CS5560 Knowledge Discovery and Management

Problem Set 5

July 3 (T), 2017

Name: Syed Moin

Class ID: 28

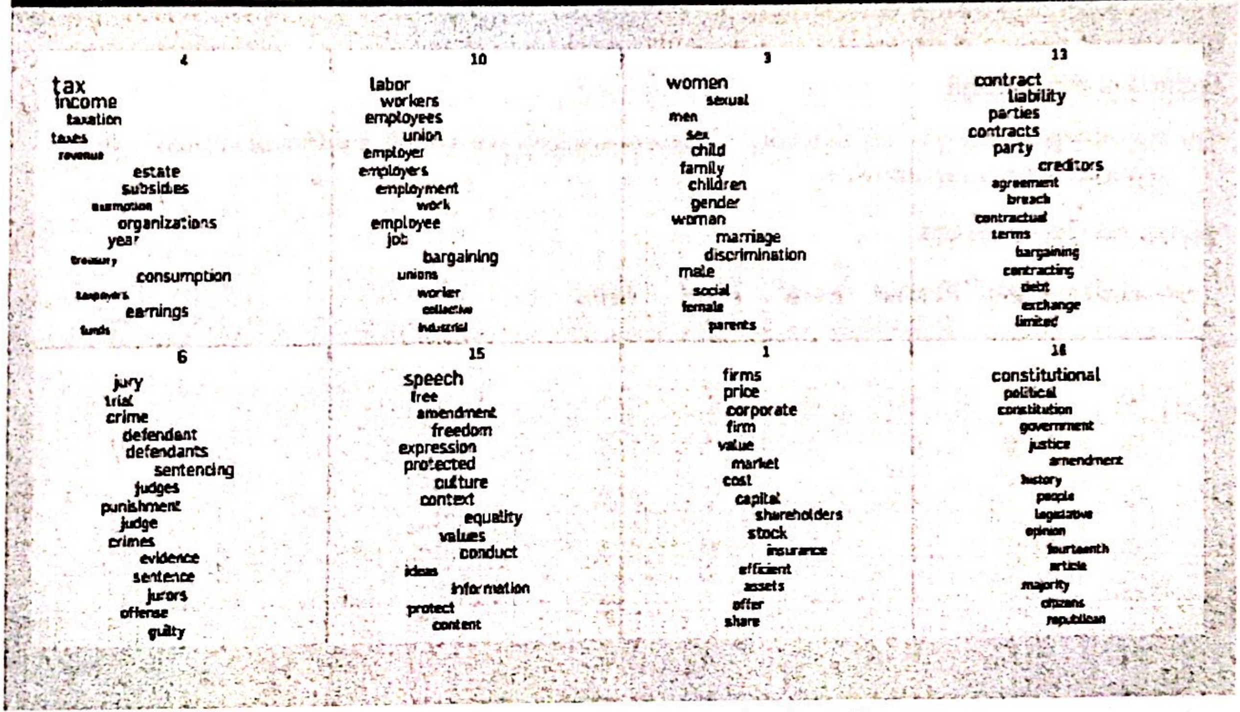
1. LDA

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."

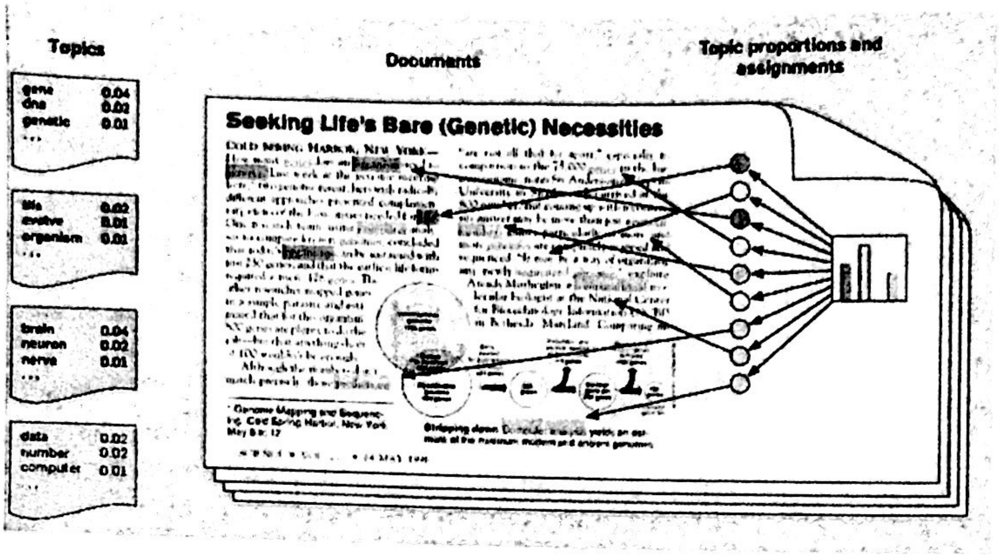


a. Describe the overall process to generate such topics from the corpus.

