CS5560 Knowledge Discovery and Management

Problem Set 5

July 3 (T), 2017

Name: SreeLakshmi Nandanamudi

Class ID: 17

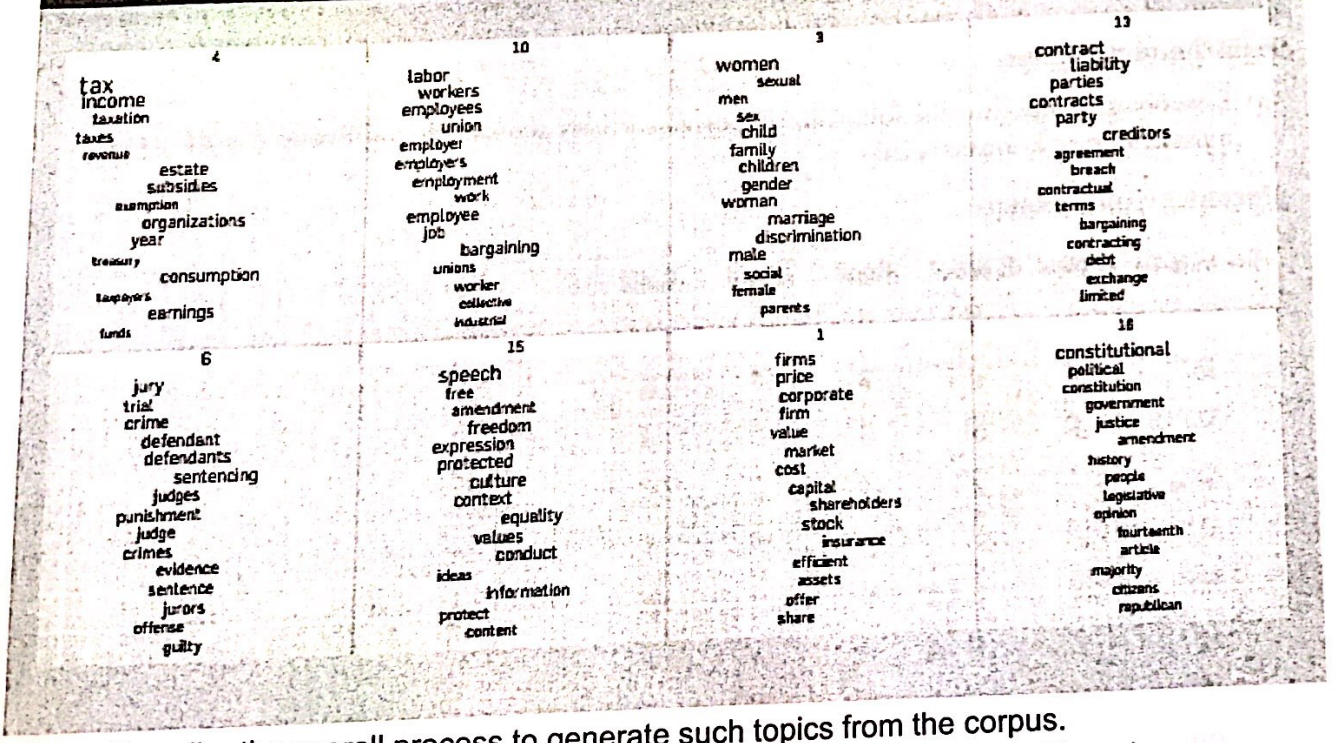
1. LDA

Read the following articles to learn more about LDA

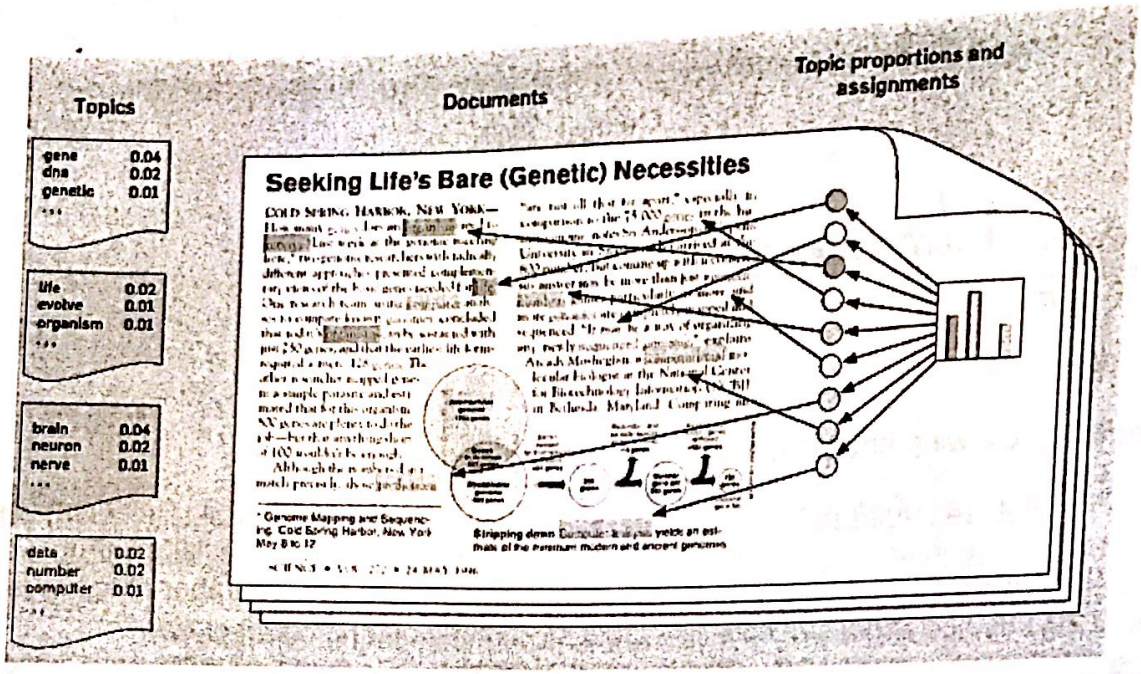
- https://algobeans.com/2015/06/21/laymans-explanation-of-topic-modeling-with-lda-2/
- http://engineering.intenthq.com/2015/02/automatic-topic-modelling-with-lda/

Consider the topics discovered from Yale Law Journal. (Here the number of topics was set to be 20.) Topics about subjects like about discrimination and contract law.

Figure 3. A topic model fit to the *Yale Law Journal*. Here, there are 20 topics (the top eight are plotted). Each topic is illustrated with its top-most frequent words. Each word's position along the *x*-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific than "tax."



a. Describe the overall process to generate such topics from the corpus.
b. Draw a knowledge graph for Topic 3 in Yale Law Journal (The First Figure).
