Multicriteria Regularization for Probabilistic Topic Modeling of Large Text Collections

Konstantin Vorontsov • Oleksandr Frei • Anna Potapenko
Murat Apishev • Peter Romov • Nikita Doykov
Andrey Shadrikov • Alexander Plavin • Marina Dudarenko
(MIPT, CC RAS, Yandex • Moscow, Russia)

Optimization and Applications in Control and Data Science
Moscow • 13–15 May 2015

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

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

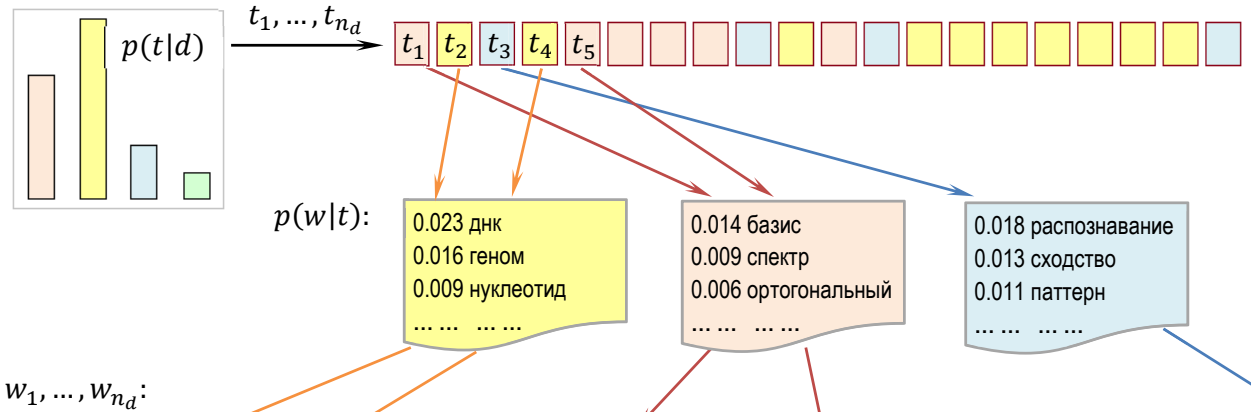
Variety of problems:

- 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:

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



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Inverse problem: text collection $\rightarrow$ 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:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \rightarrow \max_{\Phi,\Theta},$$

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

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

Topic Modeling is an ill-posed inverse problem

Topic Modeling is the problem of *stochastic matrix factorization*:

$$p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}.$$

In matrix notation $\underset{W \times D}{Z} = \underset{W \times T}{\Phi} \cdot \underset{T \times D}{\Theta}$, where

$Z = \left\| p(w|d) \right\|_{W \times D}$ is known term–document matrix,
$\Phi = \left\| \phi_{wt} \right\|_{W \times T}$ is unknown term–topic matrix, $\phi_{wt} = p(w|t)$,
$\Theta = \left\| \theta_{td} \right\|_{T \times D}$ is unknown topic–document matrix, $\theta_{td} = p(t|d)$.

Matrix factorization is not unique, the solution is not stable:
$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$ for all S such that Φ', Θ' are stochastic.

Then, regularization is needed to find appropriate solution.

Additive Regularization for Topic Modeling (ARTM)

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

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

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

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

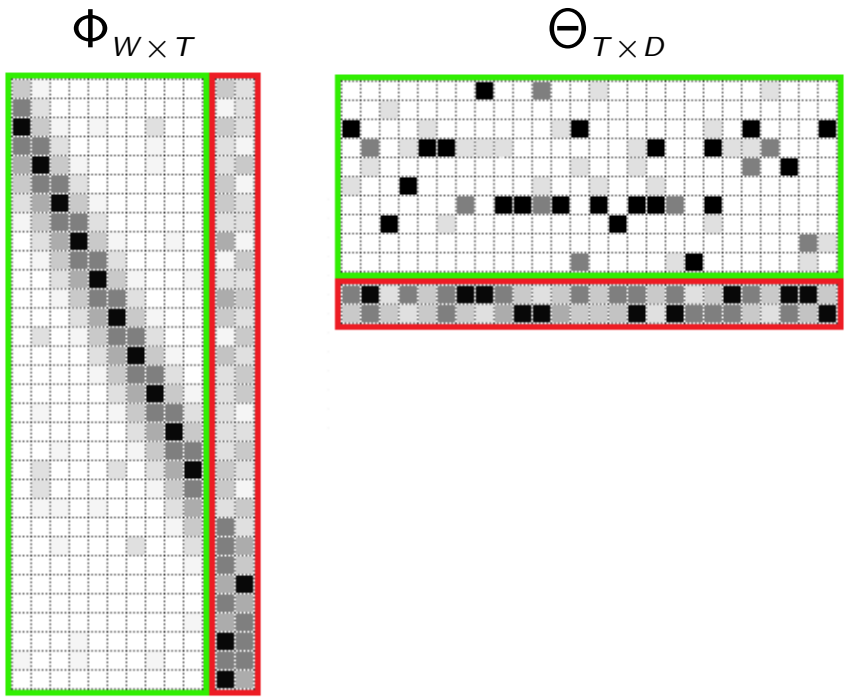
where $\tau_i > 0$ are *regularization coefficients*.

