Rohithkumar N
Class ID: 16

Knowledge Discovery & Management

(1) a) LDA (Latent Dirichlet allocation)

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

⇒ How to create the topics from the corpus?

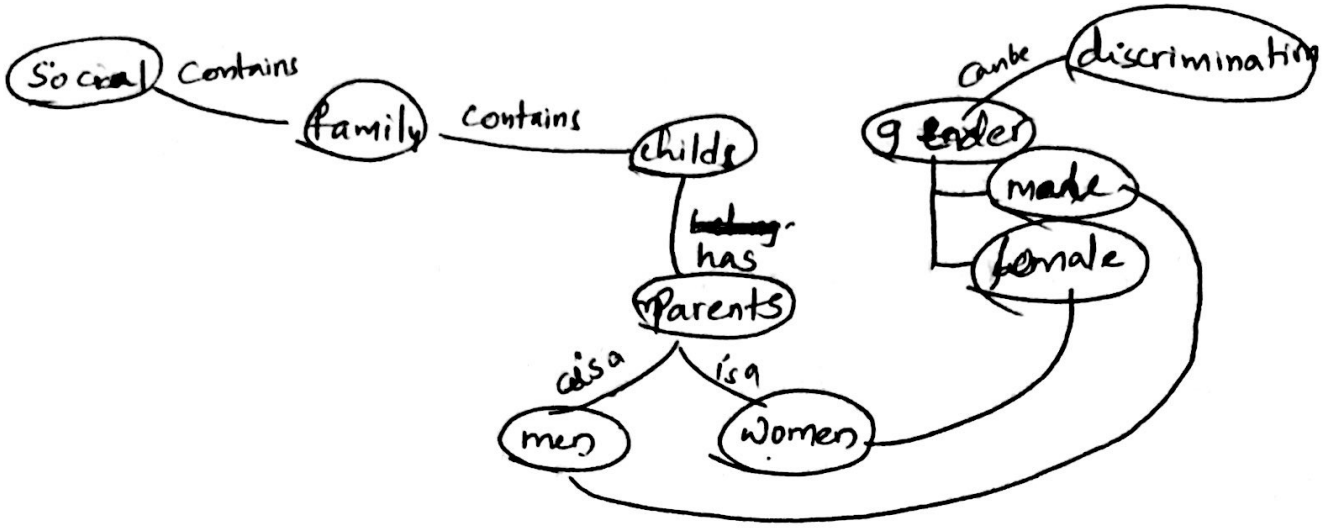
In LDA, each document is viewed as a mixture of various topics that are assigned to it by LDA model. For example we have a collection of documents talking about ten topics the algorithm estimates the probability of a token falling into each topic and assigns confident score for the token falling in to the topic.

1b) Knowledge graph for Topic 3 in Yale Law Journal

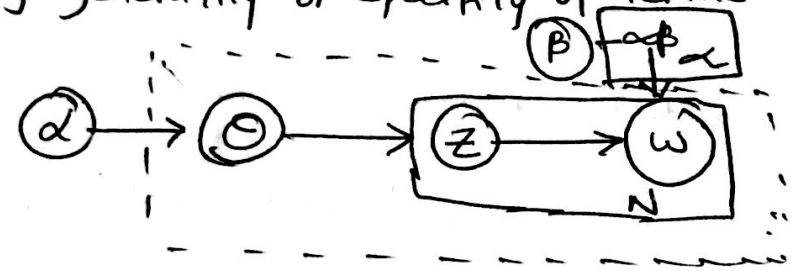
Given 8 topics and listed top-most frequent words. Each word's position along the x-axis denotes its specifity to the documents

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.



11) Determining generality or specifity of terms in a topic.



The dependencies among the many variables can be captured concisely. The boxes are places representing replicas. The outer plate represents documents, while the inner plate represents documents, while the inner plate represents the repeated choice of topics and words with in a document.

Generative process:-

Documents are represented as a mixture over latent topics where each topic is characterized by a distribution of words.

LDA assumes the following generative process for a corpus D consisting of M documents each of length N_i.

i) Choose $\theta_i \sim Dir(\alpha)$ where $i \in 1, 2 \dots M$ and $Dir(\alpha)$ is a Dirchlet algorithm.

2) Choose $\varphi_{ik} \sim Dir(\beta)$ where $k \in \{1, \dots *\}$

3) For each word positions ij where $j \in \{1 \dots N_i\}$ and $i \in \{1 \dots M\}$

The generality or specificity of the terms can be determined by the document frequency (DF) the more documents a term occurred in, the more general it was assumed to be.