c. Each topic is illustrated with its topmost frequent words. Each word's position along the x-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific than "tax." (the second figure). Describe how to determine the generality or specificity of the terms in a topic.
d. Describe the inference algorithm that was used in LDA.



2. K-means clustering vs. LDA

Read the K-means clustering for text clustering from https://www.experfy.com/blog/k-means-clustering-in-text-data

(a) Describe the steps how the following 10 documents have moved into 3 different clusters using clustered using k-means (K=3).

Document/Term Matrix

Documents	Online	Festival	Book	Flight	Delhi
D1	1	0	1	0	1
D2	2	1	2	1	1
D3	0	0	1	1	1
D4	1	2	0	2	0
D5	3	1	0	0	0
D6	0	1	1	1	2
D7	2	0	1	2	1
D8	1	1	0	1	0
D9	1	0	2	0	0
D10	0	1	1	1	1

Distance Matrix

Documents	Distance from 3 clusters				
	D2	D5	D7	Min. Distance	Movement
D1	2.0	2.6	2.2	2.0	D2
D2	0.0	2.6	1.7	0.0	
D3	2.4	3.6	2.2	2.2	D7
D4	2.8	3.0	2.6	2.6	D7
D5	2.6	0.0	2.8	0.0	
D6	2.4	3.9	2.6	2.4	D2
D7	1.7	2.8	0.0	0.0	
D8	2.6	2.0	2.8	2.0	D5
D9	2.0	3.0	2.6	2.0	D2
D10	2.2	3.5	2.4	2.2	D2

(b) Describe the difference (pro and con) of k-means clustering and the LDA topic discovery model.