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

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.

1A) .

Latend Dirichlet Algorithm (LDA) :

a) How to create the topics from the corpus?

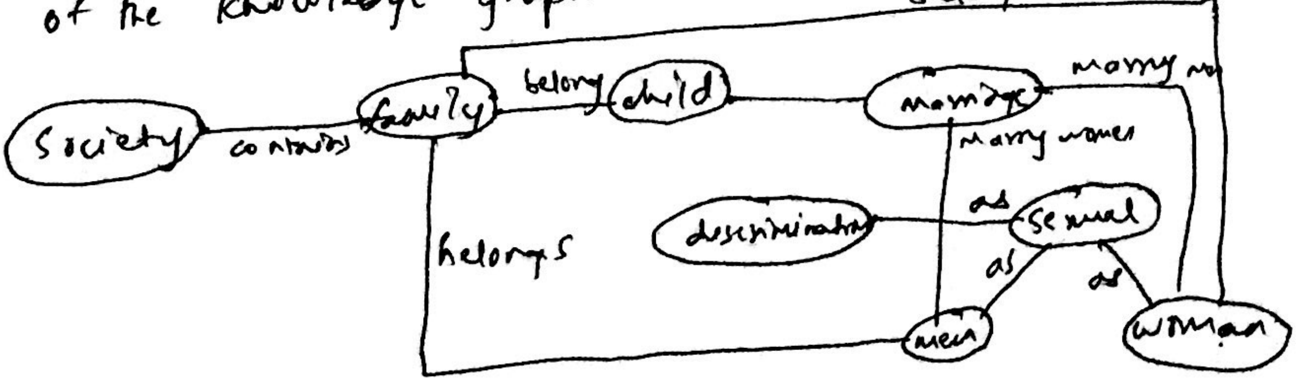
In LDA, each document may be viewer as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA. For example, an LDA model might have topics that can be classified as CAT related and DOG-related.

b) knowledge graph for topic 3 in yale law Journal?

In figure given, in the p's 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.

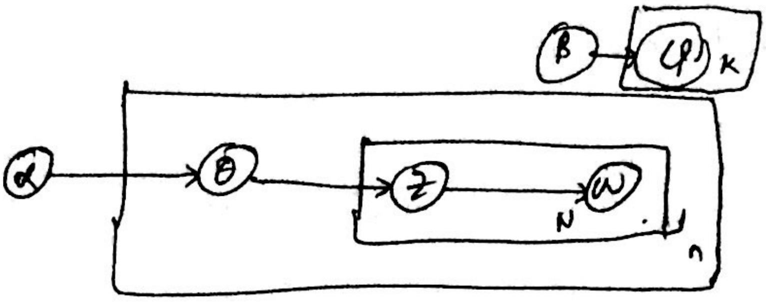
women, sexual, men, sex, child, family, children, gender,
woman, marriage, discrimination, male, social, female, parent.

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



c) Determing 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



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 consisting of M documents each of length N.

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

2) choose $\varphi_K \sim Dir(\beta)$ where $K \in \{1---K\}$

3) for each word positions i,j where $j \in \{1--N_i\}$ and $i \in \{1--M\}$

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

Scanned by CamScanner

d) Inference Algorithm in LDA

The goal of topic modelling 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.

1) You tell the algorithm how many topics we think there are

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

$$p[z, \phi, \theta | w, \alpha, \beta] = p[z, \phi, \theta, w | \alpha, \beta) / p(w | \alpha, \beta)$$

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.

$$\lambda_{kv} = \beta_{kv} + \sum_{d} \sum_{n} I[w_{dn} = v] \vartheta_{dnk}.$$

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

2) Clustering (K-means) vs LDA

a) Given the term/Document Matrix

As shown in figure, there are total 10 documents.

i) Given also the distance matrix. 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 those 3 different clusters using K-means k=3

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

The first raw 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.

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

ii) Now 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

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

iii) Group the data into clusters based on these minimum distance

$D_2 \div \{D_1, D_6, D_9, D_{10}\}$

$D_7 \div \{D_3, D_4\}$

$D_5 \div \{D_8\}$

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 differences between k-means and the LDA 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

LDA pros

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

→ feature set is reduced

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.

→ It is computationally faster

Cons :-

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

→ Applicable only when mean is specified.

→ Sensitive to the outliers.

→ Difficult to predict K-value.