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST, 2014.

Regularizer $R(\Phi, \Theta)$ comes from problem domain requirements

ARTM example: sparsing + smoothing + decorrelation

- smoothing background (common lexis) topics B in Φ and Θ
- sparsing domain-specific topics $S = T \setminus B$ in Φ and Θ
- decorrelation of topics in Φ



ARTM example: sparsing + smoothing + decorrelation

Additive combination of 5 regularizers:

- smoothing background (common lexis) topics B in Φ and Θ
- sparsing domain-specific topics $S = T \backslash B$ in Φ and Θ
- decorrelation of topics in Φ

$$R(\Phi, \Theta) = + \beta_1 \sum_{t \in B} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_1 \sum_{d \in D} \sum_{t \in B} \alpha_t \ln \theta_{td}$$

$$- \beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td}$$

$$- \gamma \sum_{t \in T} \sum_{s \in T \backslash t} \sum_{w \in W} \phi_{wt} \phi_{ws}$$

where β_0, α_0, β_1, α_1, γ are regularization coefficients.

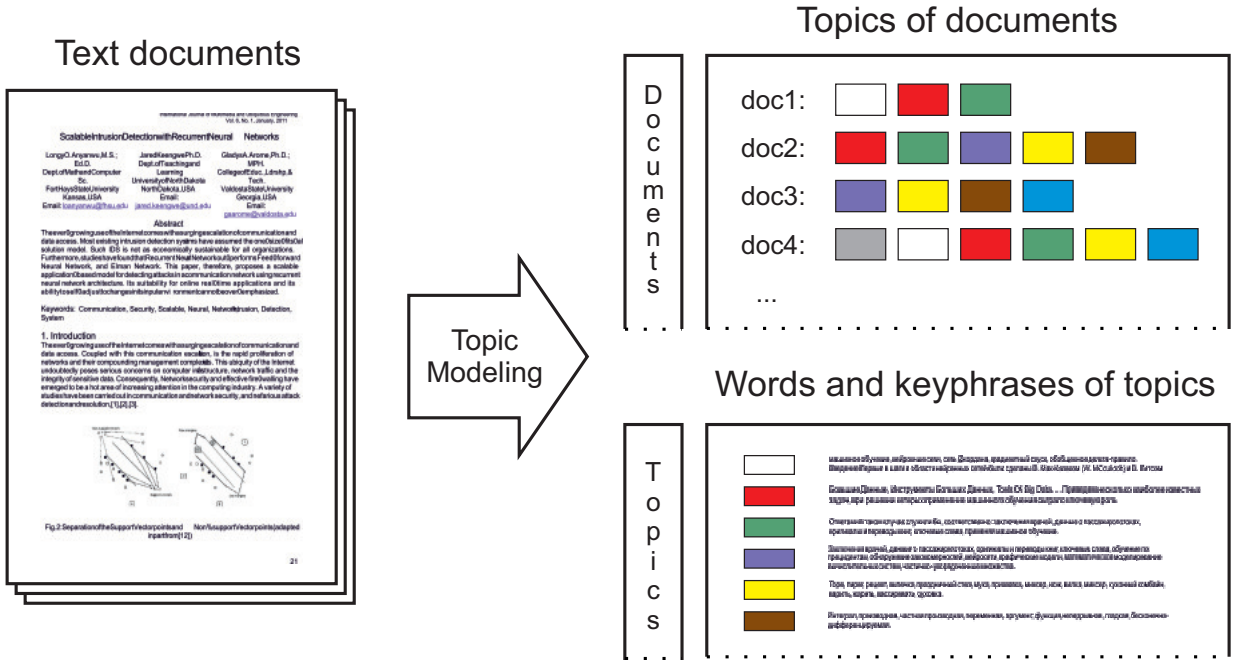
ARTM: available regularizers

- 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, 2014.

