NLP

Course of DSSC Master degree - University of Trieste

Matilde Trevisani

DEAMS

2020/05/25 (updated: 2020-06-12)

Topic detection

1. LDA (topicmodels R package)
2. Reinert's method (Iramuteq R-interface software package)

How LDA works

The **Latent Dirichlet Allocation (LDA)** is a topic-modelling algorithm.

It assumes a generative process for documents:

1. documents are generated by first picking a distribution over topics
2. and second picking words each from a topic selected according to this distribution.

One common way of modelling the contributions of different topics to a document is to treat

- each topic as a probability distribution over words,
- each document as a probabilistic mixture of topics.

If we have T topics, we can write the probability of the ith word in a given document as:

$$P(w_i) = \prod_{j=1}^{T} P(w_i|z_j)P(z_j)$$

As for the estimation problem

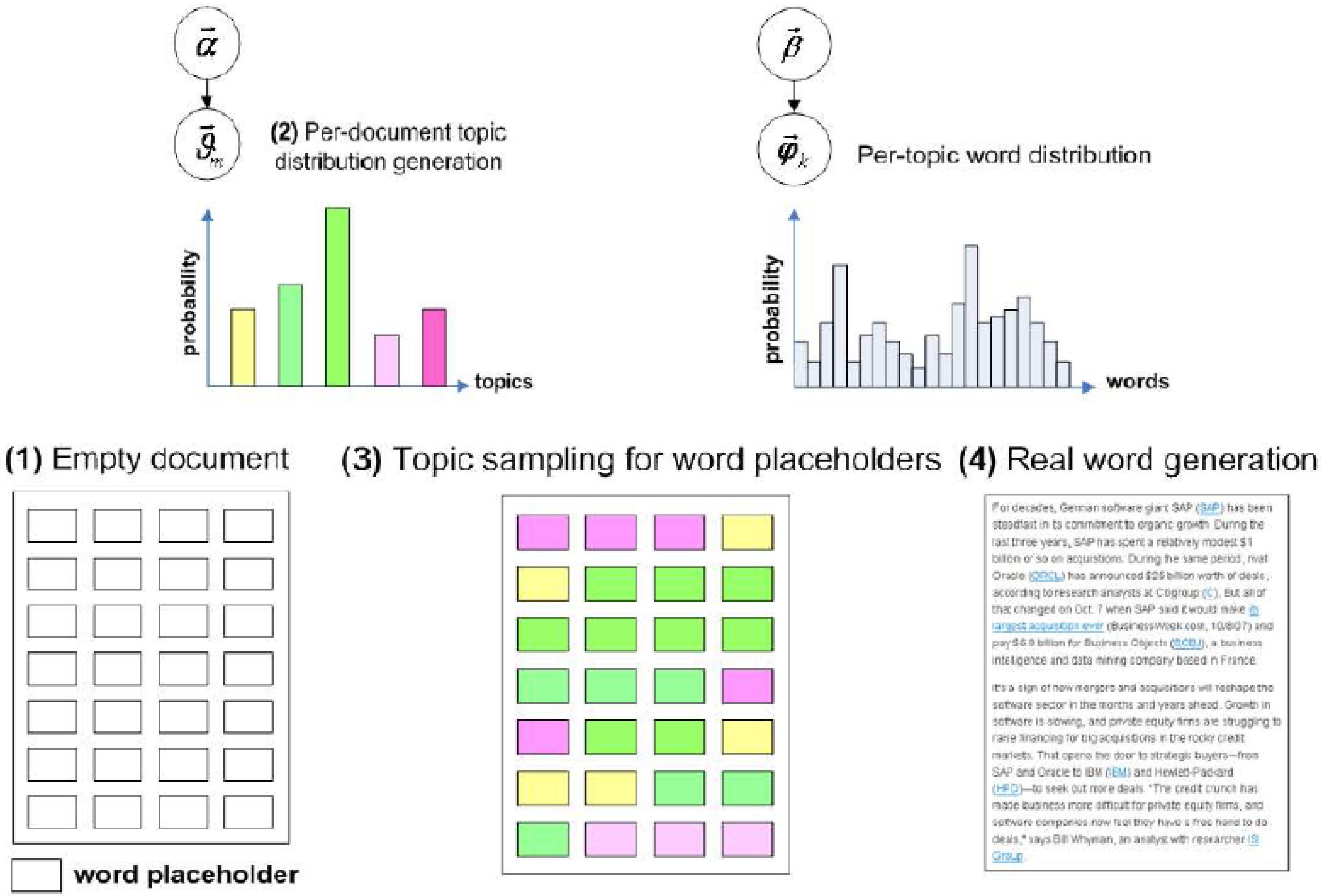
- EM algorithm (multinomial distributions)
- Gibbs sampling (Dirichlet)