1. a) Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model of a corpus. The documents are represented as a random mixtures over latent topics, where a topic is characterized by a distribution over words. Allows each document to exhibit multiple topics, but ignores the correlation between topics.

Applications:

Document modeling, text classification, image processing, Collaborative filtering, etc.

The overall process to generate 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 interpreted by the viewer as "CAT_related". The DOG_related topic likewise has the probabilities of generating each word: puppy, bark and bone might have high probability.

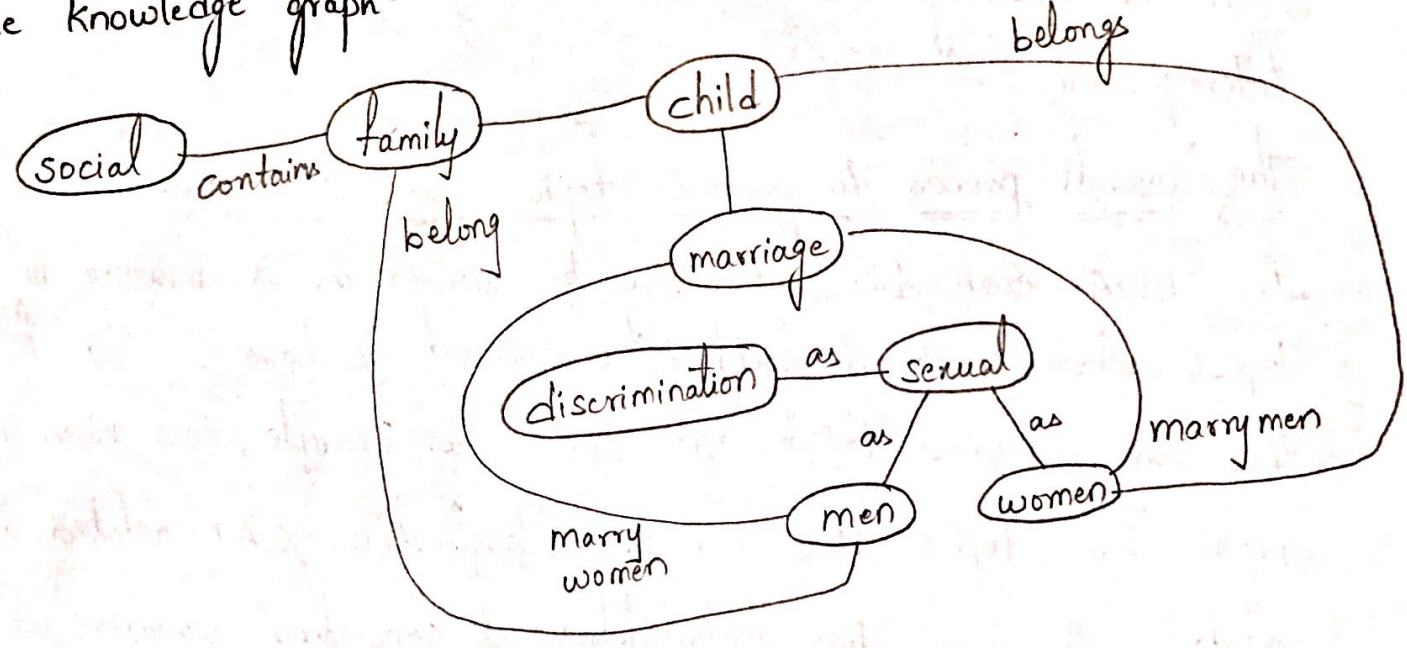
b) Knowledge graph for Topic 3 in Yale Law Journal:

In the figure 3, there are 20 topics, the top eight are plotted. Each topic is illustrated with its top most frequent

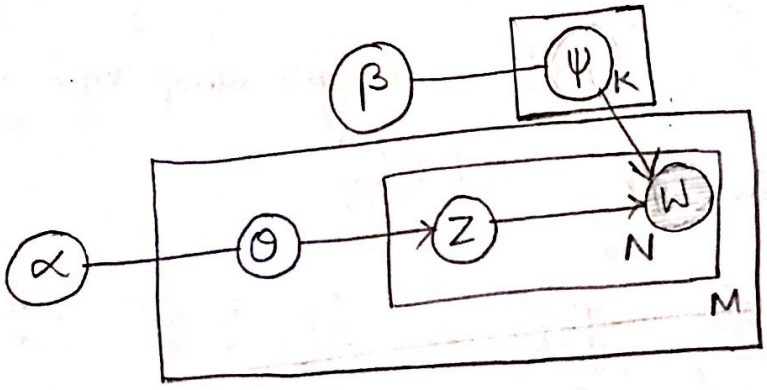
words. Each word's position along the X-axis denotes its specificity to the documents.

