CS5560 Knowledge Discovery and Management

Problem Set 5

July 3 (T), 2017

Name: SUJITHA PUTHANA
Class ID: 24

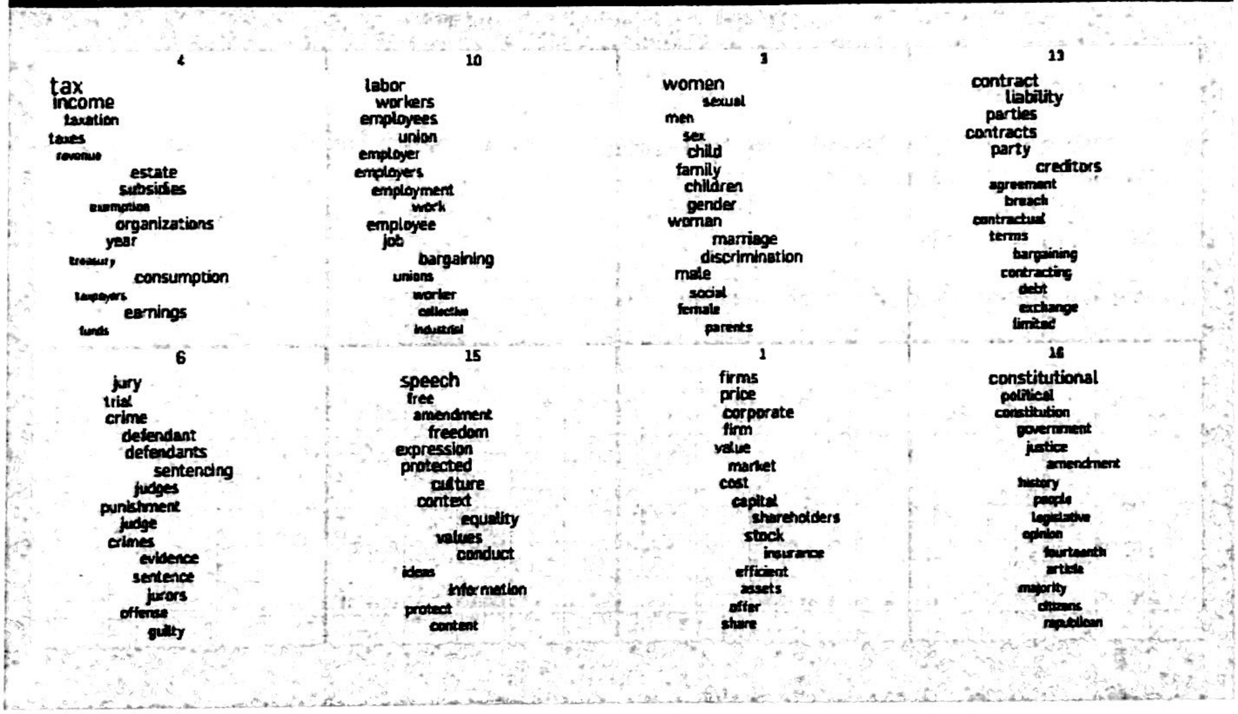
1. LDA

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.

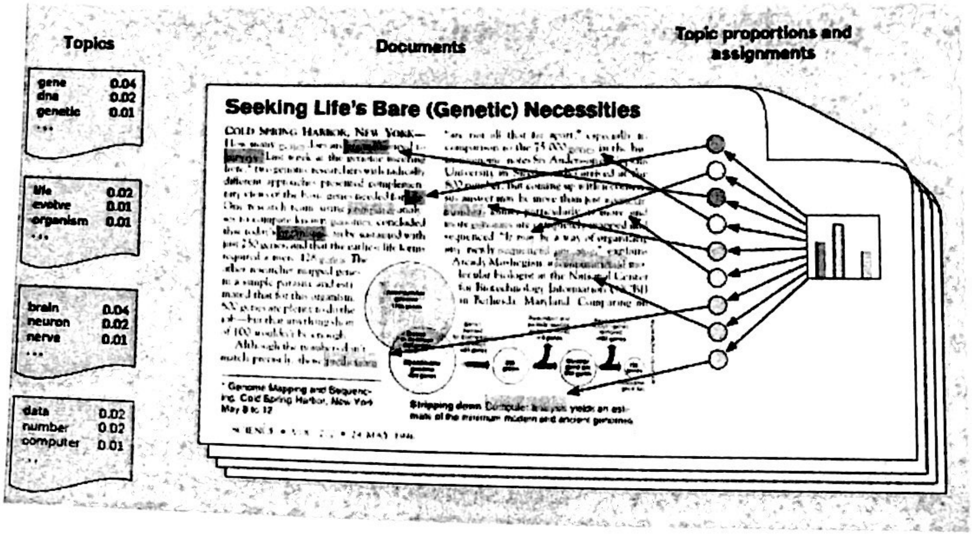
Figure 3. A topic model fit to the *Yale Law Journal*. Here, there are 20 topics (the top eight are plotted). Each topic is illustrated with its topmost frequent words. Each word's position along the x-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific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