(Griffiths & Steyvers, 2004; Blei & Jordan, 2003)

Model used in many fields such as

- collaborative filtering,
- content-based image retrieval
- bioinformatics.

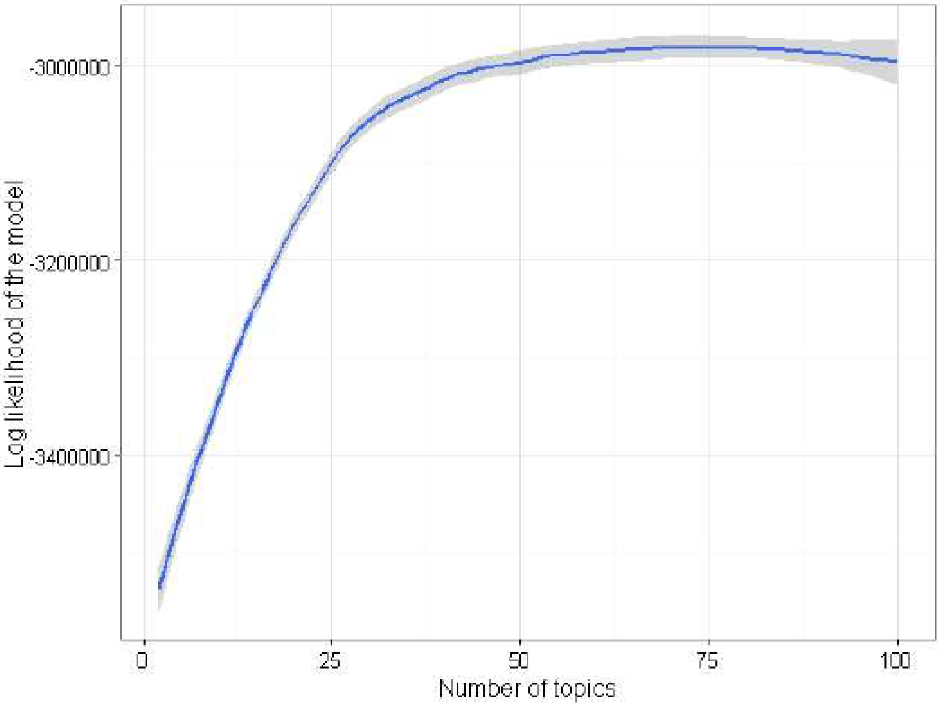
Generative process



Number of topics (in LDA)

To fit the LDA model the number of topics needs be decided in advance.

To identify the optimum number of topics, we first calculated the log-likelihood of the observed data for all models with a number of topics in a given interval. The model with the highest log-likelihood (best fit for the data) is then selected.



LDA Model

We describe a generative model for documents: LDA.

Generative models can be used to postulate complex latent structures responsible for a set of observations.

This kind of approach is particularly useful with text, where the observed data (the words) are explicitly intended to communicate a latent structure (their meaning).

This generative model postulates a latent structure consisting of a set of topics; each document is produced by choosing a distribution over topics, and then generating each word at random from a topic chosen by using this distribution.

the words that appear in a document reflect the particular set of topics it addresses .

LDA Model (continued)

- each topic is a probability distribution over words,
- each document is a probabilistic mixture of topics.

