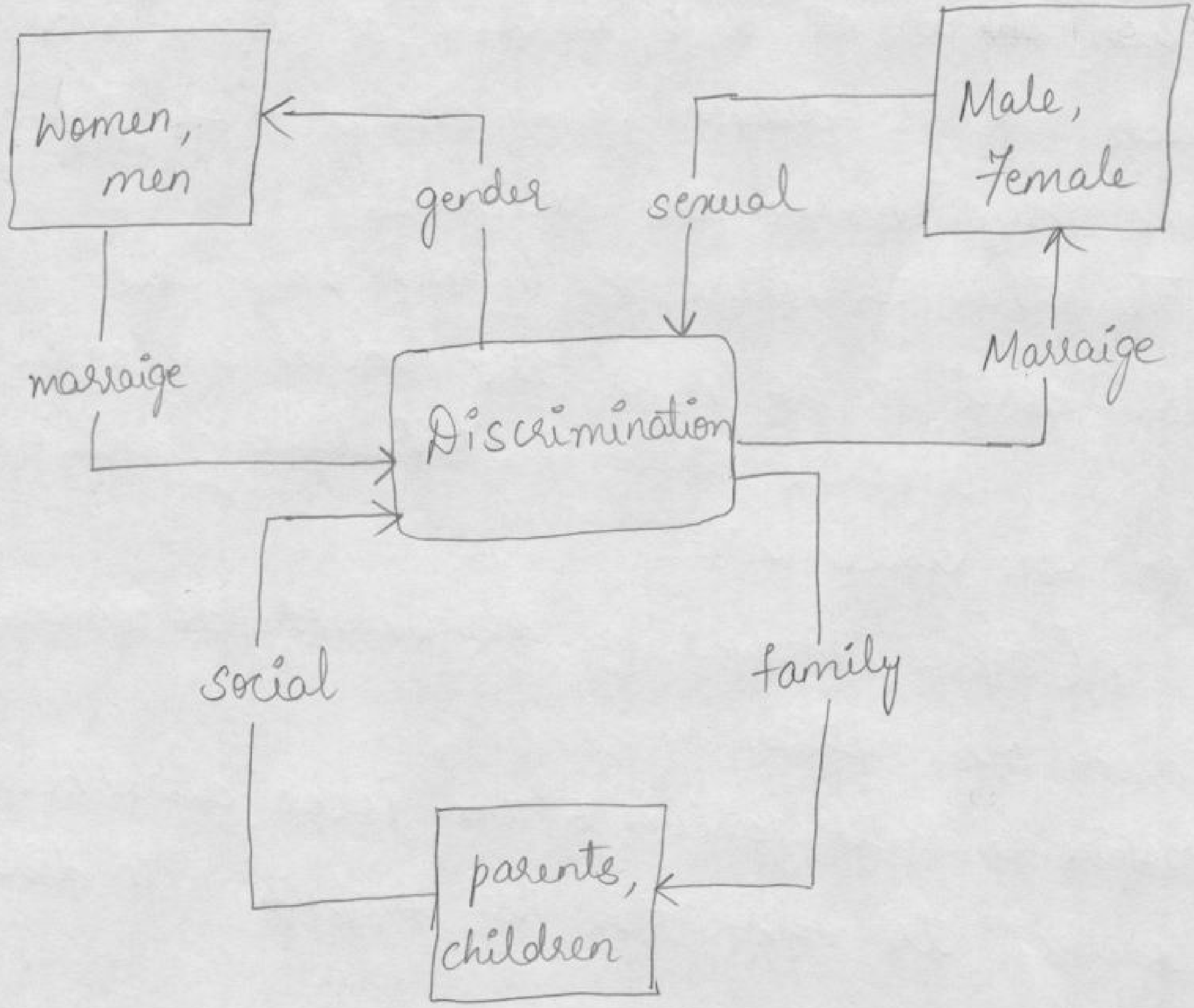
Step 1: Decide how many topics we need.

Step 2: The algorithm will assign every word to a temporary topic.

Step 3: The algorithm will check and update topic assignments.

1b)

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" word seems equally likely to be about either topic.

	Doc X		Doc Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	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 those topics.

