Ill-posed problems of non-negative matrix factorization with applications to text analysis

Konstantin Vorontsov

(MIPT • Moscow, Russia)

Quasilinear equations, inverse problems and their applications

Moscow • 12–15 September 2016

Contents

Sparse stochastic matrix factorization under KL-loss

Given a matrix $Z = \|z_{ij}\|_{n\times m}$, $(i,j) \in \Omega \subseteq \{1..n\} \times \{1..m\}$

Find matrices $X = \|x_{it}\|_{n\times k}$ and $Y = \|y_{tj}\|_{k\times m}$ such that

$$\|Z - XY\|_{\Omega,d} = \sum_{(i,j)\in\Omega} d\Big(z_{ij}, \sum_t x_{it} y_{tj}\Big) \to \min_{X,Y}$$

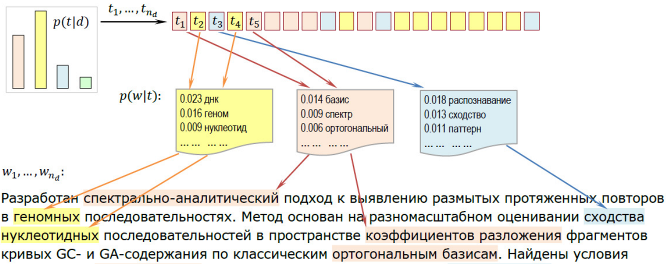
Variants of the problem:

- quadratic loss: $d(z,\hat{z}) = (z - \hat{z})^2$
- Kullback–Leibler loss: $d(z,\hat{z}) = z\ln(z/\hat{z}) - z + \hat{z}$
- nonnegative matrix factorization: $x_{it} \geqslant 0$, $y_{tj} \geqslant 0$
- stochastic matrix factorization: $x_{it} \geqslant 0$, $y_{tj} \geqslant 0$, $\sum_i x_{it} = 1$, $\sum_t y_{tj} = 1$
- sparse input data: $|\Omega| \ll nm$
- sparse output factorization X, Y

Probabilistic Topic Model (PTM) generating a text collection

Topic model explains terms w in documents d by topics t:

$$p(w|d) = \sum_t p(w|t)p(t|d)$$



Matrix factorization: $\big(p(w|d)\big)_{W\times D} = \Phi\Theta$, where:

$\Phi = \big(\phi_{wt}\big)_{W\times T}$ — term distributions of topics, $\phi_{wt} = p(w|t)$,

$\Theta = \big(\theta_{td}\big)_{T\times D}$ — topic distributions of documents, $\theta_{td} = p(t|d)$.

Inverse problem: text collection $\to$ PTM

Given: D is a set (collection) of documents
W is a set (vocabulary) of terms
n_{dw} = how many times term w appears in document d

Find: parameters $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ of the topic model

$$p(w|d) = \sum_t \phi_{wt}\theta_{td}.$$

The problem of log-likelihood maximization under constraints:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \ \to\ \max_{\Phi,\Theta},$$

$$\phi_{wt} \geqslant 0, \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1.$$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

EM-algorithm for likelihood maximization [Hofmann, 1999]

From KKT conditions for the constrained maximization problem

Theorem

Maximum of $\mathscr{L}(\Phi, \Theta)$ satisfies the system of equations with model parameters ϕ_{wt}, θ_{td} and auxiliary variables p_{tdw}, n_{wt}, n_{td}:

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} \equiv p(t|d,w) = \dfrac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}; \\ \phi_{wt} = \dfrac{n_{wt}}{\sum_{w'} n_{w't}}; \quad n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \\ \theta_{td} = \dfrac{n_{td}}{\sum_{t'} n_{t'd}}; \quad n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \end{array} \right. \\ \text{M-step:} & \end{array}$$

EM-algorithm alternates E-step and M-step until convergence.
EM-algorithm is equivalent to a simple iteration method.

