NAME :- NAGESWARARAO CLASS ID = 18

1a) LDA :

In natural language processing, latent dirichlet allocation is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. If observations are words collected into documents, it posits that each document is a mixture of small number of topics and that each word's collection is attributable to one of the documents topics.

How to create the topics from the corpus

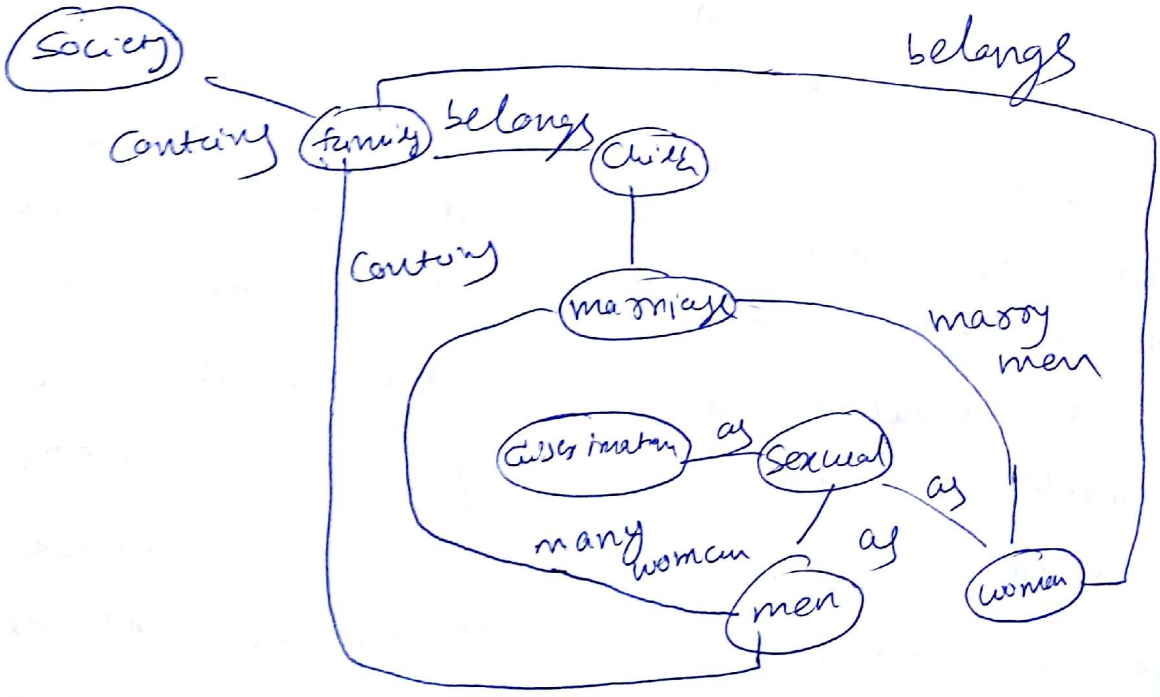
In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA. For example, an LDA model might have topics that can be classified as CAT related and dog-related. A topic has probabilities of generating various words. Such as milk, meow and kitten which can be classified and implemented by the viewer as "CAT related". The dog. The dog related topic likewise has the probability of generating each word.

1b) Knowledge graph for topic 3 in Yale law Journal

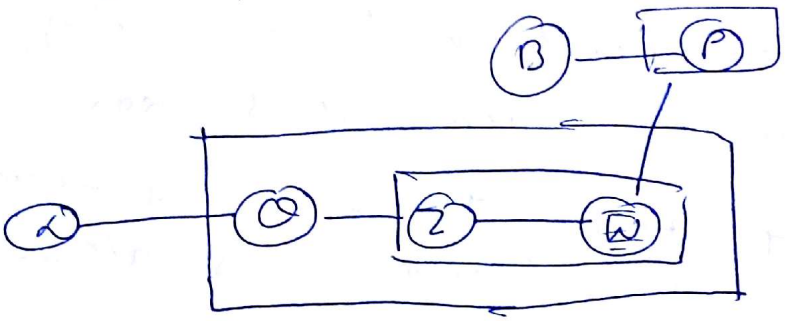
In the figure given in the problem set there are top eight topics were displayed each topic will be illustrated with its top-most frequent words. Each word's position along the x-axis denotes its specificity to the documents.

Topic 3 in the yale's law - has the following words women, Sexual, men, Sex-child, family-children, gender woman, marriage, discrimination, male, social, female, Parents.

The most important words which were spread among the x-axis is the topic 3 ax the basis for the construction of the knowledge graph



1c) Learning generality or specificity of topic models




The dependency among the many variables can be captured concisely. The boxes are plates representing replicas - the outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Generative process

Documents are represented as random mixtures over latent topics where each topic is characterised by a distribution over words. LDA assumes the following generative process for a corpus D consisting of M documents each of length $N_i$.

(1) Choose $\theta_i \sim Dir(\alpha)$ where $i \in \{1, \ldots M\}$ and $Dir(\alpha)$ is a - - dirichlet distribution.

(2) Choose $\phi_k \sim Dir(\beta)$ where $k \in \{1 \ldots K\}$

(3) for each word position $j$, where $j \in \{1 \ldots N_i\}$ and $i \in \{1 \ldots M\}$

1d) Inference algorithm in LDA :-

The goal of topic modeling is to automatically discovery the topics from a collection of documents. the documents and words are observed. the topic structure is hidden the topics, per document topic distribution. Per-document per-word topic assignment. we use observed variable to infer the hidden structure.