b)



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.

(2) Here the no. of clusters $K = 3$.

a) LDA Algorithm

Input : Words $w \in$ documents d

Output : topic assignments z and counts n_d, n_k, w and n_k

begin

randomly initialize z and increment counters

foreach iteration do

for $i = 0 \rightarrow N-1$ do

word $\leftarrow w[i]$

topic $\leftarrow z[i]$

$n_{d,topic}$ -= 1 ; $n_{word,topic}$ -= 1 ; n_{topic} -= 1

for $k = 0 \rightarrow K-1$ do

$$p(z=k|\cdot) = (n_{d,k} + \alpha_k) \frac{n_{k,w} + B_w}{n_k + B \times W}$$

end

topic $\leftarrow$ sample from $p(z|\cdot)$

$z[i] \leftarrow$ topic

$n_{d,topic}$ += 1 ; $n_{word,topic}$ += 1 ; n_{topic} += 1

end

end

return z, $n_{d,k}$, $n_{k,w}$, n_k

end.

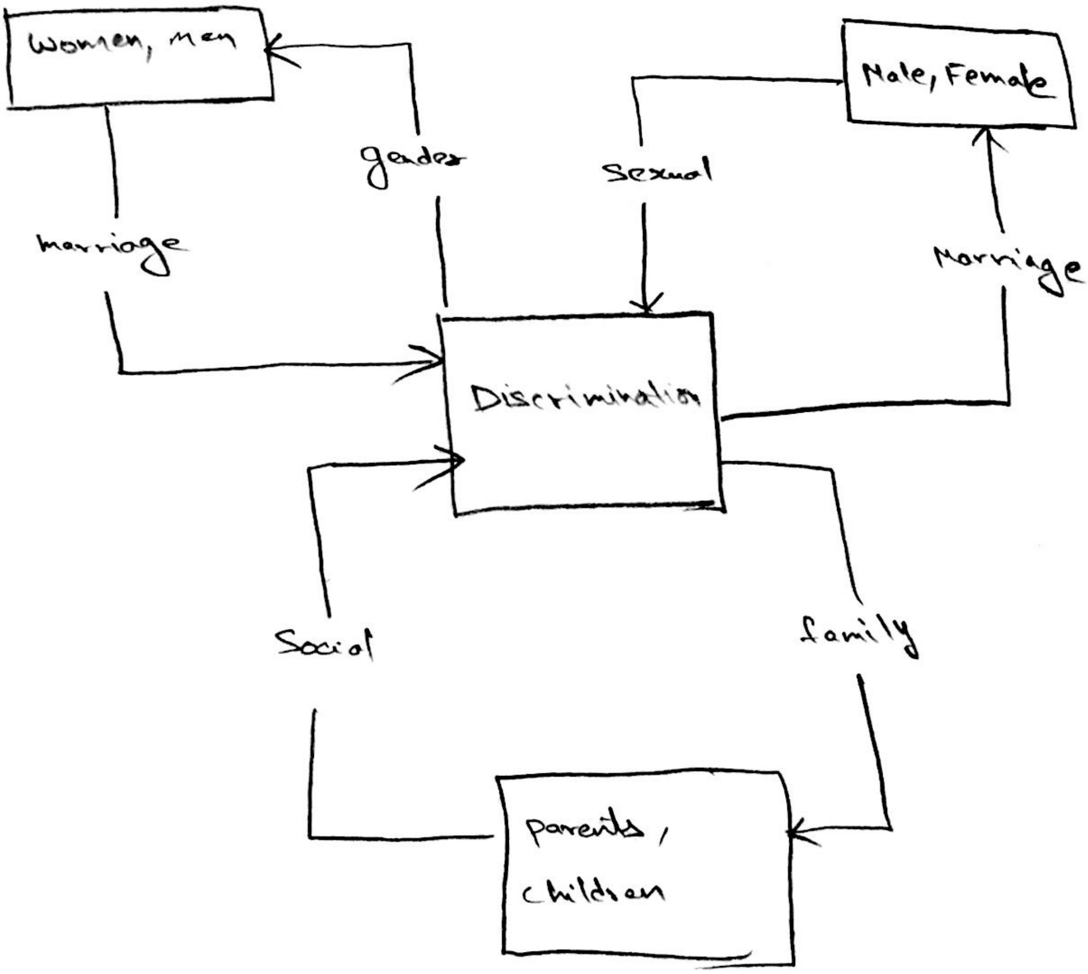
→ LDA algorithm to generate algorithm is a probabilistic iterative algorithm.

