CS5560 Knowledge Discovery and Management

Problem Set 5

July 3 (T), 2017

Name: Revanth Chakilam

Class ID: 02

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.

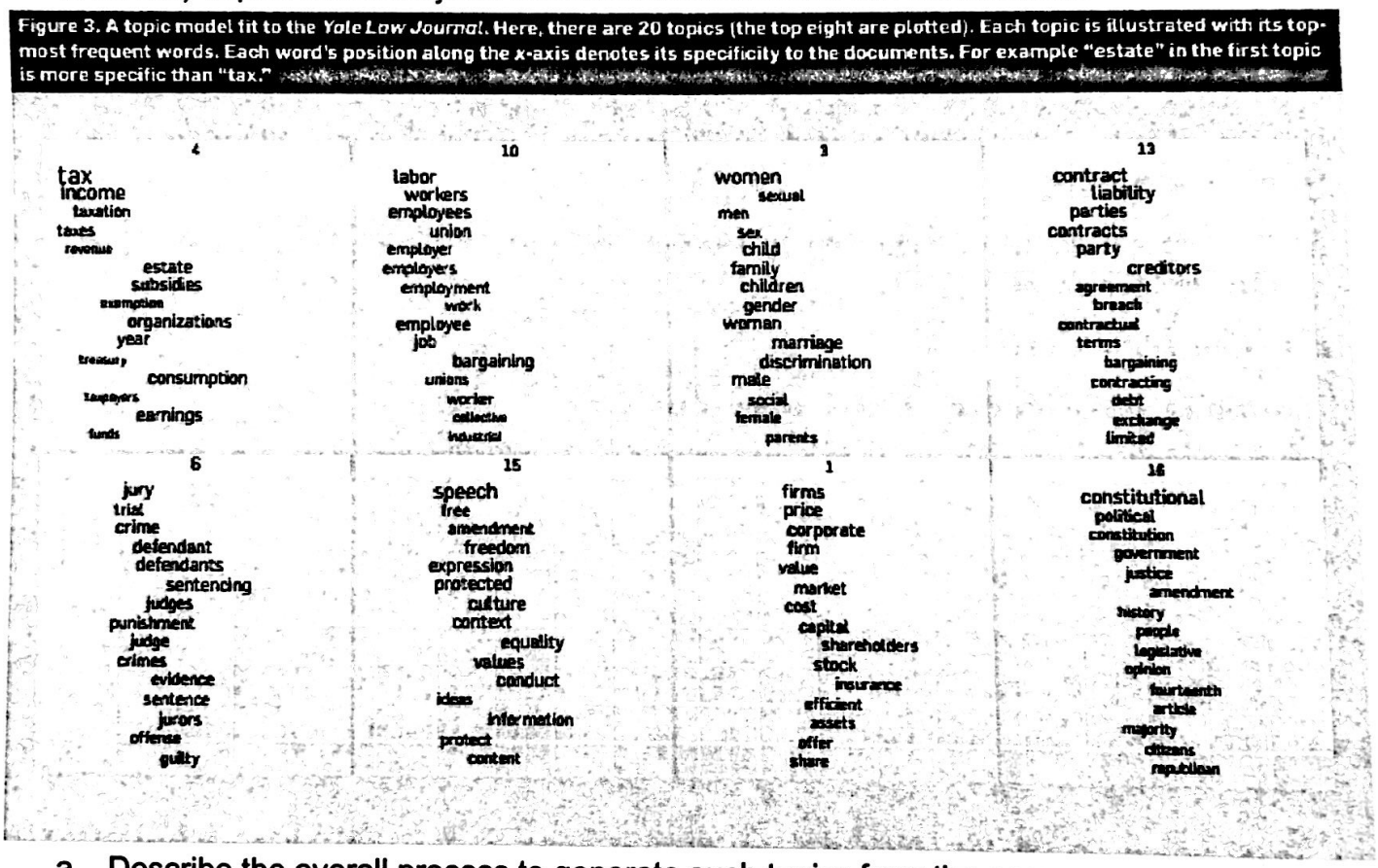
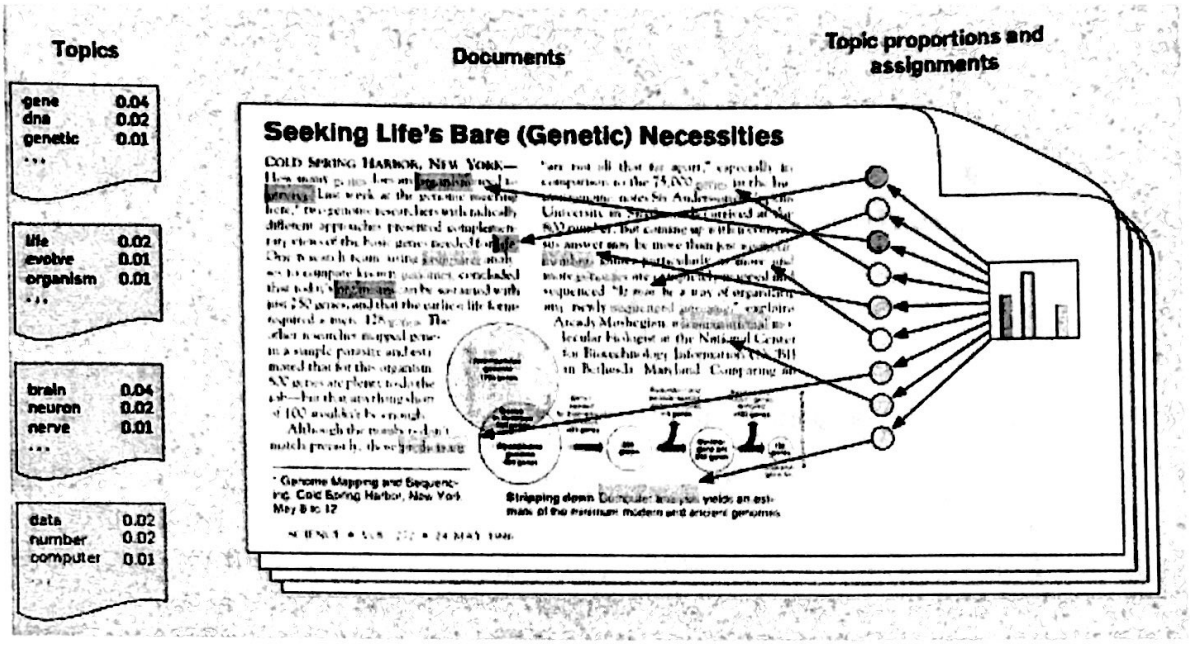