We can infer the content spread of each sentence by a word count.

step1 :- you tell the algorithm how many topics we think there are.

step2 :- the algorithm will assign every word to a temporary topic.

step3 :- the algorithm will check and update the topic assignments.

The posterior computation over hidden variable given a document.

$$P(z, \phi, \theta | w, \alpha, \beta) = \frac{P(z, \phi, \theta, w | \alpha, \beta)}{P(w | \alpha, \beta)}$$

The document represented as continous mixture

$$P(w | \alpha, \beta) = \int P(\theta | \alpha) \left( \prod_{n=1}^{N} P(w_n | \theta, \beta_i) \right) d\theta$$

for topic K, term V.

$$n_{kv} = \beta_{kv} + \sum_d \sum_n I(w_{dn} = v) \phi_{dnk}$$

2a) Given the term/document matrix

| Document | online | festival | books | figure | delhi |
|---|---|---|---|---|---|
| $D_1$ | 1 | 0 | 1 | 0 | 1 |
| $D_2$ | 2 | 1 | 2 | 1 | 1 |
| $D_3$ | 0 | 0 | 1 | 1 | 1 |
| $D_4$ | 1 | 2 | 0 | 2 | 0 |
| $D_5$ | 3 | 1 | 0 | 0 | 0 |
| $D_6$ | 0 | 1 | 1 | 1 | 2 |
| $D_7$ | 2 | 0 | 1 | 2 | 1 |
| $D_8$ | 1 | 1 | 0 | 1 | 0 |
| $D_9$ | 1 | 0 | 2 | 0 | 0 |
| $D_{10}$ | 0 | 1 | 1 | 1 | 1 |

Step 1:- Given also the distance matrix. Here are 3 clusters $D_2, D_5, D_7$ as per the diagram as we get distance as 0.0 for above 3 which indicates that $D_2, D_5, D_7$ are the centroids. The remaining documents have moved into those 3 different clusters using K-means $K = 3$

$D_2 :- D_1, D_6, D_9, P_{10}$ $D_1 :- D_3, D_4,$ $D_5 :- D_8$

the first row of the distance matrix corresponds to the distance of each object to the first centroid and the second row is the distance of each object to the second centroid and based on minimum distance grouping is done.

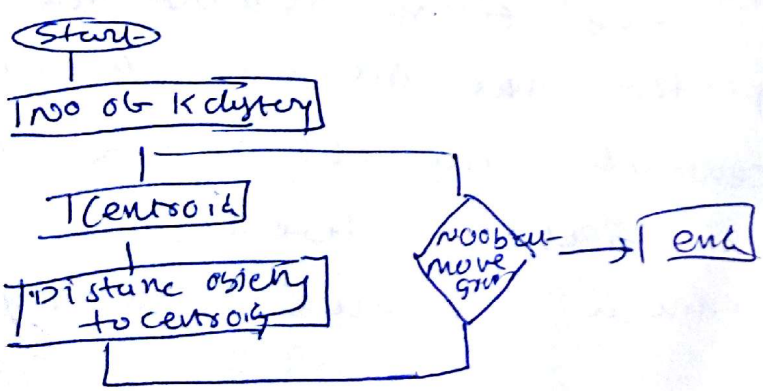
2) Clustering:

Clustering/segmentation is one of the most important techniques used in acquisition analytics. It is the process of making a group of abstract objects into classes of the of making a group of abstract objects into classes of the of making a ~~group of~~ similar objects. We will partition the observations into a cluster in such a way that they are similar in sense.

clustering is a method of unsupervised learning, and a common technique for the statistical data analysis used in many fields.

K-means Clustering:

K-means clustering is an algorithm to classify or to group your objects based on attributes/ features into K number of groups K is positive integer number. the grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.



There are 3 centroids randomly taken

$D_2(2,1,2,1)$ $D_5(3,1,0,0,0)$ $D_7(2,0,1,2,1)$

Step 2 :- Now calculate the distance for $D_1$ from $D_2, D_5, D_7$

$D_1 \rightarrow D_2$

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

$D_1 \rightarrow D_5$

$$\sqrt{(1-3)^2+(0-1)^2+(1-0)^2+(1-0)^2} = \sqrt{4+1+1+1} = \sqrt{7} = 2.6$$

$D_1 \rightarrow D_7$

$$\sqrt{(1-2)^2+(0-0)^2+(1-1)^2+0-2)^2+(1-0)^2} = \sqrt{1+0+0+1+0} = \sqrt{5} = 2.2$$

Step 3 :-

group the data into clusters based on their minimum distance.

$D_2 :- \{D_1, D_6, D_9, D_{10}\}$

$D_7 = \{D_3, D_4\}$

$D_5 = \{D_8\}$

In the above steps using the K-means algorithm we will cluster the data points based on the centroids and we will reiterate this process by calculating the new mean and new clusters

2b) The difference between k-means and LDA are very heavy

-> If both are applied to assign k topics to set of N documents. k-means is going to partition the N documents in k disjoint clusters while LDA assigns a document to a mixture of topics.

-> k-means is hard clustering while LDA is soft clustering

LDA pros:

-> LDA is on the exponential family and conjugate to the multinomial distribution

-> feature set is reduced

-> one document can be associated with multiple topics

cons:

-> unable to capture the correlation between the different topics.

k-means pros:

-> simple, easy to implement

-> easy to interpret the clustering result.

-> It is a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied

k-means cons

-> ~~difficult~~ difficult to predict k-value

-> with global clusters, it didn't work well.