If we have T topics, we can write the probability of the ith word in a given document as:

$$P(w_i) = \prod_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

where z_i is a latent variable indicating the topic from which the ith word was drawn and

- $P(w_i|z_i = j)$ is the probability of the word w_i under the jth topic.
- $P(z_i = j)$ gives the probability of choosing a word from topic j in the current document, which will vary across different documents.

Intuitively, $P(w|z)$ indicates which words are important to a topic, whereas $P(z)$ is the prevalence of those topics within a document.

Each document is characterized in terms of the contributions of multiple topics ("soft classification").

Terminology

- A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $(1, \ldots, V)$.
- A document is a sequence of n words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_n)$, where w_i is the ith word in the sequence.
- A corpus is a collection of m documents denoted by $D = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m)$.

3-level HB model

$$
\begin{aligned}
w_i|z_i, \phi^{z_i} &\sim \mathrm{Multinomial}(\phi^{z_i}) \\
z_i|\theta^{d_i} &\sim \mathrm{Multinomial}(\theta^{d_i}) \\
\theta &\sim \mathrm{Dirichlet}(\alpha) \quad \phi \sim \mathrm{Dirichlet}(\beta)
\end{aligned}
$$

α and β are hyperparameters for the priors on θ and ϕ.

LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus D:

1. Choose $n \sim \mathrm{Pois}(\xi)$ (independent)
2. Choose $\theta \sim \mathrm{Dir}(\alpha)$

3. For each of the n words w_i:

   (a) Choose a topic $z_i \sim \mathrm{Multinomial}(\theta)$.
   (b) Choose a word $w_i \sim \mathrm{Multinomial}(z_i; \beta)$

Joint and marginal distributions

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of n topics $\mathbf{z}$, and a set of n words $\mathbf{w}$ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{i=1}^{n} p(z_i | \theta) p(w_i | z_i, \beta)$$

Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{i=1}^{n} p(z_i | \theta) p(w_i | z_i, \beta) \right) d\theta$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

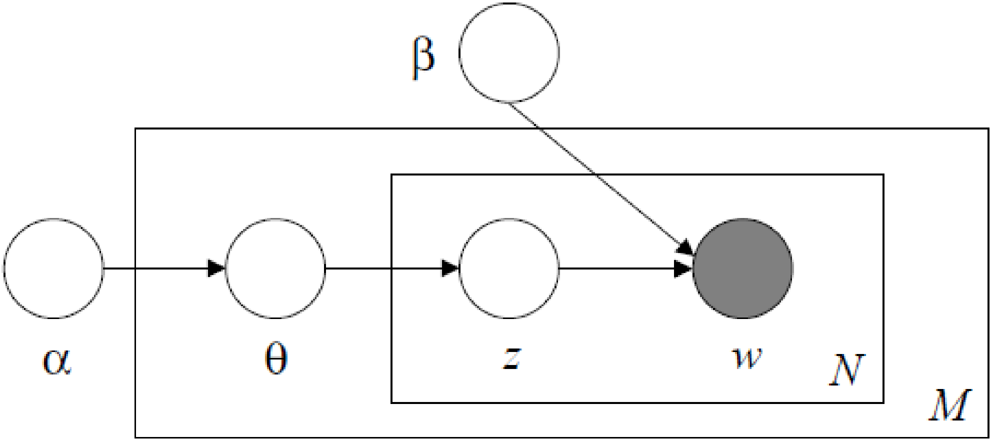
$$p(D | \alpha, \beta) = \prod_{d=1}^{m} \int p(\theta_d | \alpha) \left(\prod_{i=1}^{n} p(z_i | \theta_d) p(w_i | z_i, \beta) \right) d\theta_d$$

Simmetric Dirichlet priors conjugate to Multinomial distributions. ->
Estimation by Gibbs sampler (MCMC)

LDA probabilistic graphical model

The LDA model is represented as a probabilistic graphical model with three levels.