1d) Inference Algorithms in LDA :-

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

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

Step 1 :- You tell the algorithms 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 variables given a document

$$p(z, \theta, \theta | \omega, \alpha, \beta) = p(z, \phi, \theta, \omega | \alpha, \beta)) d\theta$$

For topic k, term V

$$\lambda_{kv} = \beta_{kv} + \sum_{d} \sum_{n} I[\omega_{dn} = v] \varphi_{dnk}$$

For each document d = $\gamma_{dk} = \alpha_k + \sum_{n} \varphi_{dnk}$

For each word n $\varphi_{dnk} \propto \exp\{ E_q [\log(\theta_{dk}) + \log(\phi_{kwan})] \}$

(2) Clustering :-

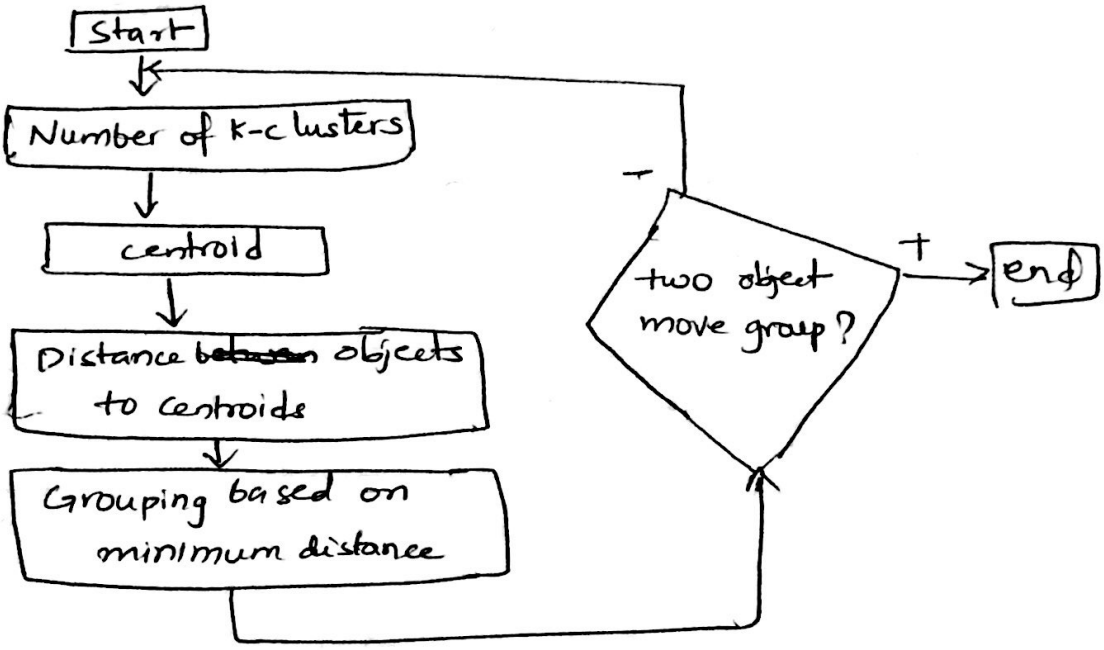
Clustering / Segmentation is one of the most important techniques used in Acquistion Analytics. It is the process of making a group of abstract objects into classes of the similar objects. We will partition the observations into a cluster in such a way that they are similar in sense.

Clustering is a ~~word~~ method of unsupervised learning and 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 group K is positive integer number.

The grouping is done by minimizing the sum of squares of Squares of distances between data and the corresponding cluster centroid.



(2 a) Given the distance matrix. There are 3 clusters D_2, D_5, D_7 as per the diagram we get distance as 0.0 for above 3 which indicate that D_2, D_5, D_7 are the centroids. The remaining documents have moved in to those 3 different clusters using k-means K=3

$D_2 :- D_1, D_6, D_9, D_{10}$ $D_7 :- D_3$ 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.

There are 3 centroids randomly taken.

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

step2:- calculating distrance for D_1 from D_2, D_5 & D_7

$D_1 \to D_2 = \sqrt{(1-2)^2+(0-1)^2+(1-2)^2+(1-0)^2+(1-1)^2} = \sqrt{1+1+1+1+0} = \sqrt{4} = 2$

$D_1 \to D_5 = \sqrt{(1-3)^2+(0-1)^2+(1-0)^2+(1-0)^2+(1-0)^2} = \sqrt{4+4} = 2\sqrt{2} =$

$D_1 - D_7 = \sqrt{(1-2)^2+(0-0)^2+(1-1)^2+(0-2)^2+(1-1)^2} = \sqrt{1+0+0+4+0} = \sqrt{5} = 2.2$

similary we calculate sum of squares of distance from each point to the centroid.

Step3:- Group the data into clusters based on these 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 datapoints based on the centroid and we will reiterate this process by calculating the new mean & new clusters.

25) (1)

The differences between K-means and the LDA are as follows

If both are applied to assign K-topics to a set of 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 in 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 subspaces where other clustering algorithms can be applied

→ The clusters are non-hierarchial and they do not overlap

→ It is computationally faster.

→ The clusters are globular.

K-means Cons:-

→ Difficult to predict K-value.

→ With global cluster, it didn't work well.

→ Doesn't work well with non-circular cluster shape- number of cluster and initial seed value need to be specified before hand.

→ Applicable only when mean is specified.

→ Sensitive to outliers.