In Yale Law Journal, the topic 3 has the following words: Women, sexual, men, sex, child, family, children, gender, woman, marriage, discrimination, male, social, female, parents.

The top-most frequent words which were spread among the X-axis is the topic 3 which are the basis for the construction of the knowledge graph.



c) Determining the generality or specificity of the terms in a topic.



the dependencies 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 the document

Generative process :

Documents are represented as random mixtures over latent topics, where each document in the topic is characterized 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 $\varphi_k \sim Dir(\beta)$ where $k \in \{1, \ldots K\}$.
3. For each word positions i, j where $j \in \{1, \ldots N\}$ and $i \in \{1, \ldots M\}$

the generality and specifity of the terms was determined by their Document frequency (DF) the more documents a term occurred in, the more generated the general it was assumed to be.

d) Inference algorithm used 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 topic structure is hidden. 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 assignments.

the posterior computation over hidden variables given a document.

$$P(z, \phi, \theta | w, \alpha, \beta) = P(z, \phi, \theta, w | \alpha, \beta) / P(w | \alpha, \beta)$$

the document represented as continuous mixture.

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

For topic k, term V

$$\lambda_{kv} = \beta_{kv} + \sum_d \sum_n I[w_{dn} = v] \psi_{dnk}$$

For each document d $\quad \gamma_{dk} = \alpha_k + \sum_n \psi_{dnk}$

For each word n

$$\psi_{dnk} \propto \exp\{ E_q[\log(\theta_{dk})] + \log(\theta_{kw\alpha n})] \}$$

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 similar objects. We will partition the observations into a cluster in such a way that they are similar in sense.

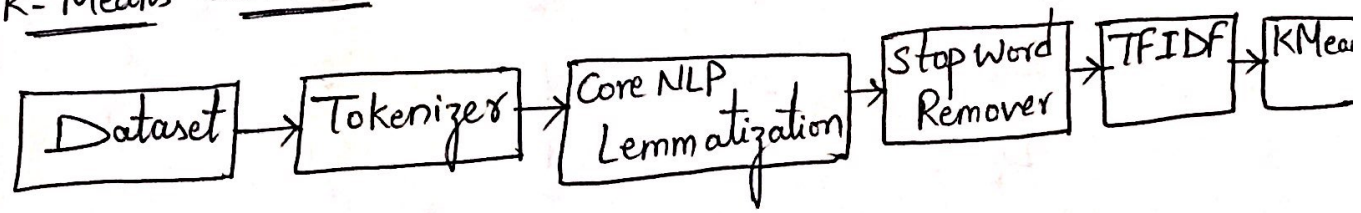
clustering is a method of unsupervised learning and a common technique for the statistical data analytics or analysis used in many fields.

kMeans clustering :