- The parameters α and β are corpus level parameters, assumed to be sampled once in the process of generating a corpus.
- The variables θ_d are document-level variables, sampled once per document.
- Finally, the variables z_{di} and w_{di} are word-level variables and are sampled once for each word in each document.



How Reinert's method works

The corpus is analysed in terms of the presence of words in units (texts or portions of texts). From this contingency table, a squared distance matrix is generated - χ^2-distance, i.e. two texts are close if they share a set of words.

A descending hierarchical clustering is performed from this distance table, which generate classes of units that best differentiate the vocabulary: It extracts classes of words that co-occur and that are best differentiated from other classes. (Reinert 1990, 1993, 1999, 2001)

The occurrence and co-occurrence of words in units is the base to assess similarity among texts.

Number of clusters (in Reinert's method)

The descending hierarchical classification method is an iterative procedure:

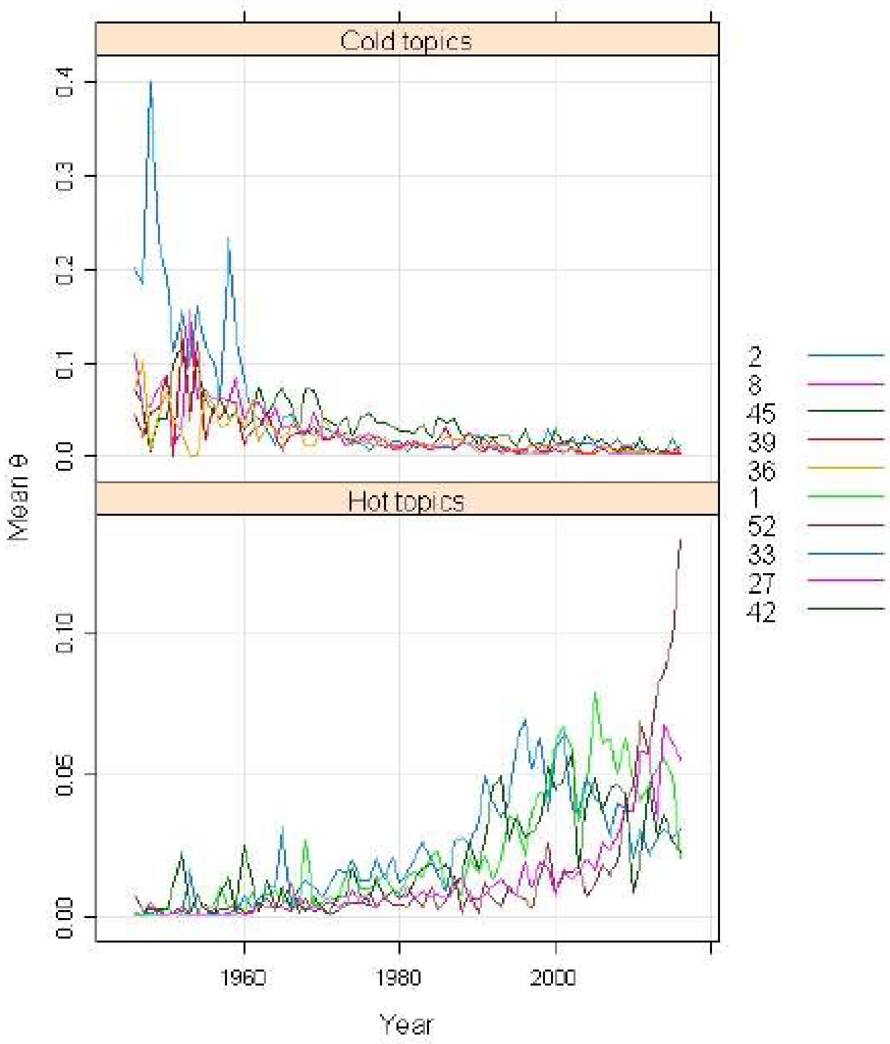
- at each descending step, the bigger class of classes X and Y is decomposed next, and so on.
- The procedure stops if a predetermined number of iterations does not result in further divisions (or when classes include a limited number of texts).

