CS5560 Knowledge Discovery and Management

Problem Set 5

July 3 (T), 2017

Name: Rashmi

Class ID: 29

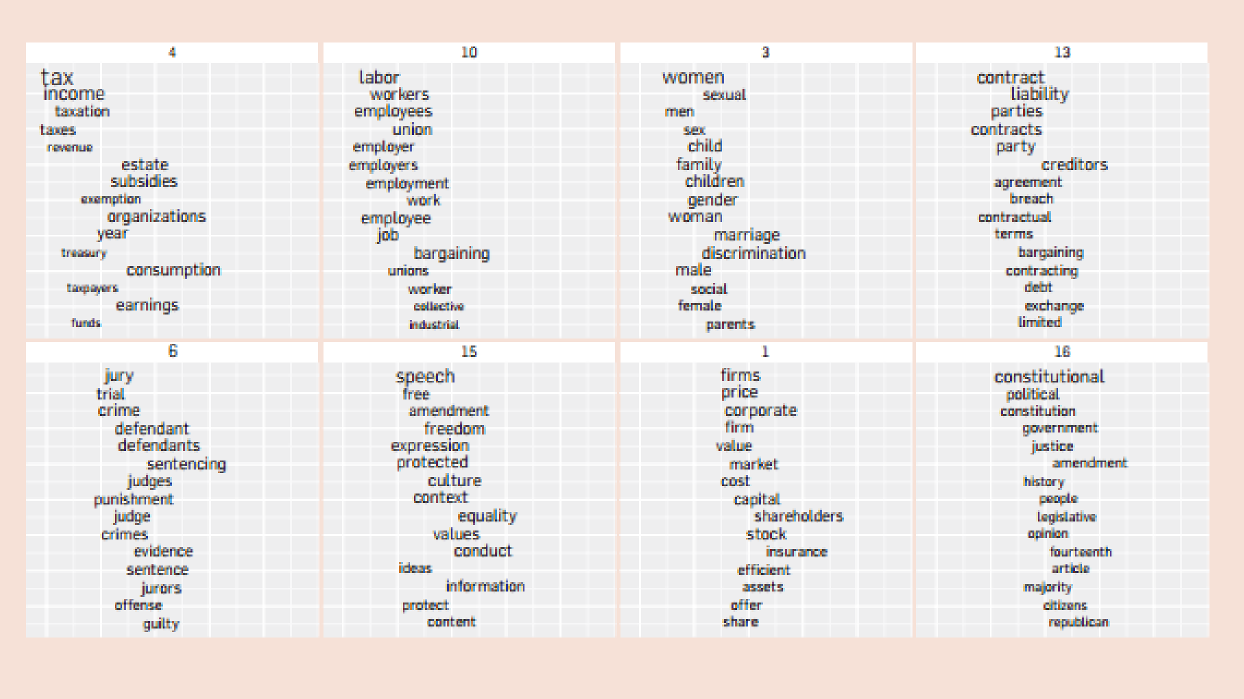
1. LDA

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.**

Latent Dirichlet allocation (LDA) is a technique that automatically discovers topics that these documents contain. Following are the steps for the same:

Step 1)First tell the algorithm how many topics you think there are.

We can either use results from previous analysis(already concluded estimates) or trial an error. The main aim is that the topic generated topic to our desired level or yielding the highest likelihood.