LDA — Latent Dirichlet Allocation [Blei, 2003]

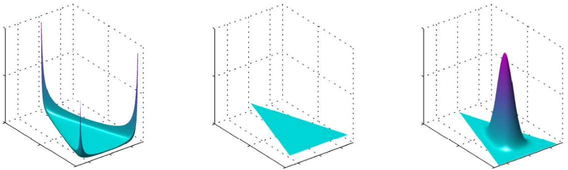
Assumption. Column vectors $\phi_t = (\phi_{wt})_{w\in W}$ and $\theta_d = (\theta_{td})_{t\in T}$ are generated from Dirichlet distributions, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod\limits_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \ \beta_t \geqslant 0;$$

$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod\limits_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t \geqslant 0;$$

Example:

$\mathrm{Dir}(\theta|\alpha)$
$|T| = 3$
$\theta, \alpha \in \mathbb{R}^3$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ $\quad$ $\alpha_1 = \alpha_2 = \alpha_3 = 1$ $\quad$ $\alpha_1 = \alpha_2 = \alpha_3 = 10$

The main difference between LDA and PLSA

The estimates of conditionals $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- in PLSA — unbiased maximum likelihood estimates:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- in LDA — smoothed Bayesian estimates:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

The difference is significant for small n_{wt}, n_{td} only.
Robust LDA and robust PLSA produce almost identical models.

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.

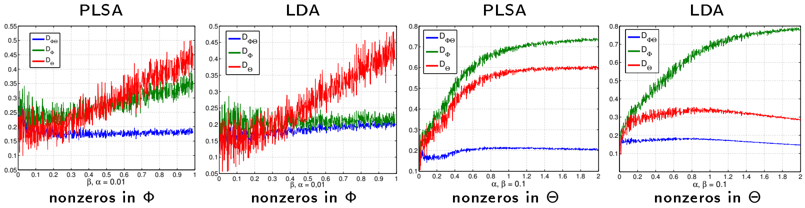
Potapenko A. A., Vorontsov K. V. Robust PLSA Performs Better Than LDA. ECIR-2013, Moscow, Russia, 24-27 March 2013. LNCS, Springer. Pp. 784–787.

Topic Modeling as an ill-posed inverse problem

The *nonuniqueness* and *instability* of matrix factorization:
$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$ for all S such that Φ', Θ' are stochastic.

Experiment: recovering known Φ, Θ on synthetic dataset,
$|D| = 500, \;\; |W| = 1000, \;\; |T| = 30.$

Result: product $\Phi\Theta$ is always recovered well, however
matrix Φ and matrix Θ are recovered if being highly sparse only:



Conclusions: Dirichlet prior is too weak as a regularizer;
more regularization is needed to ensure a stable solution.

Additive Regularization for Topic Modeling (ARTM)

Additional *regularization* criteria $R_i(\Phi, \Theta) \to \max$, $i = 1, \dots, n$.

The problem of regularized log-likelihood maximization under non-negativeness and normalization constraints:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

$$\phi_{wt} \geqslant 0; \quad \sum_{w\in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t\in T} \theta_{td} = 1$$

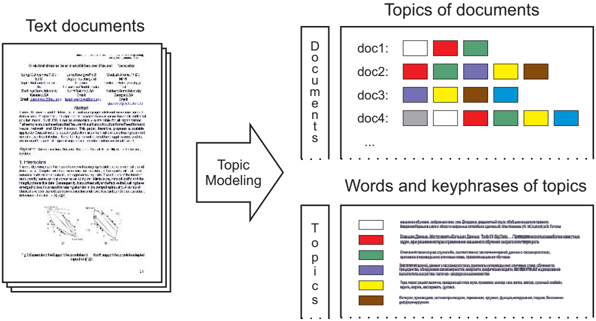
where $\tau_i > 0$ are *regularization coefficients*.