Example of LDA

Output (e.g. 60 topics) from a chronological corpus of scientific literature (Statistics discipline)

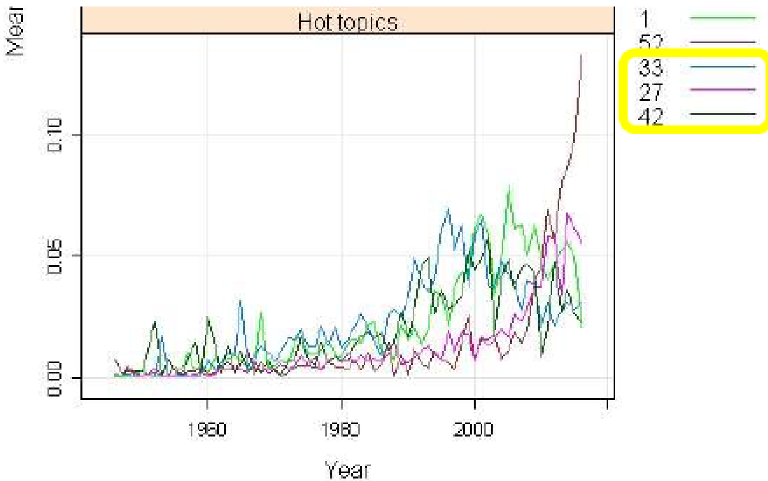
Topic 1	Topic 2	Topic 3	Topic 4	...	Topic58	Topic59	Topic60
consistent	statistics	selection	estimators	...	temperature	truncate	stratum
proposed	statistical	criterion	estimator	...	earth	singly	hartley
estimator	paper	model	sample	...	meteorological	poisson parameter	number of strata
regression	data	outliers	robust	...	climate	poisson distributions	ratio estimators
estimators	economic	based	small	...	atmospheric	hypergeometric	variance estimator
model	research	criteria	monte	...	ozone	doubly	stratify
asymptotically	problems	methods	asymptotic	...	atmosphere	binomial	horvitz
estimation	social	proposed	study	...	cool	poisson	stratified simple randor
asymptotic	statisticians	regression	properties	...	spatial	inspection	proportional allocation
estimating	labor	algorithm	finite	...	wind speed	specification limits	stratified random
propose	made	show	carlo	...	wind	lot	thompson
semiparametric	american	article	based	...	temporal	producer	population total
covariates	employment	large	efficiency	...	sea	asymptotic variances	grundy
show	presented	information	large	...	weather	simplify	multistage designs
efficient	association	approach	empirical	...	aerosol	poisson case	estimation variable
nonparametric	program	propose	samples	...	warm	e act results	stratification
function	author	data	proposed	...	climate models	life test	replacement
based	policy	robust	article	...	volcanic	moment	proportionate
article	states	outlier	robustness	...	gas	correlated bivariate poisson	sample allocation
simulation studies	article	sample	estimation	...	carbon	infinite series	balanced sampling
data	annual	stepwise	compared	...	dio	zidek	allocation
normal	united	select	breakdown	...	ocean	beta	liml
approach	business	selecting	results	...	misr	algebraic	cum
finite	analysis	prediction	means	...	wind direction	truncation	stratification variable
...	...	...	...	...	...	...	...

Hot and cold topics (with LDA)



Hot topics (with LDA)

Some examples ("hot" topics)