4) We only observe the documents

5) The other structure are hidden variables

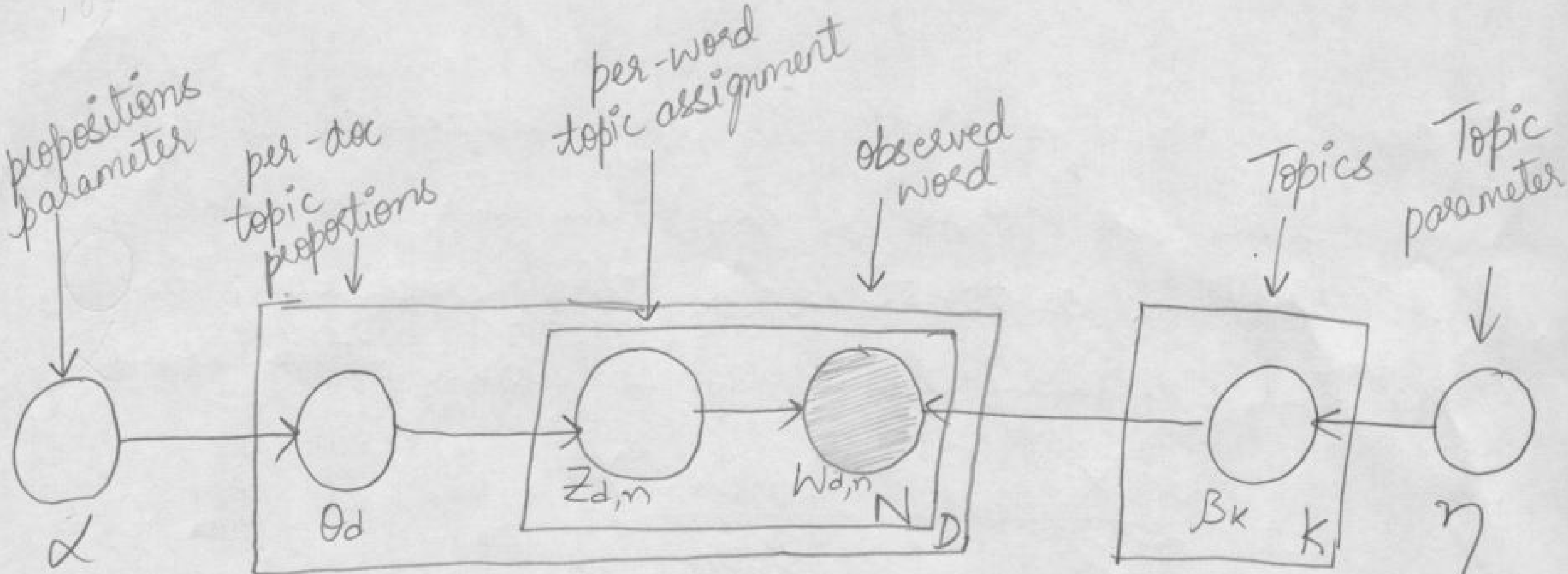
6) Our goal is to infer the hidden variables. i.e., compute their distribution conditioned on the documents

$$p(\text{topics}, \text{proportions}, \text{assignments} \mid \text{documents})$$

7) Encode assumption

8) Define a factorization of the joint distribution

9) Connect to algorithm to compute with data.



$$p(\beta, \theta, z, w) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta)\right)\left(\prod_{d=1}^{D} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid \beta_{1:K}, z_{d,n})\right)$$

2. a) We have to create $K=3$ clusters.

Let's choose D2, D5 and D7 as initial three seeds.

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

O → Online F → Festival B → Book T → Flight D → Delhi.

$\underline{D_1 \text{ to } D_2} = \sqrt{(O_1-O_2)^2+(F_1-F_2)^2+(B_1-B_2)^2+(T_1-T_2)^2+(D_1-D_2)^2}$

$= \sqrt{(1-2)^2+(0-1)^2+(1-2)^2+(0-1)^2+(1-1)^2} = \sqrt{4} = \underline{\underline{2}}$

$\underline{D_1 \text{ to } D5} = \sqrt{(1-3)^2+(0-1)^2+(1-0)^2+(0-0)^2+(1-0)^2} = \sqrt{7} = \underline{\underline{2.6}}$

$\underline{D_1 \text{ to } D7} = \sqrt{(1-2)^2+(0-0)^2+(1-1)^2+(0-2)^2+(1-1)^2} = \sqrt{5} = \underline{\underline{2.2}}$

$\underline{D_2 \text{ to } D_2} = 0$ $\underline{D_2 \text{ to } D5} = \sqrt{(2-3)^2+(1-1)^2+(2-0)^2+(1-0)^2+(1-0)^2} = \sqrt{7} = \underline{\underline{2.6}}$

$\underline{D_2 \text{ to } D7} = \sqrt{(2-2)^2+(1-0)^2+(2-1)^2+(1-2)^2+(1-1)^2} = \sqrt{3} = \underline{\underline{1.7}}$

$\underline{D_3 \text{ to } D2} = \sqrt{6} = \underline{\underline{2.4}}$ $\underline{D_4 \text{ to } D_2} = \sqrt{8} = \underline{\underline{2.8}}$ $D_7 \text{ to } D_7 = 0$

$\underline{D_3 \text{ to } D5} = \sqrt{13} = \underline{\underline{3.6}}$ $\underline{D_4 \text{ to } D5} = \sqrt{9} = \underline{\underline{3}}$ $\underline{D_7 \text{ to } D2} = \sqrt{3} = \underline{\underline{1.7}}$

$\underline{D_3 \text{ to } D7} = \sqrt{5} = \underline{\underline{2.2}}$ $\underline{D_4 \text{ to } D7} = \sqrt{7} = \underline{\underline{2.6}}$ $\underline{D_7 \text{ to } D_5} = \sqrt{8} = \underline{\underline{2.8}}$

