Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Master of Science (MSc)

Applied Information and Data Science

Institut für Kommunikation und Marketing IKM
Dr. Manuel Dömer
Externer Dozent

T direkt +41 79 551 64 17
manuel.doemer@hslu.ch

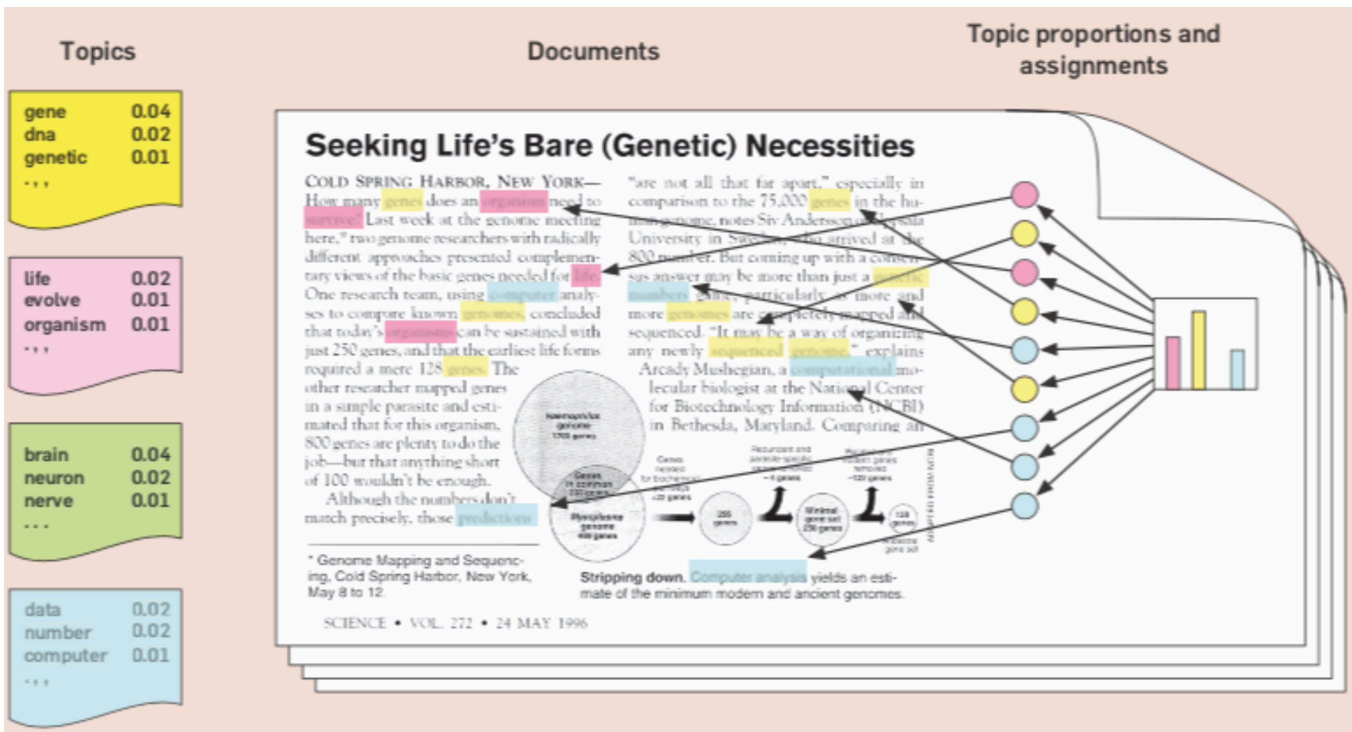
Luzern 07.05.20

Computational Language Technologies

TOPIC MODELLING

Hidden Topics

Assume the documents are composed my combining various hidden topics. Each document can cover various topics and the same topic can appear in different documents.



