Knowledge discovery and management Problem set 5

name : prudhvi sai suggala class id : 26

1) Latent Drichlet Allocation is a probabilistic used to generated the topics. LDA is the iterative model which requires 3 parameters, which are number of topics and alpha 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 Composed with the word distribution is Estimated. this distribution is then Composed 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 $w \in$ Document d

output : topic assignment z and Counts $n_{d,k}$, $n_{k,w}$ and n_k

begin

Randomly initialized z and increment Counters for Each interation do

for $i = 0 \to N-1$ 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

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

End

topic $\leftarrow$ Sample from $p(z_i)$

$z[i] \leftarrow$ topic

n_d, topic $+=1$; $n$word, topic $+=1$; $n$topic $+=1$

End

End

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

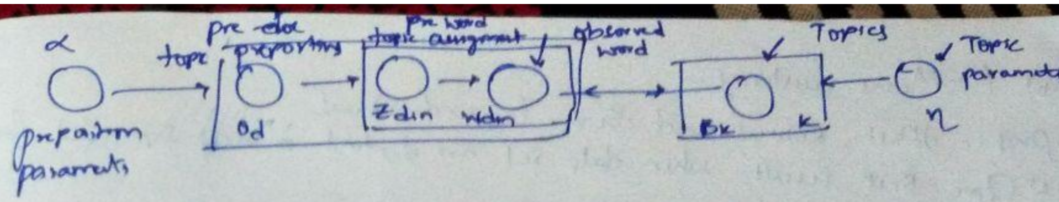
1. decide How many topics you need

2. The algorithm will assign Every word to a tempora[ry] word

3. will Check and update to[pic]

Knowledge discovery and management Problem set 5

name : prudhvi sai suggala class id : 26



$$P(\beta,\theta,Z,W) = \left(\prod_{t=1}^{k} P(\beta, |\eta)\right)\left(\prod_{d=1}^{D} P(Z_{dn}/\theta_d)\; P(W_{d,n}/\beta_{:k}\, Z_{d,n})\right)$$

2.a) K = 3 Centers D_2 D_5 D_7

D_1 to D_2 = $\sqrt{4}$ = 2

D_1 to D_5 = $\sqrt{7}$ = 2.6

D_1 to D_7 = $\sqrt{5}$ = 2.2

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

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

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

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

D_3 to D_2 = $\sqrt{13}$ = 3.6

D_3 to D_5 = $\sqrt{5}$ = 2.2

D_3 to D_7 = $\sqrt{7}$ = 2.6

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

D_5 to D_5 = 0 = 0

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

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

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

D_7 to D_2 = $\sqrt{7}$ = 2.4

D_8 , D_5 = $\sqrt{6}$ = 2.2

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

D_7 to D_2 = 2.2

D_9 to D_5 = 3.4

D_9 to D_7 = 3.2

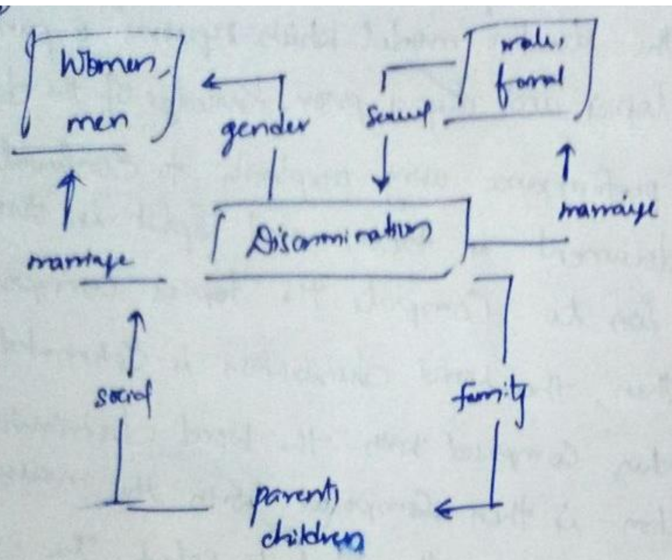
D_{10} to D_2 = 2.4

D_{10} to D_5 = 2.2

Doc	D_2	D_5	D_7	Min	Cluster
D_1	2.0	2.6	2.2	2.0	D_2
D_2	0	2.6	2.7	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	2.4	D_5
D_6	2.4	3.9	2.6	0	D_2
D_7	1.7	2.8	0	2.0	D_7
D_8	2.6	2.0	2.8	2.0	D_5
D_9	2.0	3.0	3.6	2.2	D_2
D_{10}	2.2	3.5	2.4		D_2

Knowledge discovery and management Problem set 5

name : prudhvi sai suggala class id : 26



(1c) ~~How~~ Since the words in the document Y are assigned to Topic C and Topic P in a 50-50 ratio, the remaing "fish" word seems Equality 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	Vegetably	P	Kitten

(1d)

Each topic is distribution Over words.

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 struvre are hidden Variable.
6) Our goal is to infer the hidden values i.e Compute their distribution Conditional on the document.
7) Encode assumtion
8) Define a factorization of the joint distribution

Knowledge discovery and management Problem set 5
name : prudhvi sai suggala class id : 26

2.b) K-Means clustering :-

Pros :- 1) Fast, Robust and Easier to understand.
2) Give 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.

LDA TOPIC Discovery Model.

Pros: we can infer content 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 co-relation.
4) unsupervised
5) uses BOW (words are exchangeable)