$\underline{D_5 \text{ to } D_2} = \sqrt{7} = \underline{\underline{2.6}}$ $\underline{D_6 \text{ to } D2} = \sqrt{6} = \underline{\underline{2.4}}$ $\underline{D_8 \text{ to } D_2} = \sqrt{6} = \underline{\underline{2.4}}$

$\underline{D_5 \text{ to } D_5} = 0$ $\underline{D_6 \text{ to } D5} = \sqrt{15} = \underline{\underline{3.8}}$ $\underline{D_8 \text{ to } D5} = \sqrt{5} = \underline{\underline{2.2}}$

$\underline{D_5 \text{ to } D_7} = \sqrt{8} = \underline{\underline{2.8}}$ $\underline{D6 \text{ to } D7} = \sqrt{7} = \underline{\underline{2.6}}$ $\underline{D_8 \text{ to } D7} = \sqrt{5} = 2.2$

D_9 to $D_2 = \sqrt{4} = \underline{\underline{2}}$

D_9 to $D_5 = \sqrt{9} = \underline{\underline{3}}$

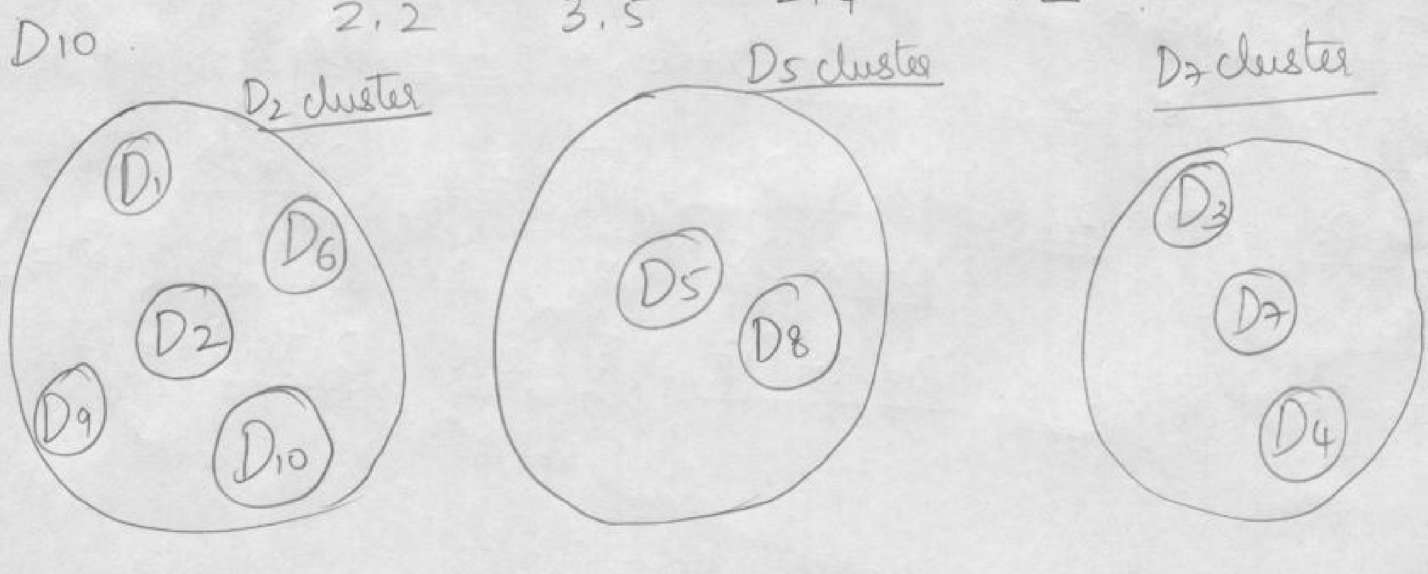
D_9 to $D_7 = \sqrt{7} = \underline{\underline{2.6}}$

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

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

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

Documents	D_2	D_5	D_7	Min dis	Cluster
D1	2.0	2.6	2.2	2.0	D_2
D2	0.0	2.6	1.7	0.0	D_2
D3	2.4	3.6	2.2	2.2	D_7
D4	2.8	3.0	2.6	2.6	D_7
D5	2.6	0.0	2.8	0.0	D_5
D6	2.4	3.9	2.6	2.4	D_2
D7	1.7	2.8	0.0	0.0	D_7
D8	2.6	2.0	2.8	2.0	D_5
D9	2.0	3.0	3.6	2.0	D_2
D10	2.2	3.5	2.4	2.2	D_2



2.b) K-Means Clustering.

Pros: computational cost → $O(K*n*d)$

① Fast, robust and easier to understand.
② Gives Best result when data set are distinct or well separated 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 infer the content spread of each sentence by a word count
② We can derive the proportions that each word constitutes in given topics.

Cons:
① We have to specify the number of topics.
② LDA's effeciency is pretty low when compared to machine learning algorithms.
③ LDA cannot capture co-relations.
④ Unsupervised (sometimes we need supervision Ex: sentiment analysis)
⑤ Uses BOW (assumes words are exchangeable)