CS5560 Knowledge Discovery and Management Problem Set 5
Name: Sai Teja Makani **July 3 (T), 2017**
Class ID: 12

1.

a) LDA(Latent Dirichlet allocation):

In natural language processing, LDA 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?**

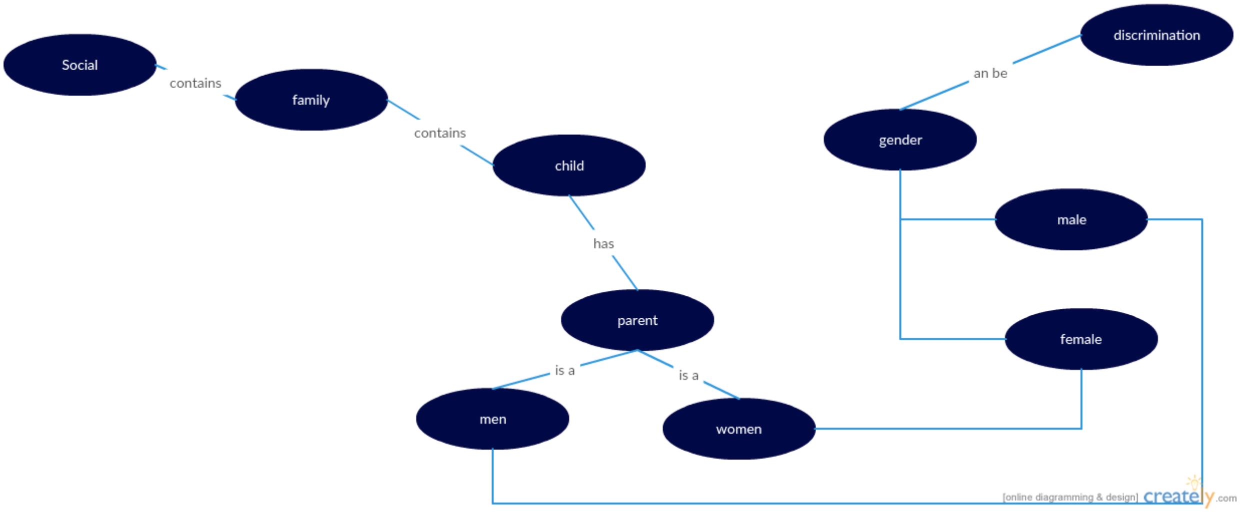
  In LDA , each document is viewed as a mixture of various topics that are assigned to it by LDA model. For example we have a collection of documents talking about ten topics the algorithm estimates the provability of a token falling into each topic and assigns confident score for the token falling into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

b) Knowledge graph for Topic3 in Yale Law Journal:

Given 8 topics and listed top most frequent word. Each word’s position along the X-axis denotes its specificity to the documents.

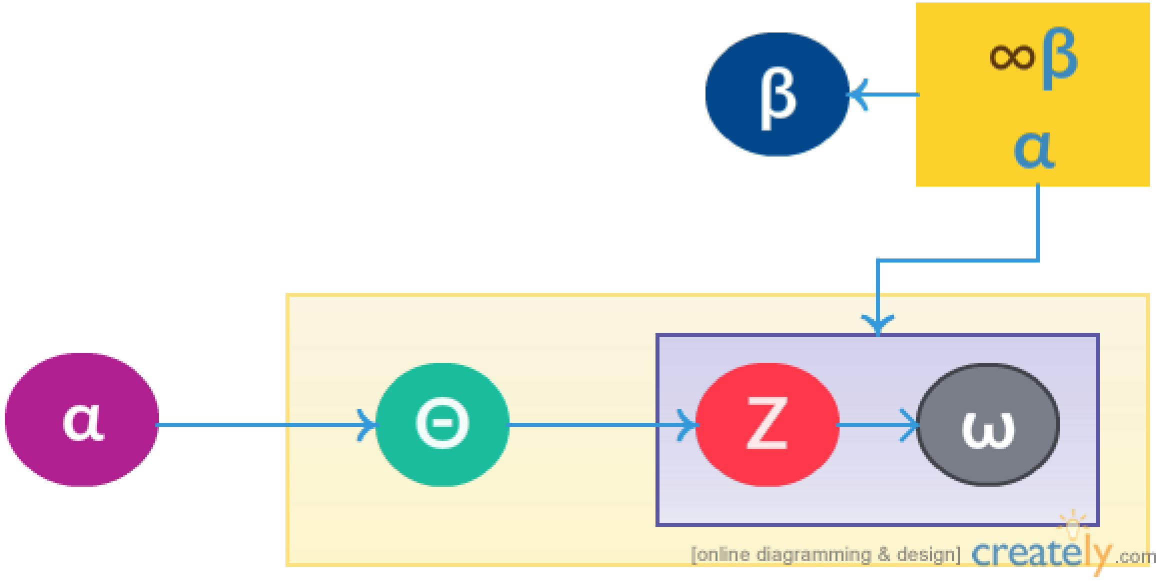
Topic3 in the Yale’s Law has the following words:

Women, Sexual, men, sex, child, family, children, gender, woman, marriage, discrimination, male, social, female, parents.



CS5560 Knowledge Discovery and Management Problem Set 5
Name: Sai Teja Makani July 3 (T), 2017
Class ID: 12

c) Determining generality or specificity 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 represents documents, while the inner plate represents the respected choice of topics and word in a document.