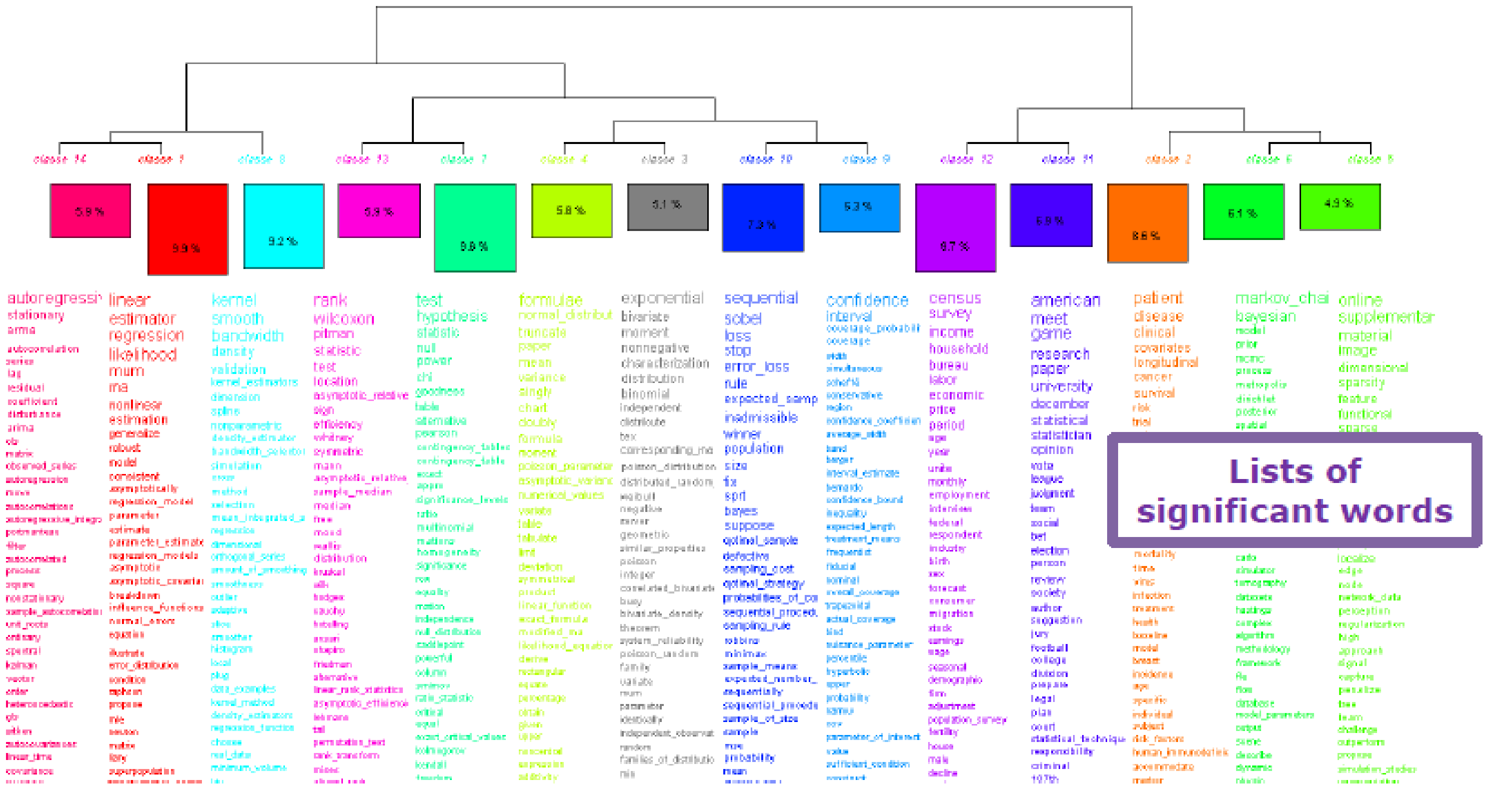
Topic33
prior
bayesian
Monte Carlo
posterior
Bayes
Markov Chain
approach
distributions
model
priors
methods
sampling
inference
parameters
Gibbs
data
hierarchical
information
MCMC
sampler
examples
frequentist
problem
article
empirical
problems
algorithm

Topic27
high dimension
functional
data
dimension
analysis
matrix
proposed
sparse
propose
methods
features
low
based
model
settings
space
predictors
approach
structured
structure
sparsity
linear
finite
demonstrate
sample
framework
applications
multidimension