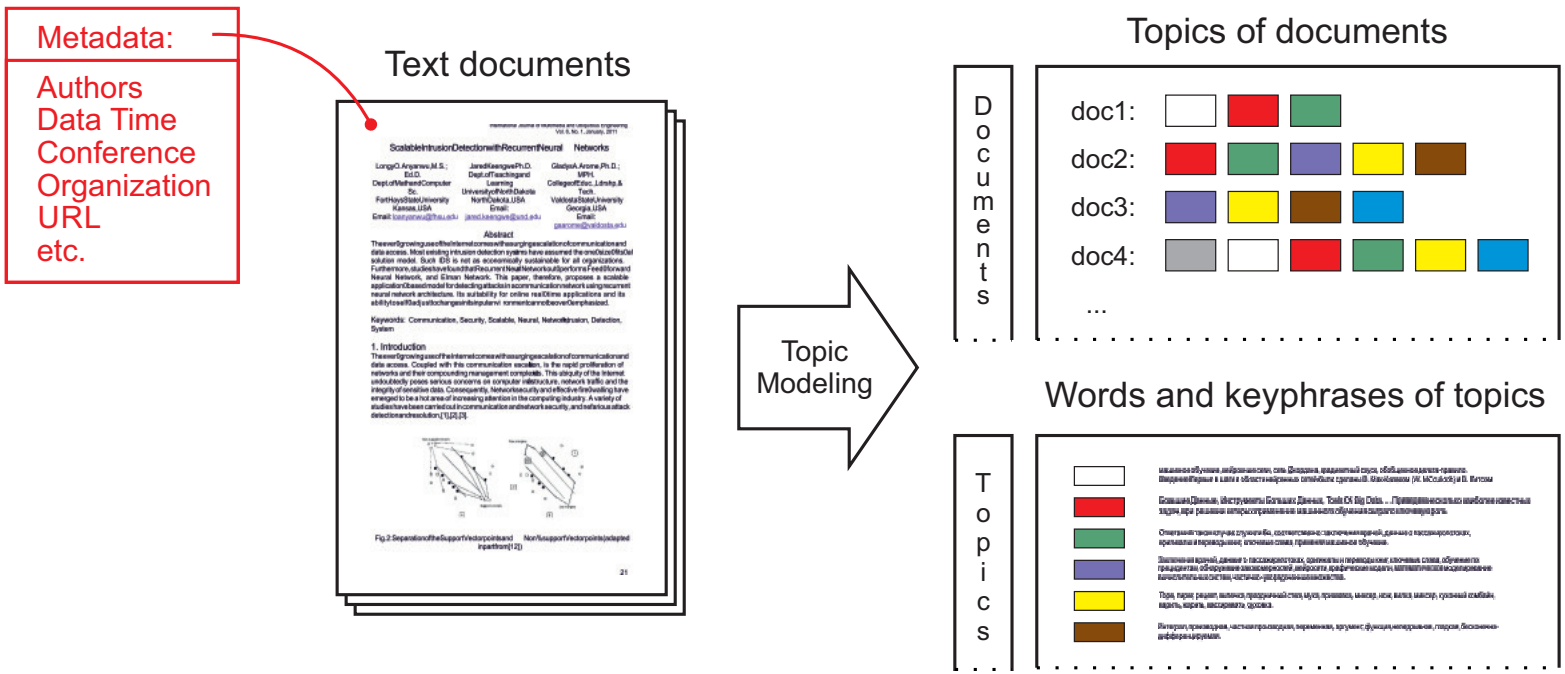
Multimodal Probabilistic Topic Modeling

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.