Blei D. M.: Probabilistic topic models, communications of the ACM vol 55, p.77 ***2012***

Topic

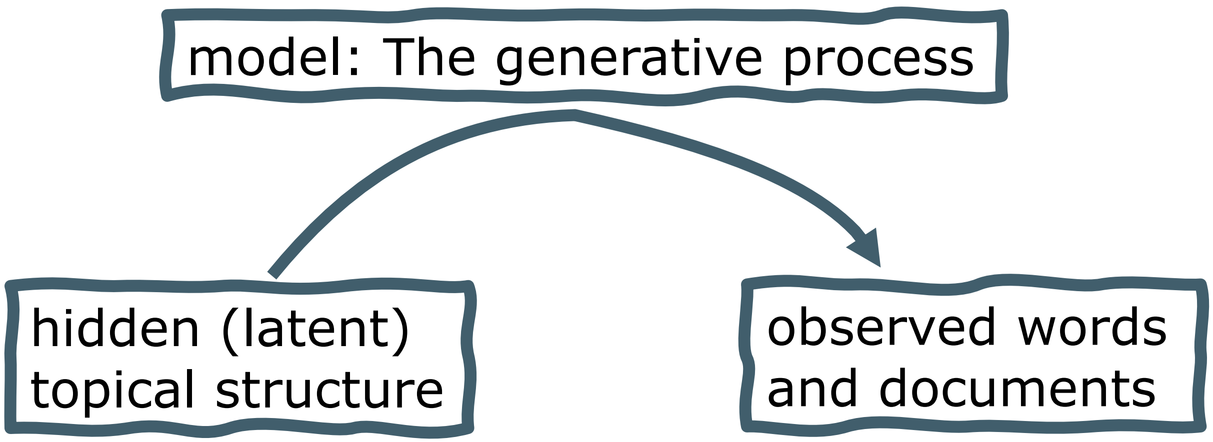
- Not strongly defined
- broad concept/theme, presented in a semantically coherent form
- e.g. politics, sports, technology, entertainment etc.

Topic in the context of probabilistic language modelling:

- the hidden structure of the text
- identified by the likelihood of word co-occurrence over a fixed vocabulary
- bag of words: order is not important
- A word may occur in several topics with different probability and different distribution of neighboring words

Probabilistic Topic Models

assume that text is generated by sampling from multinomial distributions of words, which correspond to the hidden (latent) topics:



Topic: a multinomial distribution over words

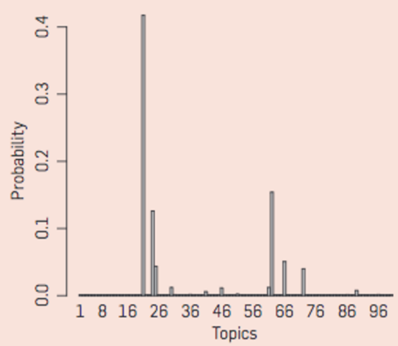
Document: joint distribution over topics and words

The ultimate goal is then, to infer the topics from the observed text based on a given model.

Applications of Topic Modelling

Answer topic-related questions by computing various kinds of posterior distributions, such as follow topic distribution over time, determine sentiment across topic distribution etc.

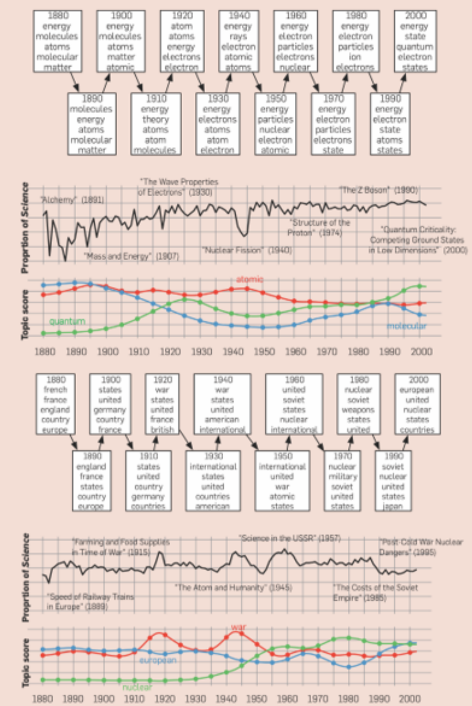
Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Blei D. M.: Probabilistic Topic Models, Communications of the ACM, vol. 55 (4), p.77 (2012)

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.