PLSA: $\quad R(\Phi, \Theta) = 0$

LDA: $\quad R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$

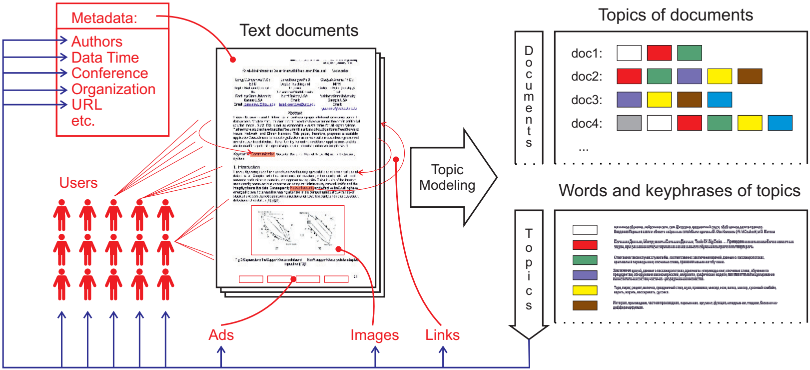
Multimodal Probabilistic Topic Modeling

Given a text document collection *Probabilistic Topic Model* finds:
$p(t|d)$ — topic distribution for each document d,
$p(w|t)$ — term distribution for each topic t.



Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distributions
$p(t|\text{author})$, $p(t|\text{time})$, $p(t|\text{category})$, $p(t|\text{tag})$, $p(t|\text{link})$,
$p(t|\text{object-on-image})$, $p(t|\text{advertising-banner})$, $p(t|\text{users})$, etc.
and binds all these modalities into a single topic model.



Multimodal ARTM: combining multimodality and regularization

M is the set of modalities
W^m is a vocabulary of tokens of m-th modality, $m \in M$
$W = W^1 \sqcup \dots \sqcup W^M$ is a joint vocabulary of all modalities

The problem of multimodal regularized log-likelihood maximization under non-negativeness and normalization constraints:

$$\sum_{m \in M} \lambda_m \underbrace{\sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{modality log-likelihood } \mathscr{L}_m(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

$$\phi_{wt} \geqslant 0, \quad \sum_{w \in W^m} \phi_{wt} = 1, \ m \in M; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1.$$

where $\lambda_m > 0$, $\tau_i > 0$ are *regularization coefficients*.

EM-algorithm for multimodal ARTM

EM-algorithm is a simple-iteration method for a system of equations

Theorem. The local maximum (Φ, Θ) satisfies the following system of equations with auxiliary variables $p_{tdw} = p(t|d,w)$:

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}} \bigl(\phi_{wt}\theta_{td}\bigr);$$

$$\phi_{wt} = \underset{w \in W^m}{\operatorname{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); \quad n_{wt} = \sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw};$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \quad n_{td} = \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw};$$

where $\underset{t \in T}{\operatorname{norm}}\, x_t = \frac{\max\{x_t, 0\}}{\sum\limits_{s \in T} \max\{x_s, 0\}}$ is nonnegative normalization;

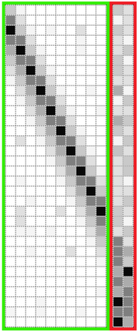
$m(w)$ is the modality of the term w, so that $w \in W^{m(w)}$.

Assumptions: what topics would be well-interpretable?

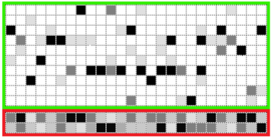
Topics $S \subset T$ contain domain-specific terms
$p(w|t)$, $t \in S$ are sparse and different (weakly correlated)

Topics $B \subset T$ contain background terms
$p(w|t)$, $t \in B$ are dense and contain common lexis words

$\Phi_{W \times T}$



$\Theta_{T \times D}$



Smoothing regularization (rethinking LDA)

The non-sparsity assumption for background topics $t \in B$:
ϕ_{wt} are similar to a given distribution β_w;
θ_{td} are similar to a given distribution α_t.

$$\sum_{t\in B} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

We minimize the sum of these KL-divergences to get a regularizer:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in B} \sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d\in D} \sum_{t\in B} \alpha_t \ln \theta_{td} \to \max.$$

The regularized M-step applied for all $t \in B$ coincides with LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w, \qquad \theta_{td} \propto n_{td} + \alpha_0 \alpha_t,$$

which is new non-Bayesian interpretation of LDA [Blei 2003].

Sparsing regularizer (further rethinking LDA)

The sparsity assumption for domain-specific topics $t \in S$: distributions ϕ_{wt}, θ_{td} contain many zero probabilities.