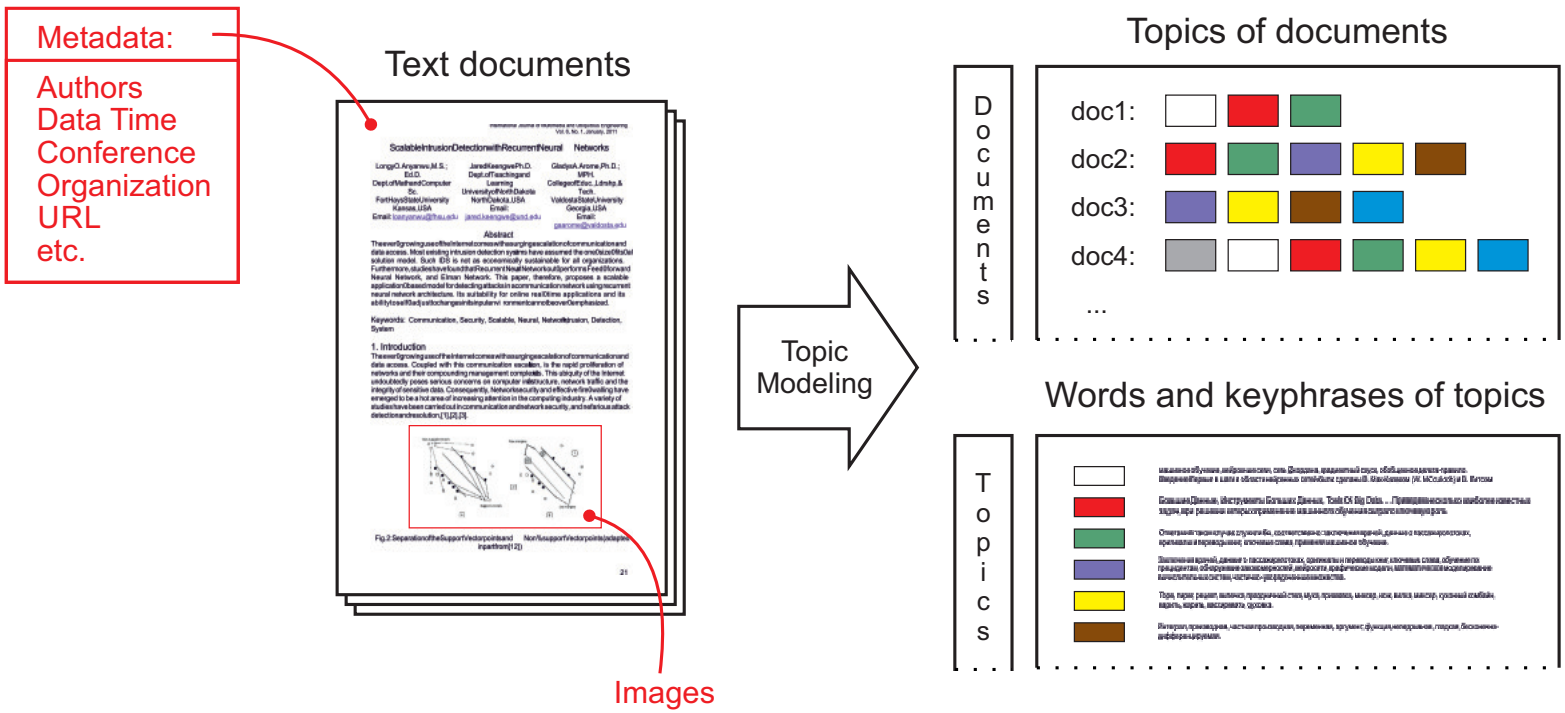
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$,



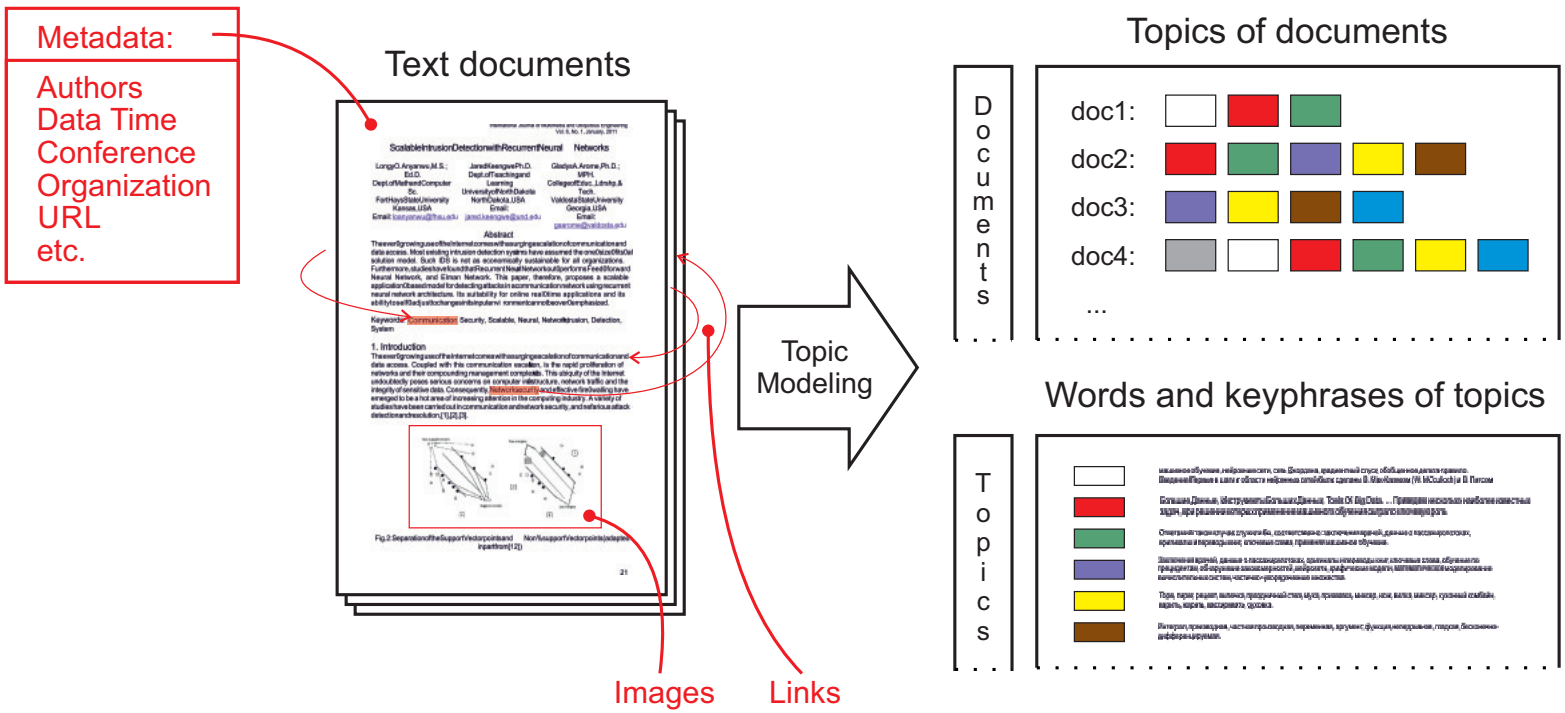
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$,