<u>Generations</u>:

Documents are represented as a number 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 Ni.

- Choose $\Theta i \sim Dir(a)$ where {1,2,3,…M} and $Dir(\alpha)$ is Dirchlet algorithm.
- Choose $\Phi k \sim Dir(\beta)$ where k∈{1,2…*}
- For each word positions I ij where J∈(1,2…Ni} and i.e. {1,2….M}.

CS5560 Knowledge Discovery and Management Problem Set 5
Name: Sai Teja Makani July 3 (T), 2017
Class ID: 12

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

d) Inference algorithm in LDA:

The goal of topic modeling is to automatically discover the topics from a collection of documents. The documents and word are discovered. 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.

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 assignment.

The posterior commutation over hidden variables given a document.

$$P(Z,Q, \Theta/\omega, \alpha, \beta) = P((Z, \Phi , \Theta, \omega/\alpha, \beta))d\Theta$$

For topic k1 term V: $\lambda_{KV} = \beta_{KV} + \sum_{d} \sum_{n} I\{\omega_{dn} = V] \phi_{dnk}$

For each document d = $\gamma_{dk} = \alpha_{k} + \sum_{n} \gamma_{dnk}$

For each word n $\theta_{dn} \propto \exp\{E_q[\log(\theta_{dk}) + \log(\phi_{k\omega_{dn}})]\}$

CS5560 Knowledge Discovery and Management Problem Set 5
Name: Sai Teja Makani July 3 (T), 2017
Class ID: 12

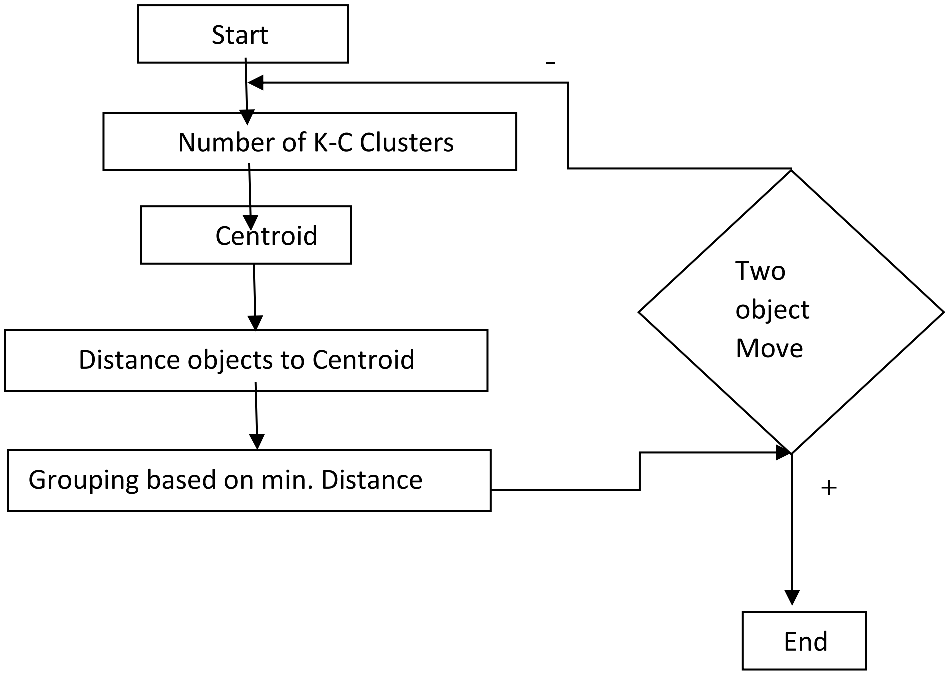
2. Clustering:

Clustering is one of the most important technique used in acquisition analysis. It is the process of making a group of abstract objects into classes of the similar objects. We will partition the observations in to a cluster in such a way that they are similar in sense. Clustering is 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 on attributes into k-number of group k is positive integer number.

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



a)

Given the distance matrix. There are three clusters D2, D5, D7 as per the diagram. We get the distance as 0 for above 3 which indicates that D2, D5 and D7 are centroids. The remaining documents have moved into those three different clusters using K-means K=3

D2: D1, D6, D9, D10
D7: D3, D4
D5: D8

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 based on minimum distance grouping is done.

There are three Centroids randomly taken:

D2 (2,1,2,1,1) D5 (3,1,0,0,0) D7 (2,0,1,2,1)

Calculating distance from D1 to D2, D5, D7

D1 -> D2 = $\sqrt{(1+1+1+1+0)}$ = 2
D1 -> D5 = $\sqrt{}(4+4)$ = $2\sqrt{}2$
D1 -> D7 = $\sqrt{}(1+0+0+4+0)$ = 2.2

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

Group the data in to clusters based on these minimum distance

D2: {D1, D6, D9, D10}
D5: {D8}
D7: {D3, D4}

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

b)

The difference between K-means and 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.

LDA Cons:

- Unable to capture the correlation between the different topics.

K-means Pros:

- Simple and easy to implement.
- Easy to interpret the clustering result.
- The clusters are non-hierarchical and they do not overlap.
- It is a great solution for pre-clustering reducing the space into disjoint smaller subspaces where other clustering algorithms can be applied.

K-means Cons:

- Difficult to predict the K-value
- With global cluster, it did not work well.
- Applicable only when mean is specified.