Name: Sadanand Kallakuri

Class ID: 9

PROBLEM SET 5

1.a. Latent Dirichlet Allocation (LDA)

It is an probabilistic is used to generate the topics. LDA is the iterative model which requires 3 parameters. which are number of topics and deep knowledge of dataset.

The performance of LDA using perplexity is evaluated. 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 is then composed with the word distribution of the 2nd half of the document. Then a measure of distance is extracted. Perplexity is often used to select the best no. 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

for each iteration do

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

word $\leftarrow$ $w[i]$

topic $\leftarrow$ $z[i]$

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

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

$$p(z = k \mid) = (n_{d,K} + a_K)\frac{n_{K,w} + \beta_w}{n_K + \beta \times W}$$

end

topic $\leftarrow$ sample from $p(z \mid)$

$z[i] \leftarrow$ topic

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

end

end

return $z, n_{d,K}, n_{K,w}, n_K$

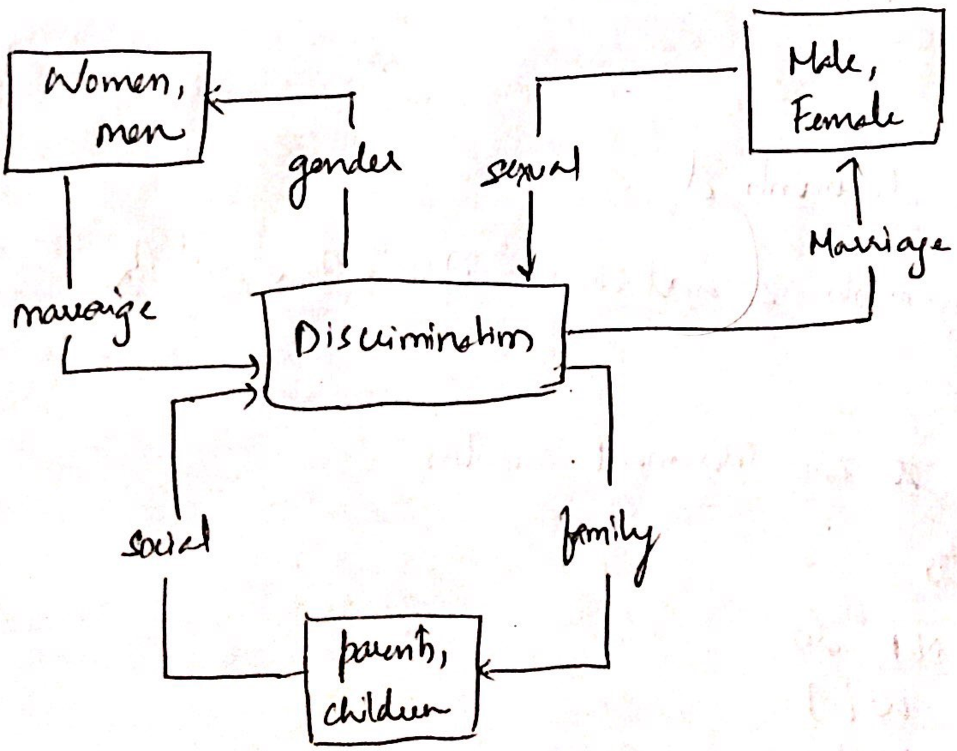
end

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) <u>Knowledge Graph</u>



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

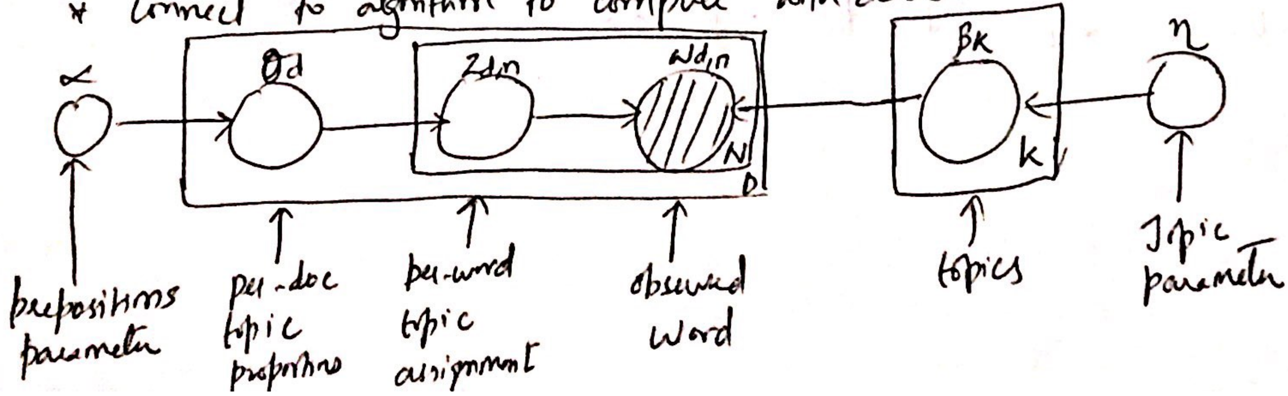
Since the words in Doc Y are assigned to Topic F and Topic P in a 50-50 ratio, the remaining "fish" words seems equally likely to be about either topic.