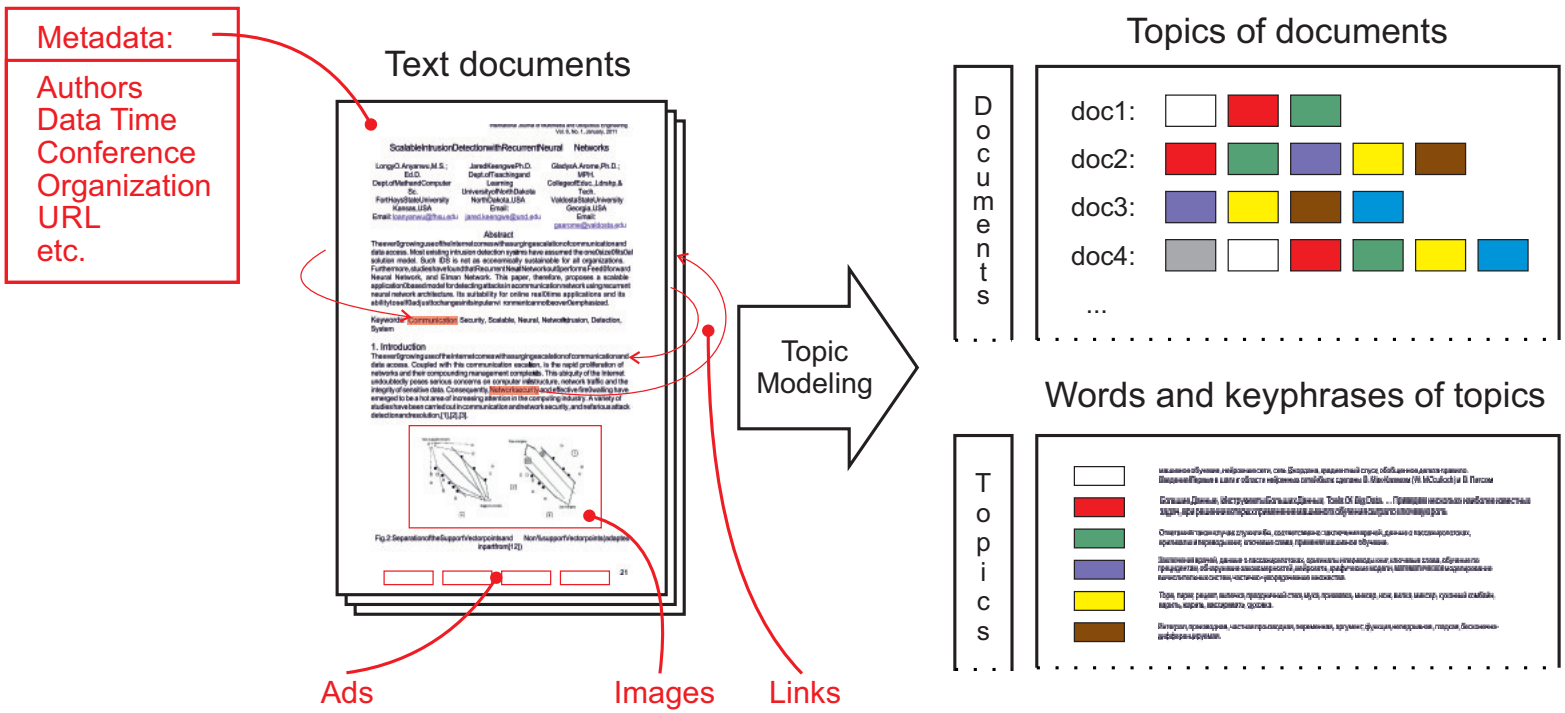
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$,