→ We should describe "no. of topics to be generated", "The algorithm will assign every word to a temperory topic". "The algorithm will check and again update topics.

b) Knowledge Graph for Topic 3

-> Given the topics are the words related to the "sexual discrimination" context.

-> If we observe in details, all the main reasons and impact of the discrimination are collected as words i.e., topics



-> Here Discrimination would be the main topic so, I am placing it in middle and all the other subtopics are around the main topic

c) In the LDA algorithm were the algorithm will check and update topic assignments, looping through each word in every document. For each word, its topic assignment is updated based on two criteria.

1. How prevalent is the word across topics?
2. How prevalent are topics in the document?

We evaluate performance of the LDA using perplexity. To evaluate the LDA model, one document is taken and split into two. The first half is fed into LDA to compute the topics composition; from that composition, then the word distribution is estimated. This distribution is then compared with the word distribution of the second half of the document. A measure of distance is extracted.

Perplexity is often used to select the best number of topics of the LDA used.

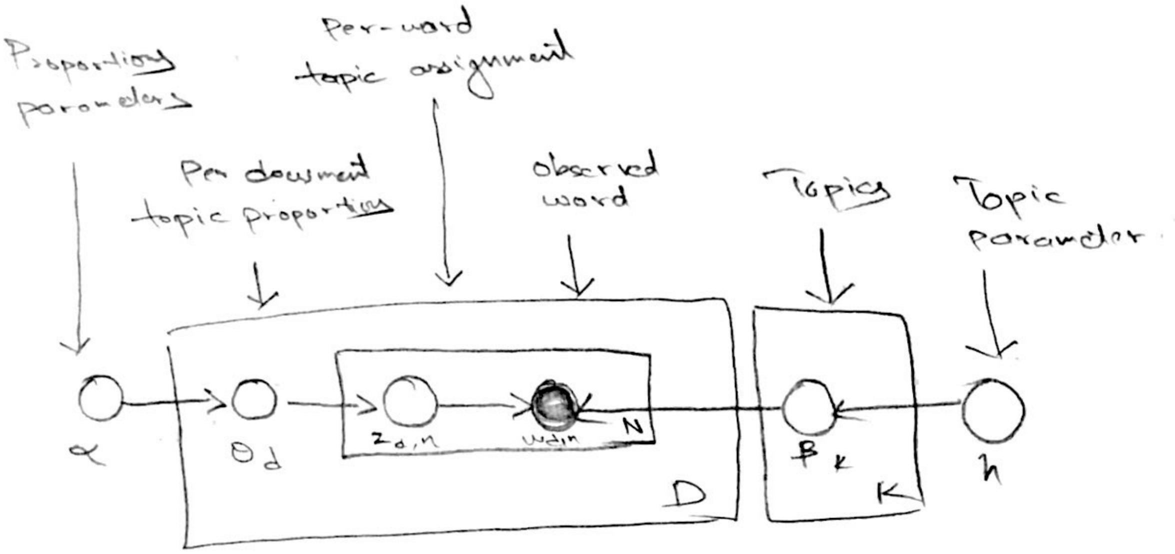
d) -> Each topic is a distribution over words
-> Each document is a mixture of corpus wide topics
-> Each word is drawn from one of these words.
-> We only observe the document.
-> The other structure are hidden variables.
-> Our goal is to infer the hidden variables i.e., compute their distribution conditioned on the document

P (topics, proportions, assignments | document)

→ Encode assumption.