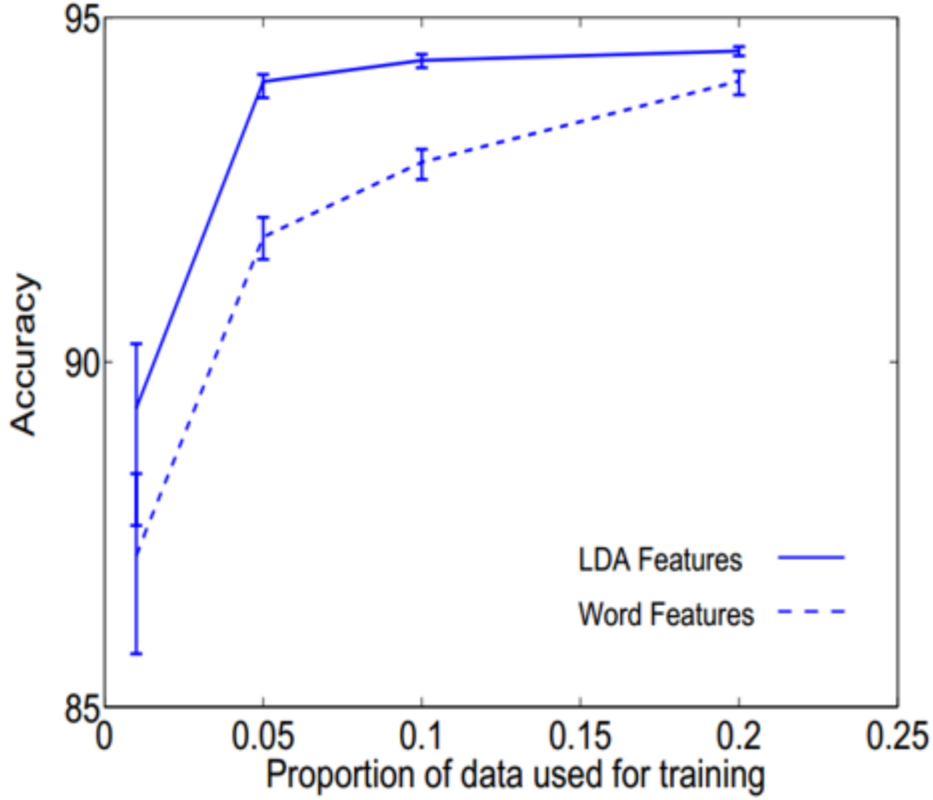
Applications of Topic Modelling

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."

Blei D. M.: Probabilistic topic models, communications of the ACM vol 55, p.77 ***2012***

Applications of Topic Modelling

Use it for dimensionality reduction prior to document classification to produce better/more stable features