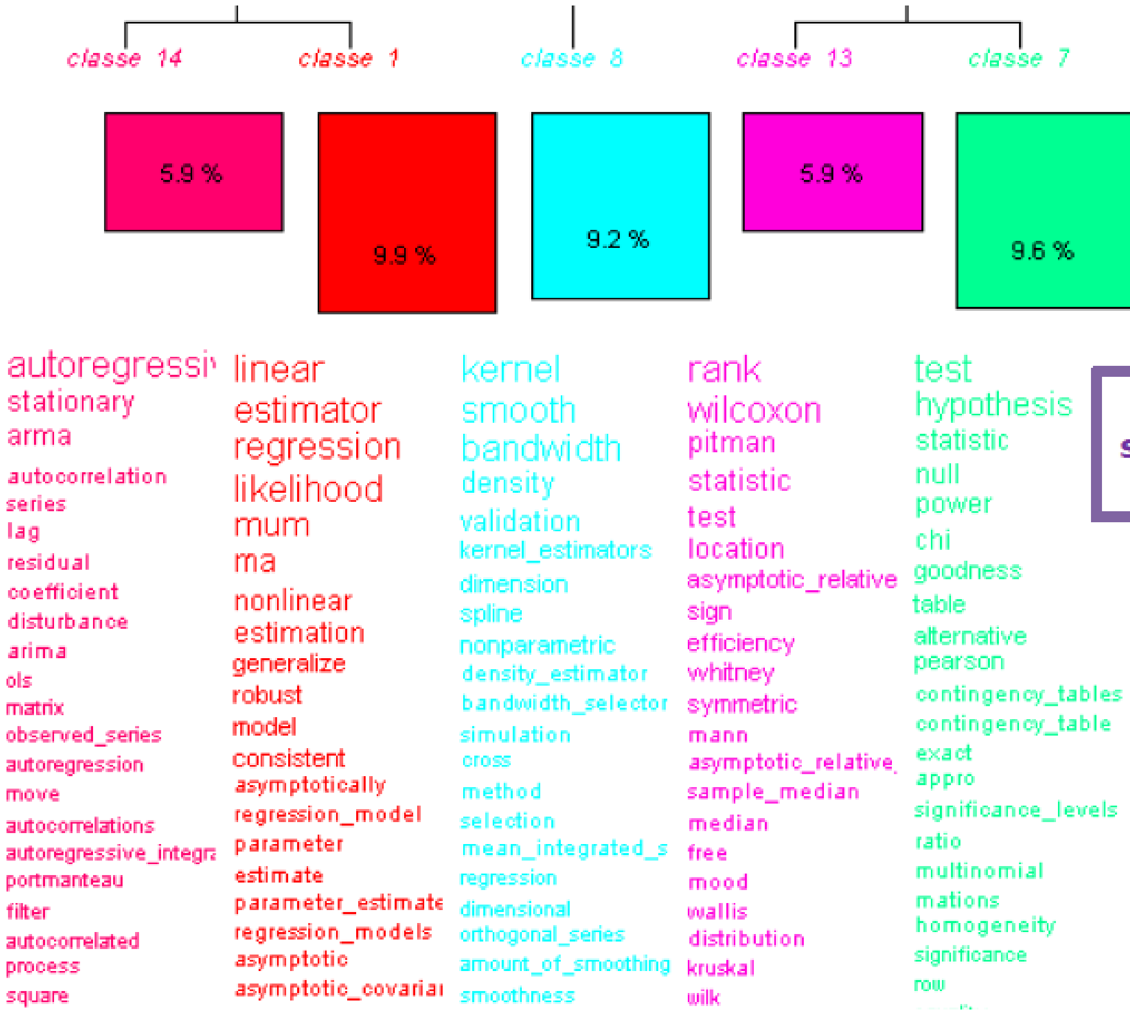
Topic42
time
data
survival
failure
model
risk
cancer
disease
censoring
study
event
hazard
censored
follow
events
covariates
patients
proportional
methods
dependent
subject
specific
hazards
function
proposed
breast
based
Cox

Example of Reinert's method

An example at a first glance: 14 topics (Iramuteq)