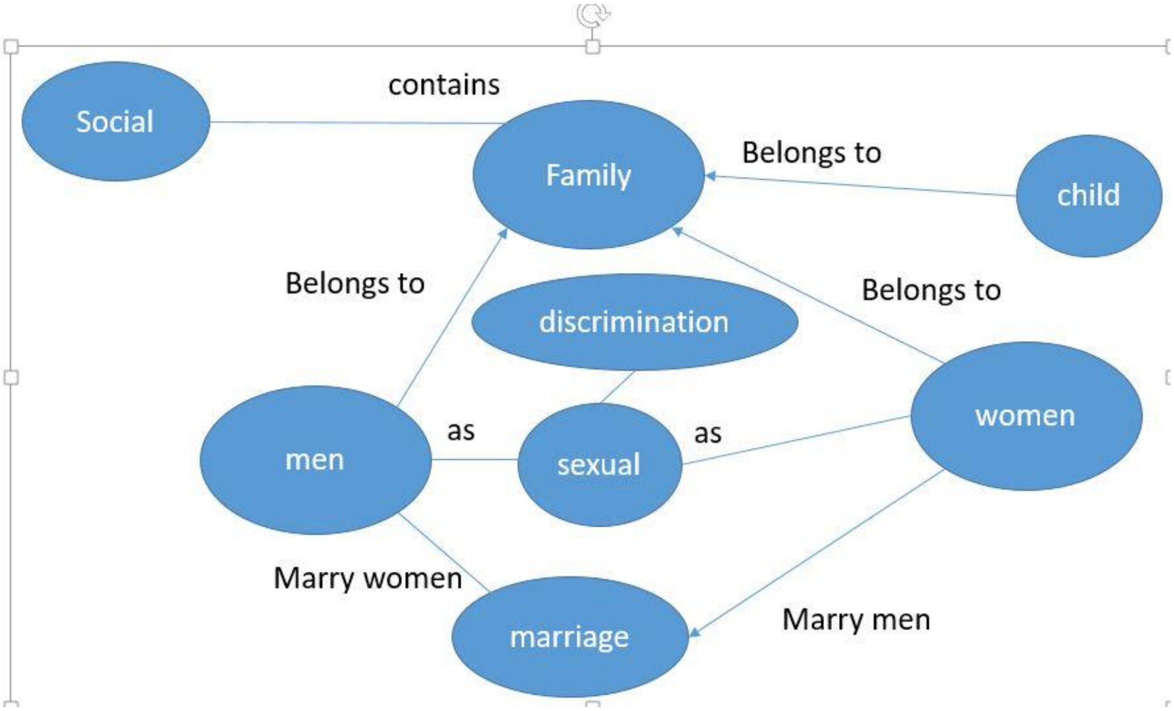
Step 2)The algorithm will assign every word to a temporary topic.

It is temporary and will be updated in step 2. Temporary topics are assigned in semi-random manner which means that even the same word may belong to different documents. We also remove unwanted stop words like the, and, my before topic allocation.

Step 3 (iterative) The algorithm check and update topic assignments, looping through each word in every document. For each word, its topic assignment is updated based on two criteria:

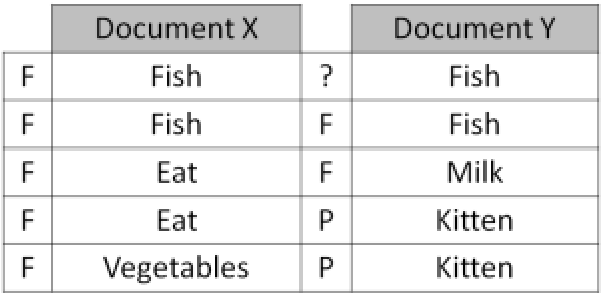
- How prevalent is that word across topics?
- How prevalent are topics in the document?

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.**

Suppose the documents are as follows:

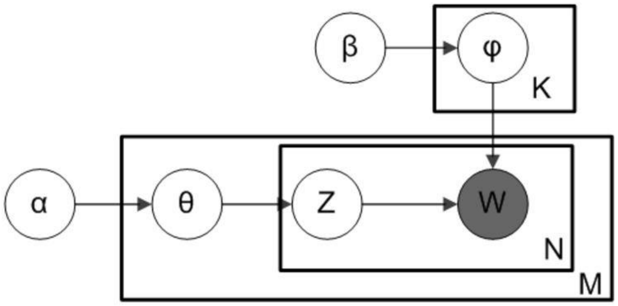


	Document X		Document Y
F	Fish	?	Fish
F	Fish	F	Fish
F	Eat	F	Milk
F	Eat	P	Kitten
F	Vegetables	P	Kitten

Generality: We first check how much is word let us day w appearing in both the documents. Since “fish” words across both documents comprise nearly half of remaining Topic F words but 0% of remaining Topic P words, a “fish” word picked at random would more likely be about Topic F.

Specificity: Since the words in Doc Y are assigned to Topic F and Topic P in a 50-50 ratio, the remaining “fish” word seems equally likely to be about either topic.

d. **Describe the inference algorithm that was used in LDA.**
One of the most advanced algorithms for doing topic-modelling is **Latent Dirichlet Allocation** (or **LDA**). This is a probabilistic model developed by Blei, Ng and Jordan in 2003. LDA is an iterative algorithm which requires only three parameters to run: when they’re chosen properly, its accuracy is pretty high



Parameters of the model:

- Boxed:
  - K is the number of topics
  - N is the number of words in the document
  - M is the number of documents to analyse
- α is the Dirichlet-prior concentration parameter of the per-document topic distribution
- β is the same parameter of the per-topic word distribution
- φ(k) is the word distribution for topic k
- θ(i) is the topic distribution for document i
- z(i,j) is the topic assignment for w(i,j)
- w(i,j) is the j-th word in the i-th document

φ and θ are Dirichlet distributions, z and w are multinomials.