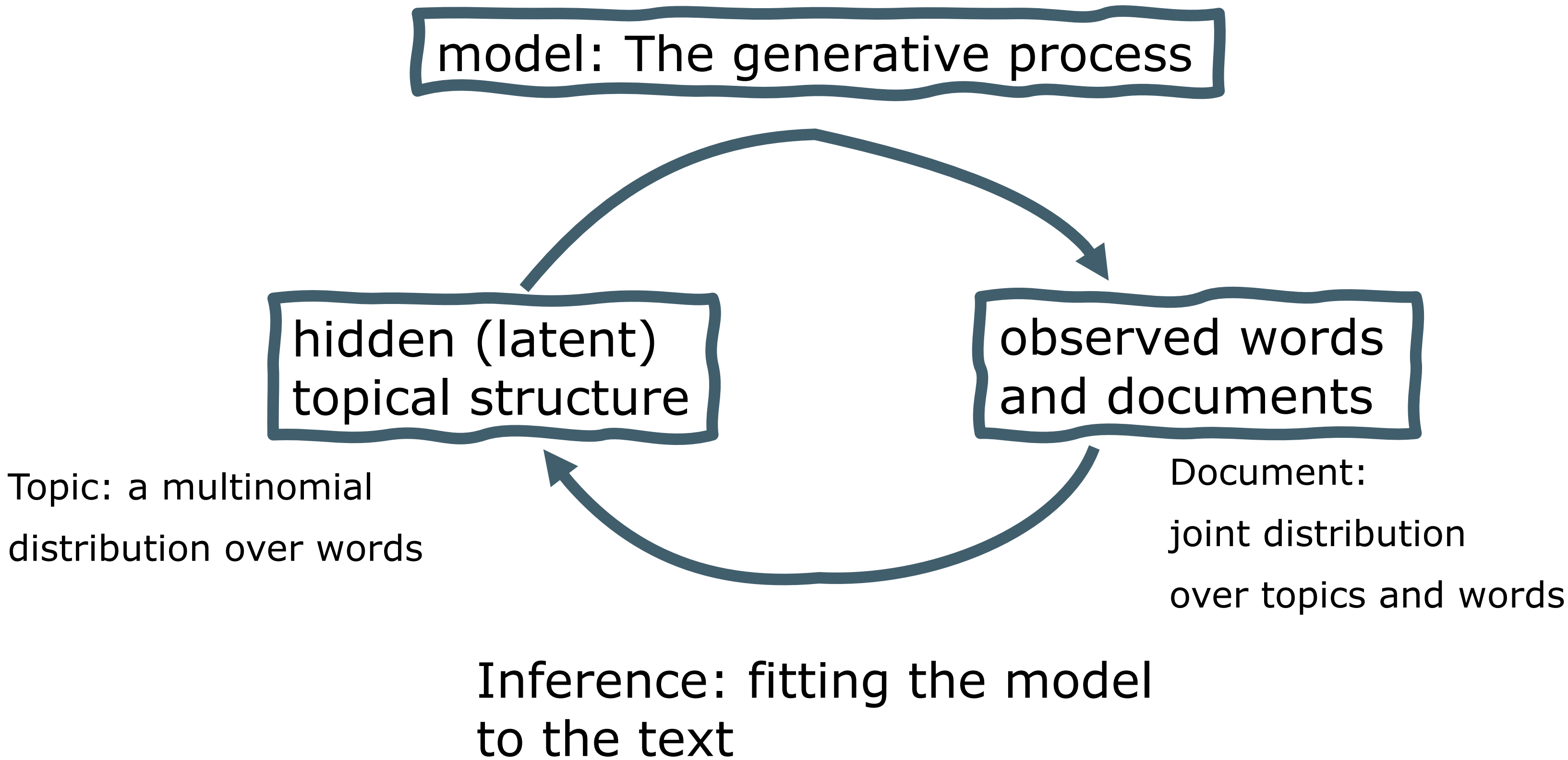
LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation (LDA)

- Generative Model
- David Blei, Andrew Ng, and Michael I. Jordan (2003), Journal of Machine Learning Research 3 993-1022
- Generalisation of the probabilistic latent semantic analysis (pLSA)

- Literature Recommendation:
  Steyves M., Griffiths T.: Probabilistic Topic Models in *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum (2007)