Figure 3. A topic model fit to the *Yale Law Journal*. Here, there are 20 topics (the top eight are plotted). Each topic is illustrated with its topmost frequent words. Each word's position along the x-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific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

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

	Distance from 3 clusters				
Documents	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.

Revanth Chakilam

1. (a) LDA :- (Latent Dirchlet Allocation)

When we consider NLP - Natural Language Processing, LDA 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 posts that each document is a mixture of small number of topics and that each words collection is attributable to one of the doc topics.

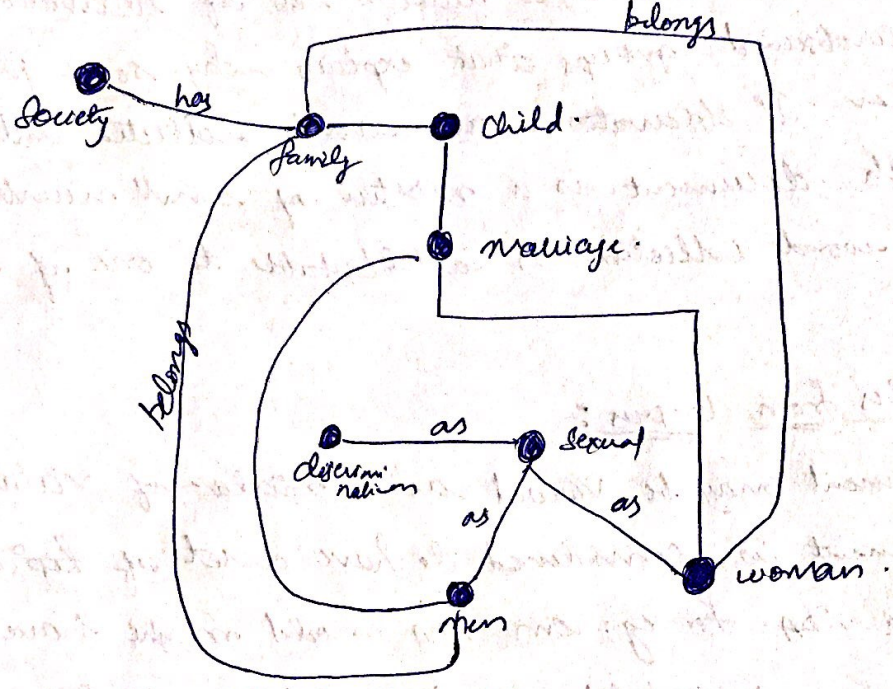
→ Creation of topics from Corpus :-

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 eg: an LDA model might have topics that can be classified as CAT related & DOG - Related. A topic has probabilities of generating various words which can classified and interpreted by the viewer as "CAT - related". And the other related topic like wise has the probabilities of generating each word. Puppy, bark & bone might have high probability.