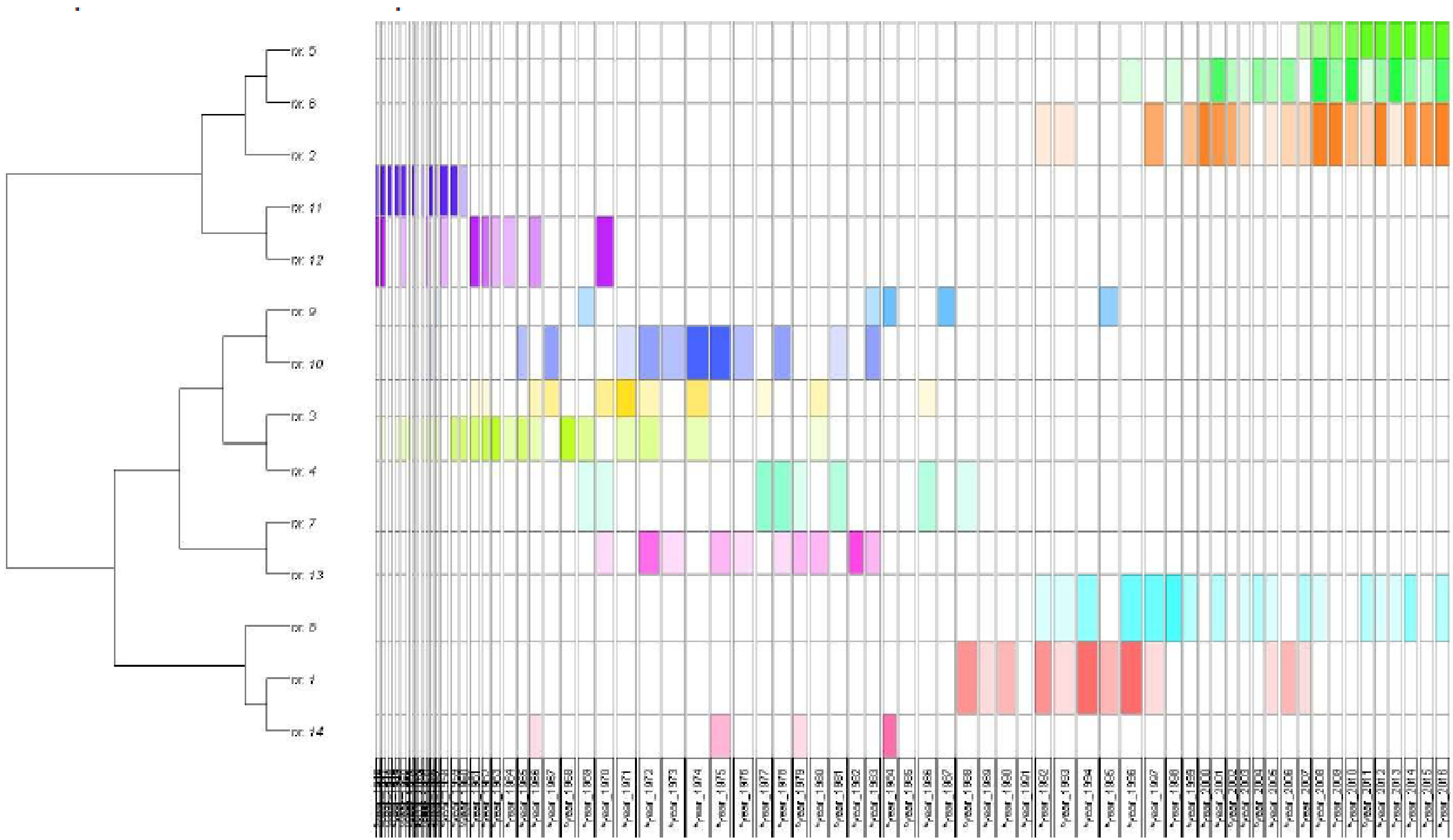


A zoom



Lists of significant words (ZOOM)

Temporal evolution of topic (Reinert)



Hot and cold topics (Reinert)