We maximize the sum of KL-divergences $\mathrm{KL}(\beta \| \phi_t)$ and $\mathrm{KL}(\alpha \| \theta_d)$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \to \max .$$

The regularized M-step gives “anti-LDA”, for all $t \in S$:

$$\phi_{wt} \propto \big(n_{wt} - \beta_0 \beta_w\big)_+, \qquad \theta_{td} \propto \big(n_{td} - \alpha_0 \alpha_t\big)_+ .$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Regularization for topics decorrelation

The dissimilarity assumption for domain-specific topics $t \in S$: if topics are interpretable then they must differ significantly.

We maximize covariances between column vectors ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max .$$

The regularized M-step makes columns of Φ more distant:

$$\phi_{wt} \propto \left(n_{wt} - \tau \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws} \right)_{+} .$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Regularization for topic selection

Let us maximize KL-divergence: $\mathrm{KL}\bigl(\frac{1}{|T|} \parallel p(t)\bigr) \to \max$
to make distribution over topics $p(t)$ sparse:

$$R(\Theta) = -\tau n \sum_{t \in S} \frac{1}{|T|} \ln \underbrace{\sum_{d \in D} p(d)\theta_{td}}_{p(t)} \to \max .$$

The regularized M-step formula results in Θ row sparsing:

$$\theta_{td} = \operatorname*{norm}_{t \in T} \left(n_{td} \left(1 - \tau \frac{n}{n_t |T|} \right) \right).$$

The row sparsing effect:
if $n_t < \tau \frac{n}{|T|}$ then all values in the t-th row turn into zeros.

ARTM: available regularizers

- topic smoothing ($\Leftrightarrow$ Latent Dirichlet Allocation)
- topic sparsing
- topic decorrelation
- topic selection via entropy sparsing
- topic coherence maximization
- supervised learning for classification and regression
- semi-supervised learning
- using documents citation and links
- modeling temporal topic dynamics
- using vocabularies in multilingual topic models
- etc.

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models. Machine Learning Journal. Springer, 2015.

BigARTM project

BigARTM features:

- Parallel + Online + Multimodal + Regularized Topic Modeling
- Out-of-core processing of Big Data
- Built-in library of regularizers and quality measures

BigARTM community:

- Open-source https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Documentation http://bigartm.org

BigARTM license and programming environment:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

BigARTM vs Gensim vs Vowpal Wabbit

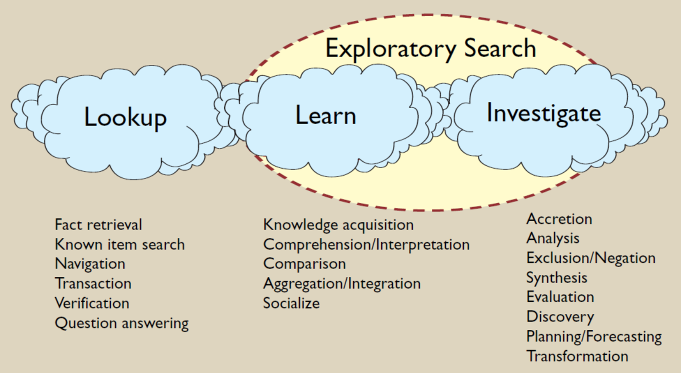
- 3.7M articles from Wikipedia, 100K unique words

	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *procs* = number of parallel threads
- *inference* = time to infer θ_d for 100K held-out documents
- *perplexity* is calculated on held-out documents.

Exploratory Search for learning, knowledge acquisition and discovery

- what if the user doesn't know which keywords to use?
- what if the user isn't looking for a single answer?



Gary Marchionini. Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41–46.