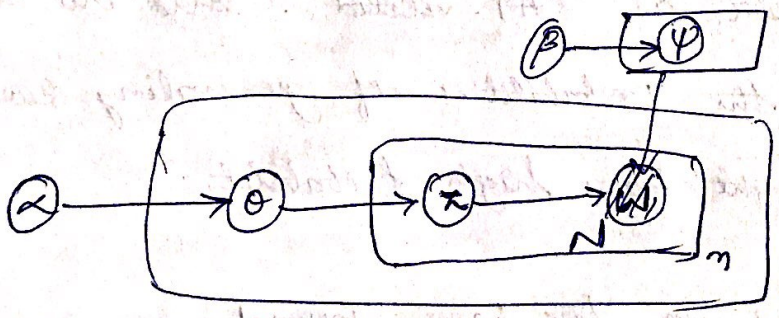
(b) KG for the topic-3 in Yale law Journal :-

If we consider the diagram in the question, there are around 8 topics were displayed. Each topic will be illustrated with its top-most frequent words. Each words position along the X-axis denotes its specificity to the documents.

The most important words which will spread among the X-axis is the topic. 3 are the basis for the Construction of knowledge Graph.



(c) To Determine generality (or) specificity of 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 with in a document.

Revanth Chakilam

Generative process:

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

LDA assumes the following generative process for a corpus D.

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

→ Choose $\varphi_k \sim Dir(\beta)$ where $k \in \{1...K\}$

→ For each word position i,j where $j \in \{1...N_i\}$ and $i \in \{1...M\}$

(D) Inference Algorithm in LDA:

The goal of the topic modelling is to automatically discover the topics from a collection of documents. The document and words are observed. The topic structure is hidden. The topics, per-document topic distribution, we observed variable to infer the hidden structure.

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

Step-①: The algorithm tells how many topics we think are there

②: The algorithm will assign every word to a temporary topic.

③: The algorithm will check and update the topic assignments.

The posterior computation over hidden variables given a document.

$$P[z, \Phi, \theta | w, \alpha\beta] = P(z, \Phi, \theta, w | \alpha, \beta) / P(w) \alpha\beta$$

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

$$\lambda_{kw} = \beta_{kv} + \sum_{d} \sum_{n} I(w_{dn} = v) \varphi_{dnk}$$

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

2Q] K-means Clustering Vs LDA :-

Clustering: Clustering / Segmentation is one of the 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. It method of unsupervised learning, and a common technique for the statistical data analysis and in many fields.

K means clustering: K-means clustering is an algorithm to classify or to group your objects based on attribute / feature into K-number of group. K is positive Integer number. The grouping is done by minimizing the sum of squares of distances b/w data and the corresponding cluster centroid.

(a) Document Text matrix:

In the given figure there are total 10 documents.

→ Distance matrix is also provided. There are 3 Clusters D2, D5, D7 as per the diagram as we get distance as 0.0 for above 3 which indicates that D2, D5, D7 are the centroids. The remaining documents have moved into these 3 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 and based on minimum distance grouping is done :-

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

→ Lets Calculate the distance for D1 from D2, D5, D7

D1 → D2

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

D1 → D5

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

D1 → D7

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

Same way Calculating the distance from each point to centroid.

→ Lets now group the data into clusters based on there minimum distance.

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

Using K-means algorithm we will cluster the data points based on the Centroid and we will re-iterate this process by calculating the new mean and new clusters.

(b) Differences b/w k-means and the LDA are follows:-

→ If both are applied to assign k-topics to a set of N-documents, k-mean is going to partition the N-documents into k-disjoint clusters while LDA assigns a document to a mixture of topics.

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

Advantages: (LDA)

→ LDA is in the exponential family and conjugate to the multinomial distributions.

→ feature set is reduced
→ One document can be associated with multiple topics.

Disadvantages: (LDA)

→ Unable to capture the correlation b/w the different topics.

Advantages: (K-means)

→ 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.
→ The clusters are non-hierarchial and they don't overlap.
→ It is computationally faster.

Disadvantages: (K-means)

→ Difficult to predict k-value.
→ With Global cluster, it didn't work well
→ Doesn't work well with non-circular cluster shape-num
of cluster and initial seed value need to be specified before hand.
→ Applicable only when mean is specified.
→ Sensitive to the outliers.