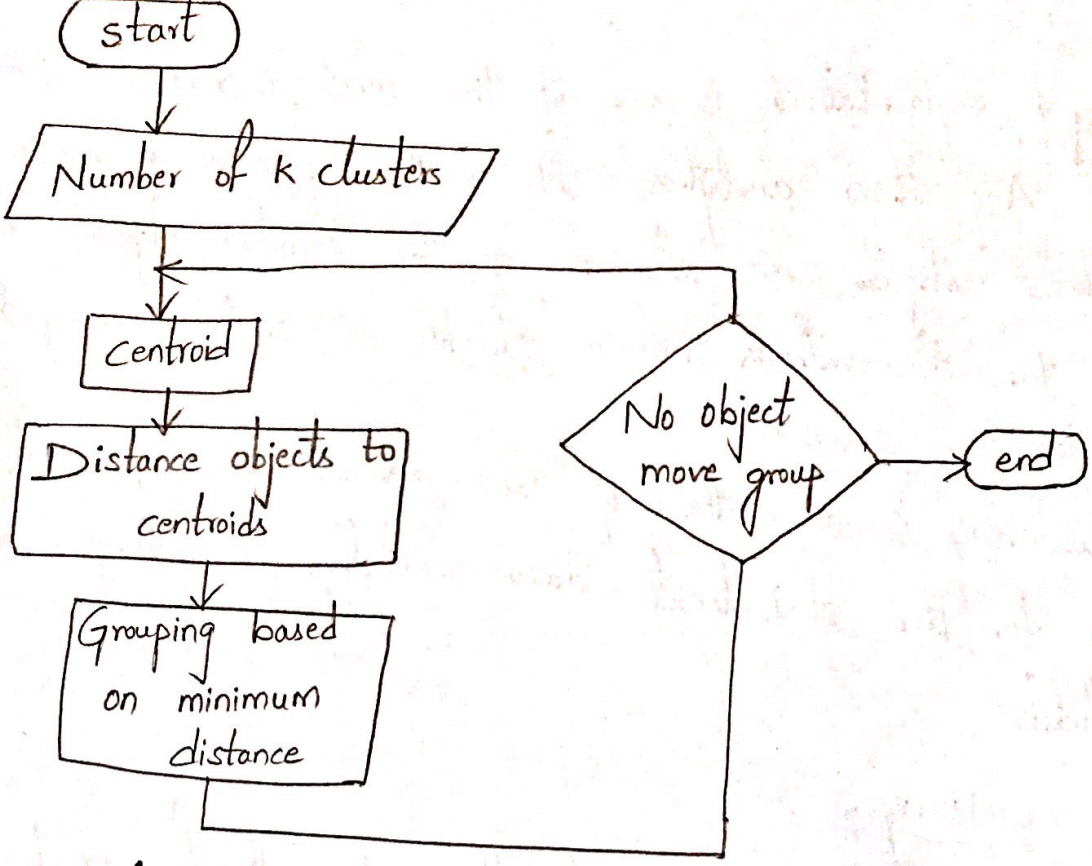
K-Means clustering is an algorithm to classify or to group your objects based on attributes / features into K number of group. K is positive integer number.

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

K-Means workFlow :



The below figure describes how the K-Means clustering Algorithm works.



a) Steps for the given 10 documents which have moved into 3 different clusters using clustered using k-means (k=3)

Documents	online	Festival	Book	Flight	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. There are 3 clusters D_2, D_5, D_7 as per the diagram. As we got 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, D_{10} $\quad D_7 : D_3, D_4 \quad D_5 : D_8$

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.

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

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

$D_1 \rightarrow D_2$

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

$D_1 \rightarrow D_5$

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

$D_1 \rightarrow D_7$

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

Likewise we will calculate the sum of squares of distance from each data point to the centroid.

step 3 : Group the data into clusters based on these 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 centroid and we will reiterate this process by calculating the new mean and new clusters.

b) The difference (pro and con) of K-Means clustering and the LDA topic discovery model are as follows :

If both are applied to assign K topics to a 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.

K-Means pros :

1) Simple, easy to implement
2) Easy to interpret the clustering result.
3) It is a great solution for pre-clustering, reducing the

space into disjoint smaller sub-spaces where other clustering algorithms can be applied.

4. The clusterings are non-hierarchial and they do not overlap.
5. It is computationally faster.
6. The clusters are globular.

K-Means Cons:

1. Difficult to predict k-value.
2. With global cluster, it didn't work well.
3. Doesn't work well with non-circular cluster shape - number of cluster and initial seed value need to be specified before hand.
4. Applicable only when mean is specified.
5. Sensitive to the outliers.

LDA pros:

1. The Dirichlet distribution is in the exponential family and conjugate to the multinomial distribution -- variational inference is tractable.
2. θ are document-specific, so the variational parameters of θ could be regarded as the representation of a document -- feature set is reduced.

3. Z are sampled repeatedly within a document -- one document can be associated with multiple topics.

LDA Cons:

1. Because of the independence assumption implicit in the Dirichlet distribution, LDA is unable to capture the correlation between different topics.