Assuming one has K topics, a corpus D of M = |D| documents, and a vocabulary consisting of V unique words:

```
• For $j \in [1, \ldots, M]$,
    – $\theta_j \sim Dirichlet(\alpha)$
    – For $t \in [1, \ldots, |d_j|]$
        * $z_{j,t} \sim Multinomial(\theta_j)$
        * $w_{j,t} \sim Multinomial(\phi_{z_{j,t}})$
```

In words, this means that there are K topics φ1,...,K that are shared among all documents, and each document dj in the corpus D is considered as a mixture over these topics, indicated by θj . Then, we can generate the words for document dj by first sampling a topic assignment zj,t from the topic proportions θj , and then sampling a word from the corresponding topic φzj,t . zj,t is then an indicator variable that denotes which topic from 1, . . . K was selected for the t-th word in dj .

It is important to point out some key assumptions with this model. First, we assume that the number of topics K is a fixed quantity known in advance, and that each φk is a fixed quantity to be estimated. Furthermore, we assume that the number of unique words V fixed and known in advance (that is, the model lacks any mechanisms for generating “new words”). Each word within a document is independent (encoding the traditional “bag of words” assumption), and each topic proportion θj is independent. In this formulation, we can see that the joint distribution of the topic mixtures Θ, the set of topic assignments Z, and the words in the corpus W given the hyperparameter α and the topics Φ is given by

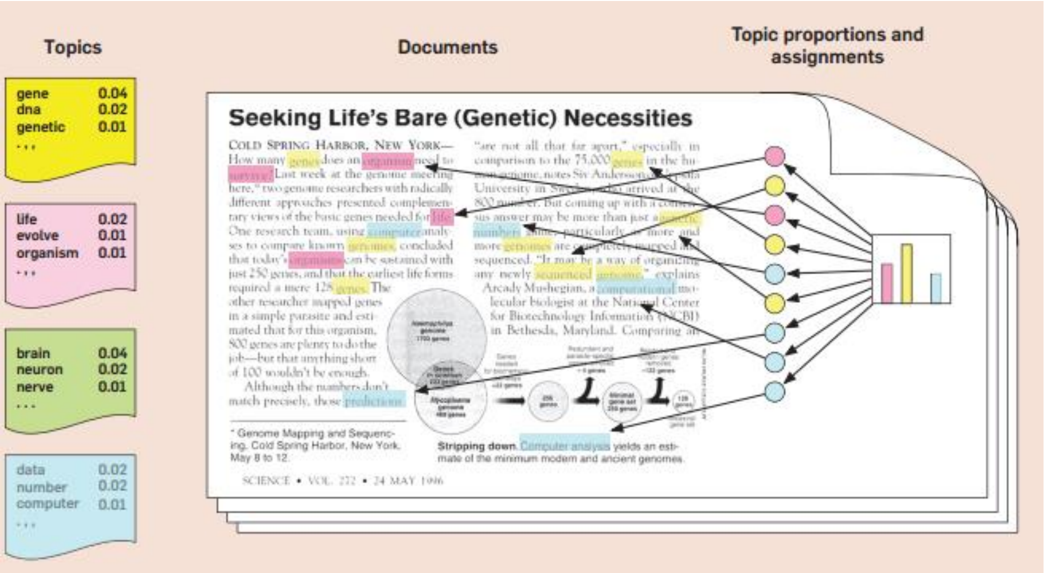
$$P(\mathbf{W}, \mathbf{Z}, \Theta \mid \alpha, \Phi) = \prod_{j=1}^{M} P(\theta_j \mid \alpha) \prod_{t=1}^{N_j} P(z_{j,t} \mid \theta_j) P(w_{j,t} \mid \phi_{z_{j,t}}).$$

Weakness : it does not also place a prior on the each φk—since this quantity is not modeled in the machinery for inference, it must be estimated using maximum likelihood.