Latent Dirichlet Allocation

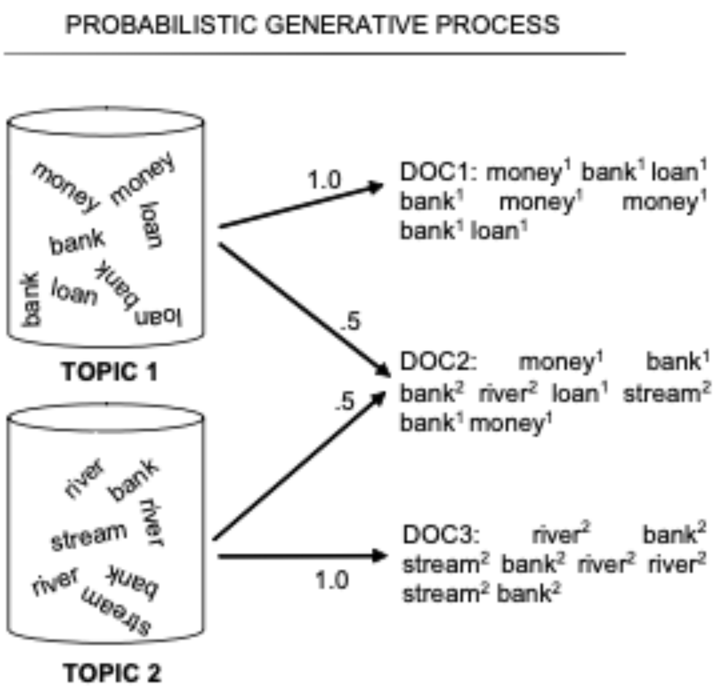


LDA – The Generative Model

Topics are multinomial distributions over words. Documents are mixtures of topics generated according to the following procedure:

1. For each topic determine a multinomial distribution over words
2. Decide on the number of words for the document
3. Choose a distribution of topics for the document
4. Generate each token in the document by:
    1. First picking a topic (from the distribution chosen in 2.)
    2. Using the topic to generate the word itself (according to the topic’s multinomial distribution).

LDA is a Bag-of-Words Model



LDA – The Generative Model

K: number of topics

V: Size of vocabulary

$\mathrm{Dir}_K(\vec{\alpha})$: K dimensional Dirichlet

$\mathrm{Dir}_V(\vec{\beta})$: V dimensional Dirichlet

Word-to-topic distribution

	Topic 0	…	Topic k	…	Topic K
word 0					
word 1					
…					
word v					
…					
word V					

probability of each word per topic

ϕ_k

Doc-to-topic distribution

	Topic 0	…	Topic k	…	Topic K
doc 0					
doc 1					
…					
doc d					
…					
doc D					

θ_d

pick a topic according to its probability

1) For each topic k
   a) Draw a distribution over words $\boxed{\overrightarrow{\phi_k} \sim \mathrm{Dir}_V(\vec{\beta})}$
2) For each document d
   a) draw a distribution over topics $\boxed{\overrightarrow{\theta_d} \sim \mathrm{Dir}_K(\vec{\alpha})}$
   b) for each token position i in the document:
      i) draw a topic assignment $Z_{d,i} \propto \mathrm{Mult}(\overrightarrow{\theta}_d), Z_{d,i} \in \{1, ...K\}$
      ii) draw a word $W_{d,i} \propto \mathrm{Mult}(\overrightarrow{\phi}_{z_{d,i}}), W_{d,i} \in \{1, ...V\}$

look up topic Z and pick word according to its probability

Why Dirichlet?

- The Dirichlet distribution is the conjugate prior of the multinomial distribution
- If the prior distribution of the multinomial parameters is Dirichlet

$$(p_1, \dots p_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$$

then the posterior distribution is also a Dirichlet distribution (with parameters different from those of the prior)

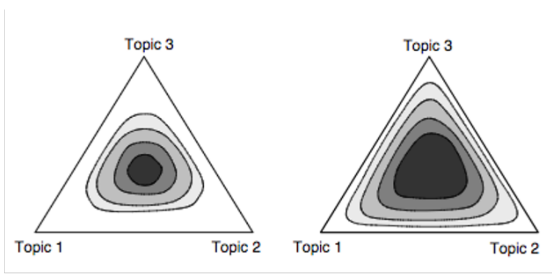
$$(p_1, \dots p_K) | (x_1, \dots x_K) \sim \text{Dirichlet}(\alpha_1 + x_1, \dots, \alpha_K + x_K)$$