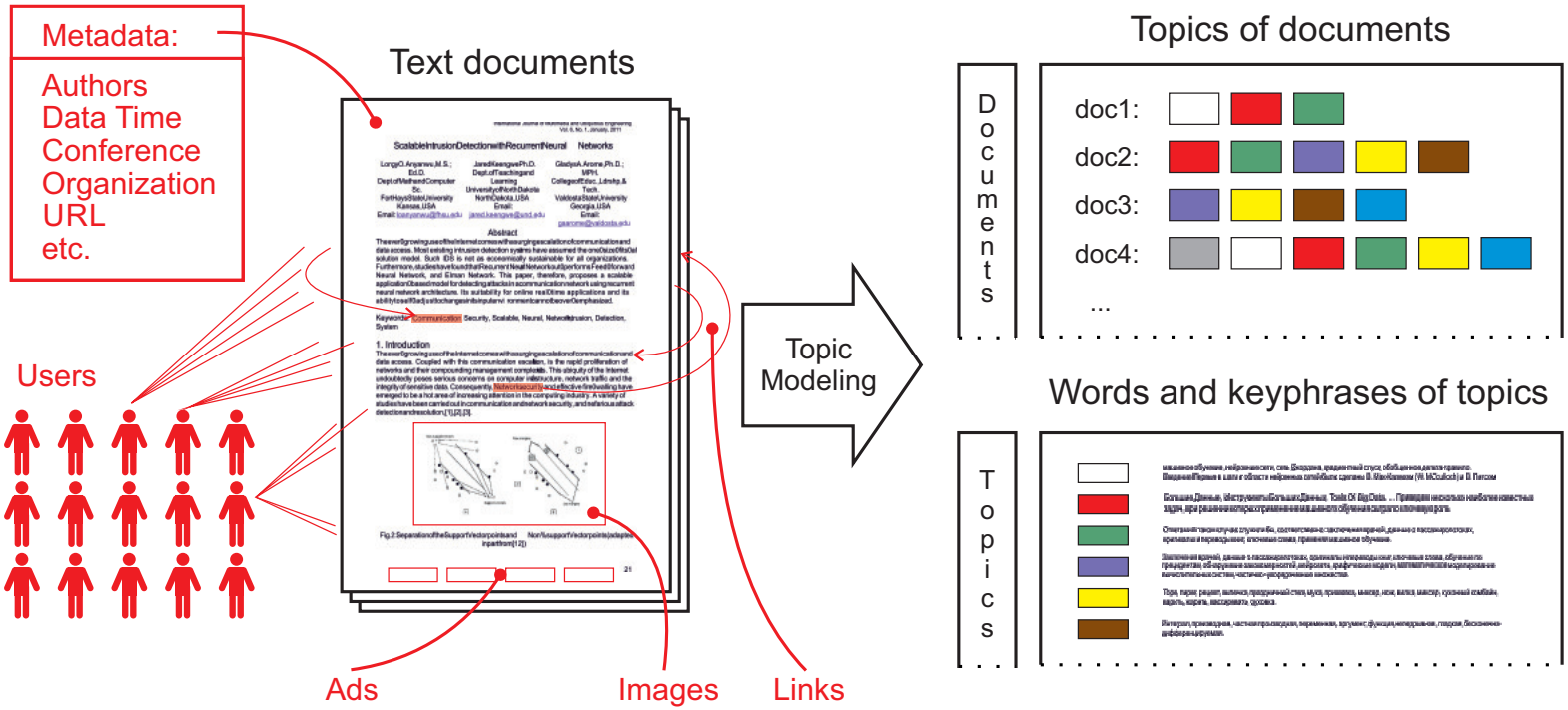
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$, advertising banners $p(b|t)$,