→ Define a factorization of the joint distribution.

→ Connect to algorithm to compute with data.



$$p(\beta, \theta, Z, W) = \left(\prod_{i=1}^{K} p(\beta_i|\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}|\theta_d)\, p(w_{d,n}|\beta_{1:K}, z_{d,n})\right)$$

→ This joint defines a posterior, $p(\theta, Z, \beta | W)$

→ From a collection of documents, infer

- per word topic assignment $z_{d,n}$.
- per corpus topic distribution β_K
- Per document topic proportions θ_d

(2) Here the no. of clusters $K = 3$.

-> Let D2, D5, D7 be the three seeds.

-> Next we have to calculate Euclidean distance of other documents from D2, D5 & D7.

O -> online, F -> festival, B -> book, T -> Flight, D -> Delhi

-> D_1 to $D_2 = \sqrt{(O_1-O_2)^2 + (F_1-F_2)^2 + (B_1-B_2)^2 + (T_1-T_2)^2 + (D_1-D_2)^2}$

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

-> D_1 to $D_5 = \sqrt{(1-3)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{7} \simeq 2.6$

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

-> D_2 to $D_2 = 0$

-> D_2 to $D_5 = \sqrt{(2-3)^2 + (1-1)^2 + (2-0)^2 + (1-0)^2 + (1-0)^2} = \sqrt{7} \simeq 2.6$

-> D_2 to $D_7 = \sqrt{(2-2)^2 + (1-0)^2 + (2-1)^2 + (1-2)^2 + (1-1)^2} = \sqrt{3} = 1.7$

-> D_3 to $D_2 = \sqrt{6} = 2.4$

-> D_4 to $D_2 = \sqrt{8} = 2.8$

-> D_7 to $D_7 = 0$

-> D_3 to $D_5 = \sqrt{13} = 3.6$

-> D_4 to $D_5 = \sqrt{9} = 3$

-> D_7 to $D_2 = \sqrt{3} = 1.7$

Documents	D_2	D_5	D_7	Minimum	Clusters
D_1	2.0	2.1	2.2	2.0	D_2
D_2	0.0	2.6	1.7	0.0	D_2
D_3	2.4	3.6	2.2	2.2	D_7
D_4	2.8	3.0	2.6	2.6	D_7
D_5	2.6	0.0	2.8	0.0	D_5
D_6	2.4	3.9	2.6	2.4	D_2
D_7	1.7	2.8	0.0	0.0	D_7
D_8	2.6	2.0	2.8	2.0	D_5
D_9	2.0	3.0	3.6	2.0	D_2
D_{10}	2.2	3.5	2.4	2.2	D_2

→ D_2 Cluster – $\{D_1, D_2, D_6, D_9, D_{10}\}$

→ D_5 cluster – $\{D_5, D_8\}$

→ D_7 cluster → $\{D_3, D_4, D_7\}$

b) K-means Clustering

Pros :-

1. Computational cost → O(K * n * d) ⇒ fast, robust
2. Easier to understand.
3. Gives best result when data set are distinct or well seperated from each other.

4. Works great for spherical clusters
5. It is a great resolution for pre-clustering

Cons

1. Does not work well with clusters of different size and different density.
2. K-value is difficult to predict

LDA topic :-

Cons

① Difficult to predict the no. of topics to be generated.
② Efficiency is low than Machine learning algorithms.
③ LDA cannot capture correlations
④ Uses bow i.e, assumes that words are exchanged.

Pros

① Each topic proportions can be easily derived.
② We can infer the content speed of each sentence by a word count.