From *close reading* to *distant reading*

Information Seeking Mantra [B.Shneiderman, 1996]

«Overview first, zoom and filter, details on demand»

Distant reading [Franco Moretti, 2005]

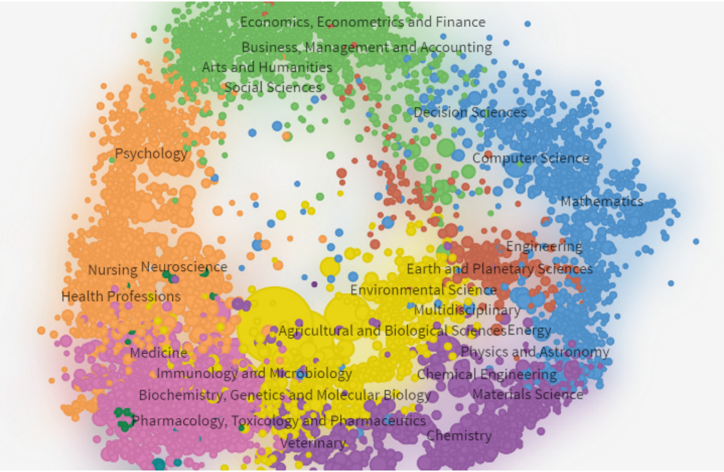
«*Distant reading* is not an obstacle but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.»

B.Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Visual Languages, 1996.

F.Moretti. Graphs, Maps, Trees: Abstract Models for a Literary History. 2005.

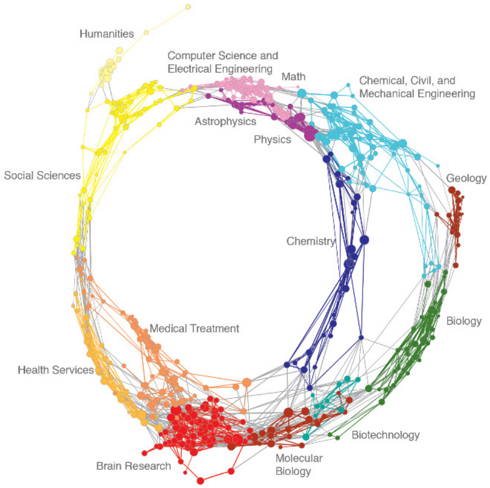
S.Janicke, G.Franzini, M.F.Cheema, G.Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis, 2015.

Example #1: the map of science



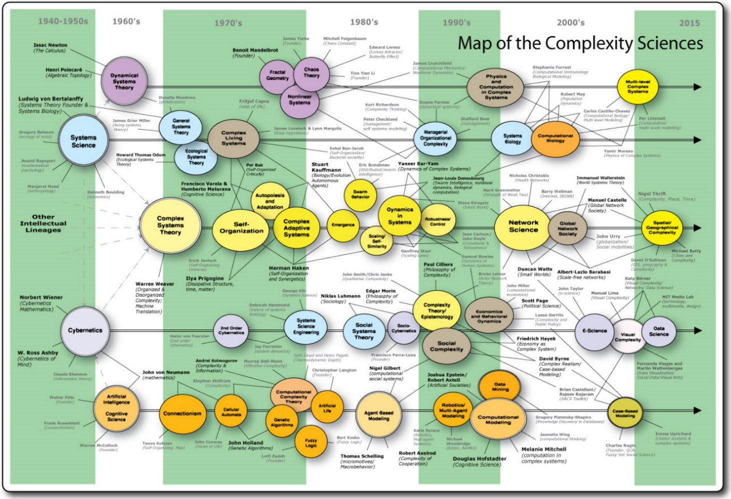
http://onlinelibrary.wiley.com/browse/subjects

Example #2: the map of science



http://scimaps.org

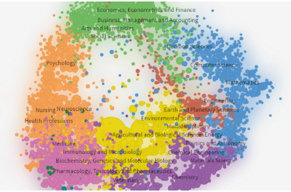
Example #3: hand-made time–topics map of Complexity Theory



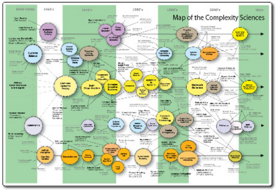
http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences

Challenge for Topic Modeling coming from exploratory search

How to build maps of science fully automatically?



$\rightarrow$



- ARTM makes the model temporal, hierarchical, multimodal, multilanguage, multigram, well-interpretable at once
- BigARTM helps to learn such model effectively from millions of documents