so we directly have access to the posterior distributions from the observations

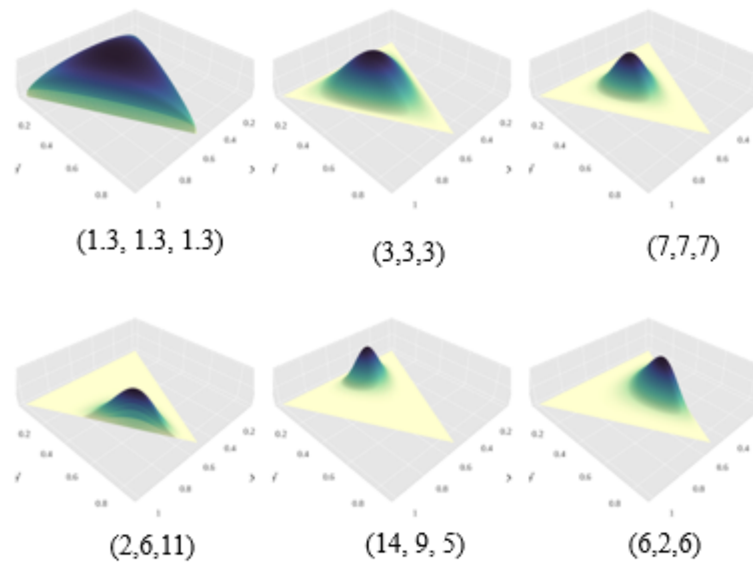
The Dirichlet Distribution

The probability density of a K dimensional Dirichlet distribution

$$\mathrm{Dir}(\alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1}$$

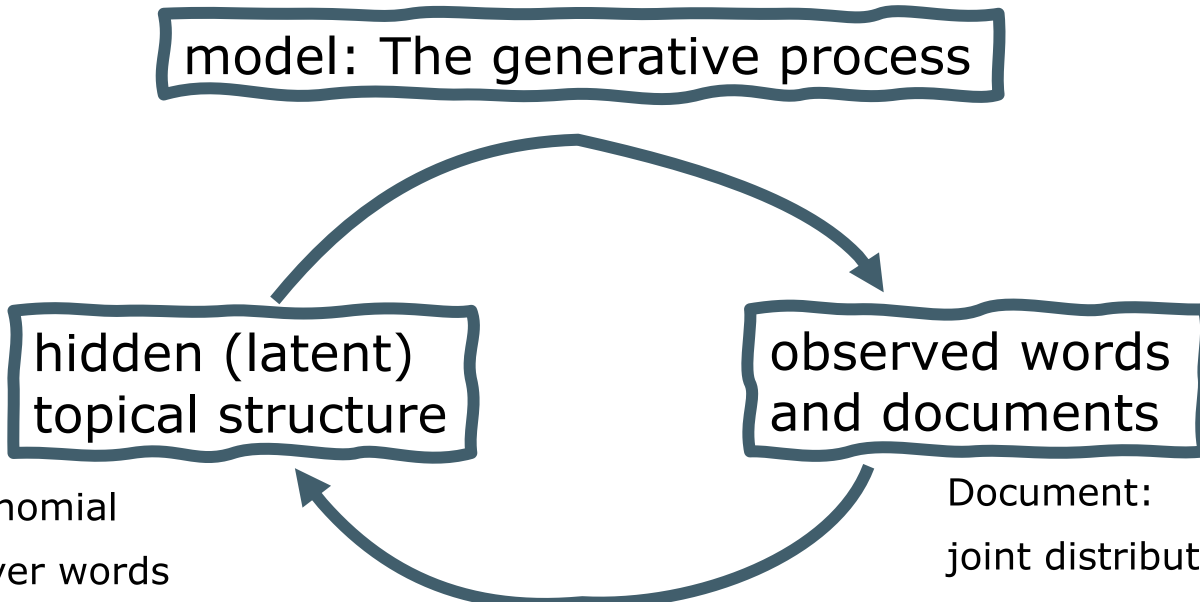


Steyvers et al: Probabilistic Topic Models,LatentSemantic Analysis: A Road to I



https://commons.wikimedia.org/wiki/File:Dirichlet-3d-panel.png

Inference

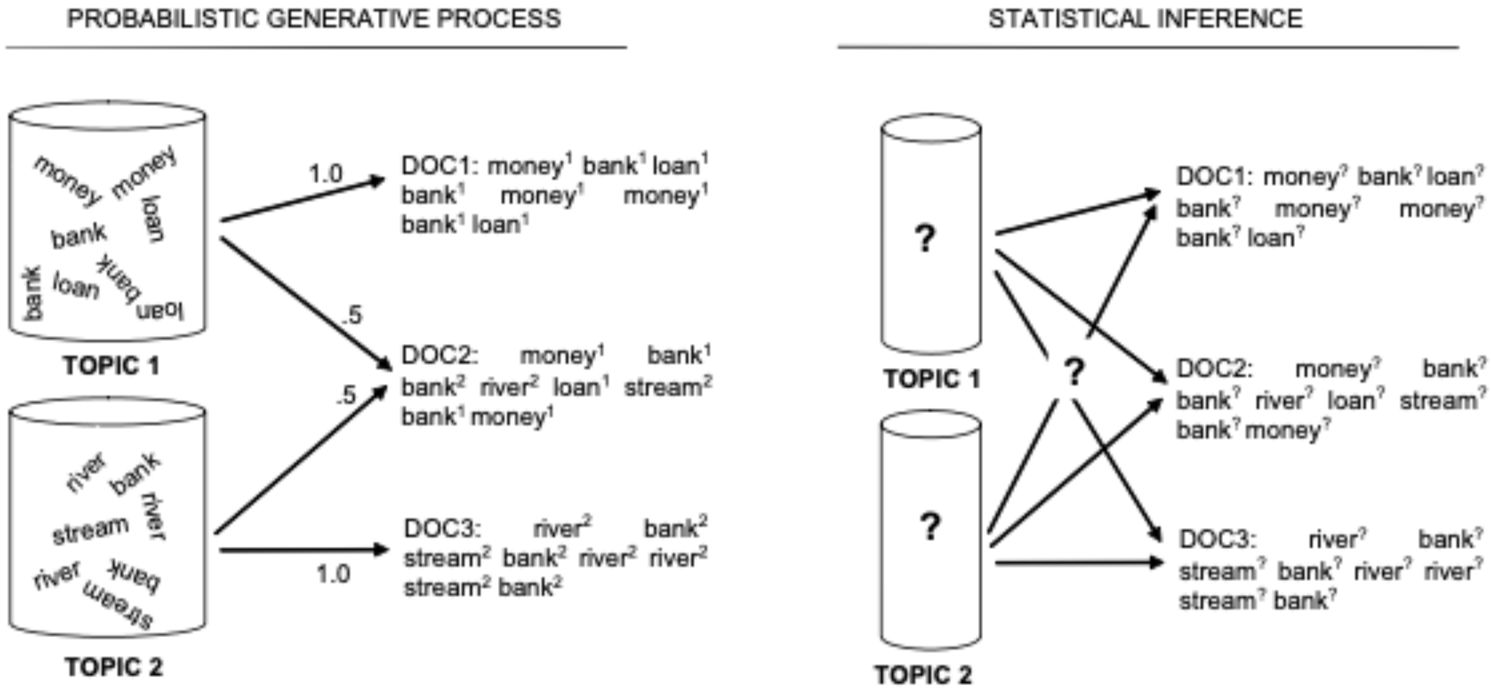


Topic: a multinomial distribution over words

Document: joint distribution over topics and words

Inference: fitting the model to the text
Given the observed documents and tokens and assuming they were produced by the generative process defined previously, LDA now provides an approximate algorithm to deduce the set of topics that are likely to have generated them

Inference



PROBABILISTIC GENERATIVE PROCESS
STATISTICAL INFERENCE
TOPIC 1
TOPIC 2
1.0
.5
.5
1.0
DOC1: money¹ bank¹ loan¹ bank¹ money¹ money¹ bank¹ loan¹
DOC2: money¹ bank¹ bank² river² loan¹ stream² bank¹ money¹
DOC3: river² bank² stream² bank² river² river² stream² bank²
?
DOC1: money? bank? loan? bank? money? money? bank? loan?
DOC2: money? bank? bank? river? loan? stream? bank? money?
DOC3: river? bank? stream? bank? river? river? stream? bank?