Choosing another Dirichlet parameterized by β as the prior for each φk , the generative model becomes:

1. For $i \in [1, \ldots K]$, $\phi_i \sim Dirichlet(\beta)$
2. For $j \in [1, \ldots, M]$,
   - $\theta_j \sim Dirichlet(\alpha)$
   - For $t \in [1, \ldots, |d_j|]$
     - $z_{j,t} \sim Multinomial(\theta_j)$
     - $w_{j,t} \sim Multinomial(\phi_{z_{j,t}})$

$$P(\mathbf{W}, \mathbf{Z}, \Theta, \Phi \mid \alpha, \beta) = \prod_{i=1}^{K} P(\phi_k \mid \beta) \times \prod_{j=1}^{M} P(\theta_j \mid \alpha) \prod_{t=1}^{N_j} P(z_{j,t} \mid \theta_j) P(w_{j,t} \mid \phi_{z_{j,t}}).$$



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

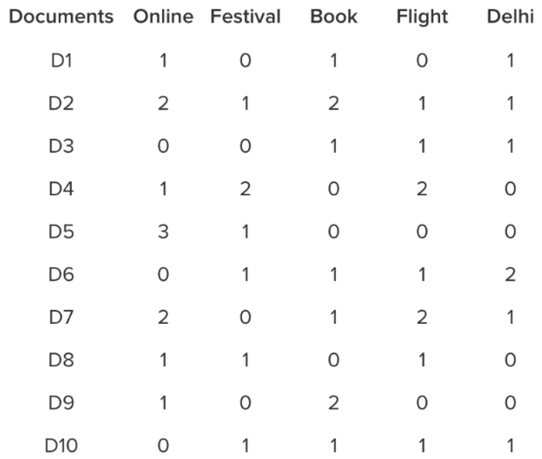
SCIENCE • VOL. 272 • 24 MAY 1996

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).

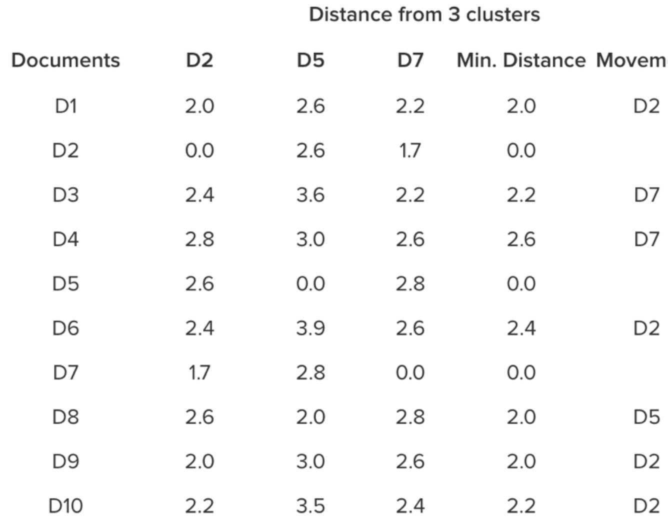
Below is the document term matrix for this dataset. It shows for how many times one word has appeared in the document. For example, in document 1 (D1), the words *online, book* and *Delhi* have each been mentioned once



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

- First, three seeds should be chosen. Suppose, D2, D5 & D7 are chosen as initial three seeds.
- The next step is to calculate the Euclidean distance of other documents from D2, D5 & D7.
- Assuming: U=Online, V= Festival, X=Book, Y=Flight, Z=Delhi. Then the Euclidean distance between D1 & D2 would be:

  ((U1-U2)^2 + (W1-W2)^2+(X1-X2)^2+ (Y1-Y2)^2+(Z1-Z2)^2)^0.5



		Distance from 3 clusters			
Documents	D2	D5	D7	Min. Distance	Movem
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

Clusters	**# of Observations**
D2	5
D5	2
D7	3

Hence, 10 documents have moved into 3 different clusters. Instead of Centroids, Medoids are formed and again distances are re-calculated to ensure that the documents who are closer to a medoid is assigned to the same cluster.

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

K means	**LDA**
unsupervised learning algorithms	unsupervised learning algorithms
user needs to decide a priori the parameter K, respectively the number of clusters	user needs to decide a priori the parameter K, respectively the number of topics
K-means is going to partition the N documents in K disjoint clusters	LDA assigns a document to a mixture of topics.
Each document is characterized by just one topic/cluster	Each document is characterized by one or more topics (e.g. Document D belongs for 60% to Topic A, 30% to topic B and 10% to topic E).
can give less realistic results	can give more realistic results

References

http://times.cs.uiuc.edu/course/598f16/notes/lda-survey.pdf