	Doc X		Doc Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	kitten
F	Vegetables	P	kitten

1d) Inference algorithm used in LDA

* Each topic is a distribution over words.
* Each document is a mixture of corpus-wide topics.
* Each word is drawn from one of the topics
* Observe the documents.
* Goal is to infer the hidden variables, compute distribution conditioned on the documents.

  P(topics, proportions, assignments / documents)
* Encode assumption.
* Define a factorization of the joint distribution.
* 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) k-means clustering vs LDA

We have to create $k = 3$ clusters.

Lets 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 F → Flight D → Delhi

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

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

D_1 to $D_5 = \sqrt{(1-3)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{7} = 2.6$

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 $D_2 = 0$

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

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_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_7 to $D_7 = 0$

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

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

D_8 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_9 to $D_2 = \sqrt{4} = 2$

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

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

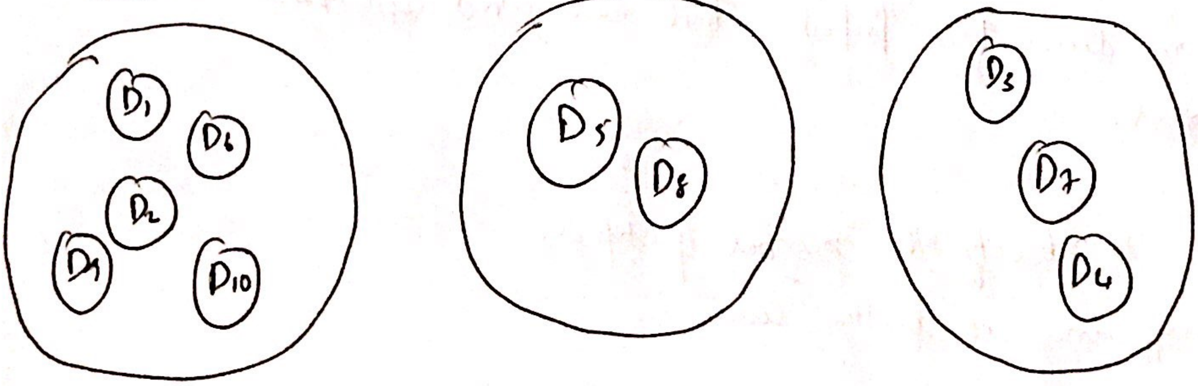
D_{10} to $D_2 = \sqrt{5} = 2.2$

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	Min distance	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
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	0.0	2.8	0.0	D_5
D_6	2.4	3.9	2.6	2.4	D_2
D_7	1.7	2.8	0.0	0.0	D_7
D_8	2.6	2.0	2.8	2.0	D_5
D_9	2.0	3.0	3.6	2.0	D_2
D_{10}	2.2	3.5	2.4	2.2	D_2

D_2 Cluster　　　D_5 Cluster　　　D_7 Cluster



2. b) Pros & Cons of K-Means Clustering:

Pros:

1. Fast → Computational cost → O(k×n×d)
2. Robust and easier to understand.
3. Gives best result when data set are distinct or well separated from each other.
4. It is a great solution for pre-clustering
5. Works great for spherical clusters.

Cons:

1. K-value is not known and is difficult to predict.
2. Does all 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 efficiency is pretty low.
3. LDA cannot capture co-relations.
4. Unsupervised
5. Uses BOW.

Scanned by CamScanner