Inference – Commonly used Algorithms

- Deterministic approximation
  - Variational inference
- Markov Chain Monte Carlo
  - Collapsed Gibbs Sampling

Collapsed Gibbs Sampling – The Count Matrices

The probability distributions are estimated from the counts of the topic assignments to the tokens among all the documents of the corpus

C^{VK} the word-topic counts matrix
C^{VK}_{vk}: the number of times a token of word v is assigned to topic k

	Topic 0	...	Topic k	...	Topic K
word 0					
word 1					
...					
word v					
...					
word V					

C^{DK} document-topic counts matrix
C^{DK}_{dk}: number of times topic k is assigned to some token in document d

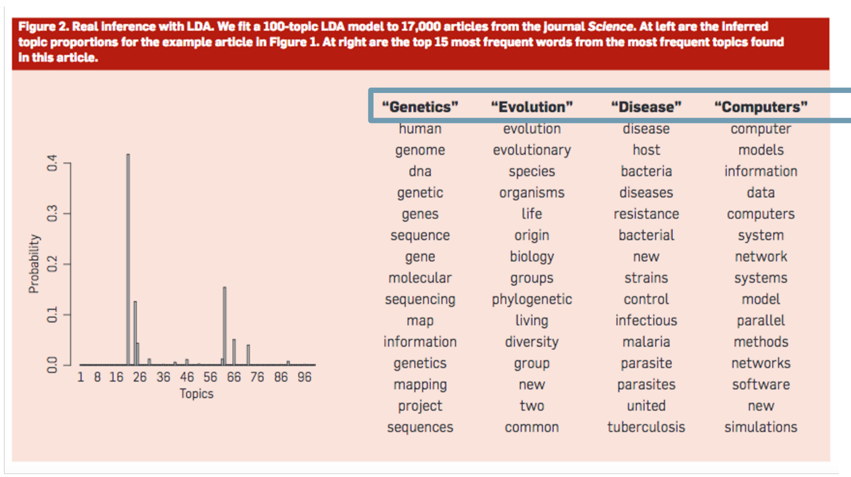
	Topic 0	...	Topic k	...	Topic K
doc 0					
doc 1					
...					
doc d					
...					
doc D					

Collapsed Gibbs Sampling – High Level Description

- Go through each document, and randomly assign each word in the document to one of the K topics -> initial guess for the document-topic and word-topic distributions
- Improve iteratively the document-topic and word-topic distributions:
  until convergence criteria satisfied:
  for each document d…
  - Go through each word token t_i (corresponding vocabulary word v) in d…
    - compute:
      1) p(topic k|document d) = the proportion of tokens in document d that are currently assigned to topic k
      2) p(word v|topic k) = the proportion of assignments to topic k over all documents and tokens with corresponding vocabulary word v.
    - Re-assign token t_i a new topic $z_i = l$, where we choose topic l with probability p(word v|topic l) * p(topic l|document d)

$$P(z_i = l | z_{-i}, v, d, \cdot) \propto \frac{C_{vl}^{VK} + \beta}{\sum_{v=1}^{V} C_{vl}^{VK} + V\beta} \frac{C_{dl}^{DK} + \alpha}{\sum_{k=1}^{K} C_{dk}^{DK} + K\alpha}$$

LDA as an Unsupervised Machine Learning Method



Blei D. M.: Probabilistic Topic Models, Communications of the ACM, vol. 55 (4), p.77 (2012)

- Number of Topics: Hyperparameter
  Compare to Clustering
- No labelled training set needed
- Computes the document-topic and word-topic distributions
- Topic labels have to be assigned "manually"