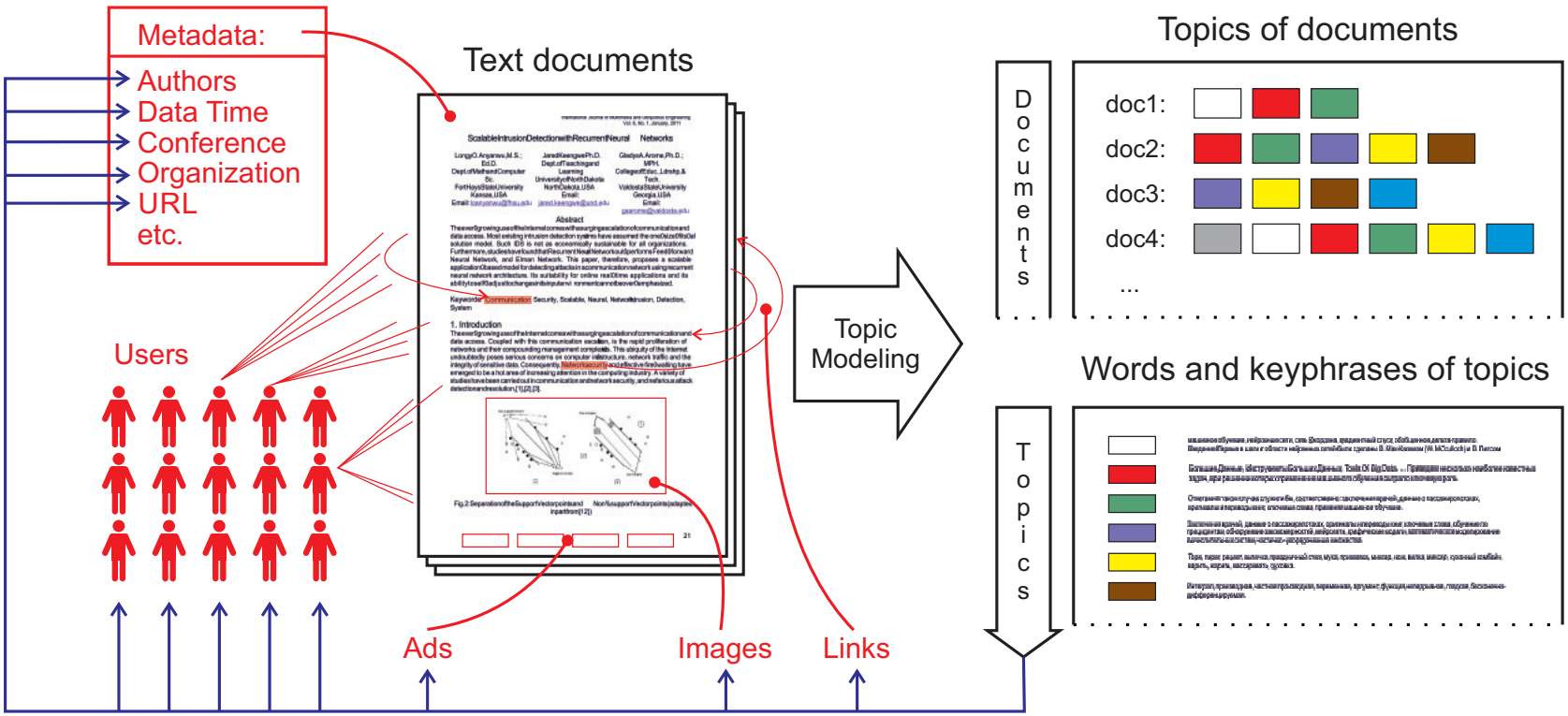
Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$, advertising banners $p(b|t)$, users $p(u|t)$,



Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$, advertising banners $p(b|t)$, users $p(u|t)$, 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, $m \in M$
$W = W^1 \sqcup \cdots \sqcup W^M$ is a joint vocabulary of all modalities

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

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

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

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 $p_{tdw} = p(t|d,w)$:

$$p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right);$$

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

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

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

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

Summary of ARTM approach

EM-algorithm is computationally effective:

- It has linear time complexity $O(n \cdot |T| \cdot \text{nIter})$
- Its online version makes only one pass through big collection
- Parallelism is possible for both multi-core CPUs and clusters

ARTM reduces barriers to entry into PTM research field:

- General EM-algorithm for many models and their combinations
- PLSA, LDA, and 100s of PTMs are covered by ARTM
- Combining multiple modalities and regularizers is easy
- No complicated Bayesian inference and graphical models

Open problem / Under development:

- Adaptive optimization of regularization coefficients τ_i, λ_m

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

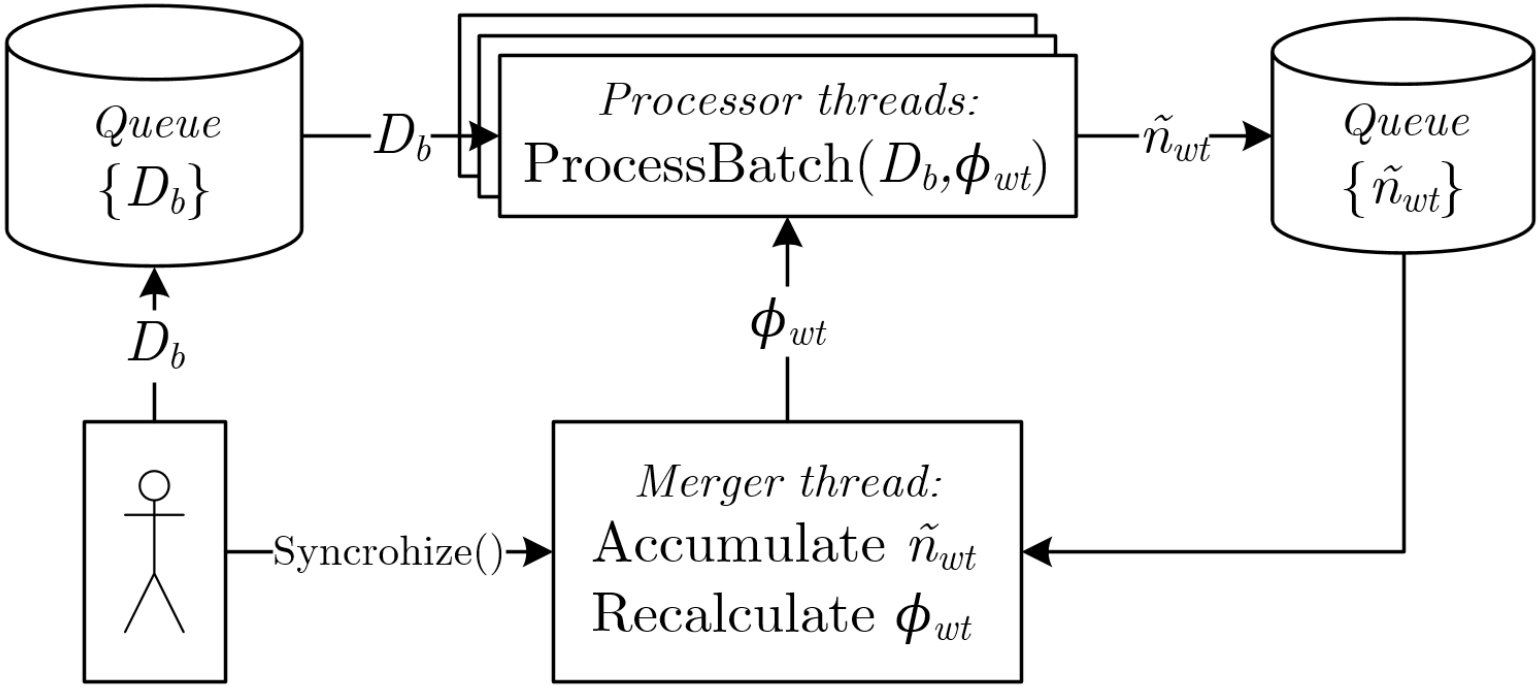
BigARTM community:

- 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

The BigARTM project: parallel architecture



- Concurrent processing of batches $D = D_1 \sqcup \cdots \sqcup D_B$
- Simple single-threaded code for *ProcessBatch*
- User controls when to update the model in online algorithm
- Deterministic (reproducible) results from run to run

Fast online EM-algorithm for regularized multimodal PTMs

Input: collection D split into batches D_b, $b = 1, \ldots, B$;
Output: matrix Φ;

initialize ϕ_{wt} for all $w \in W$, $t \in T$;
$n_{wt} := 0$, $\tilde{n}_{wt} := 0$ for all $w \in W$, $t \in T$;
for all batches D_b, $b = 1, \ldots, B$
    iterate each document $d \in D_b$ at a constant matrix Φ:
    $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \textbf{ProcessBatch}\,(D_b, \Phi)$;
    if (synchronize) **then**
        $n_{wt} := n_{wt} + \tilde{n}_{dw}$ for all $w \in W$, $t \in T$;
        $\phi_{wt} := \underset{w \in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)$ for all $w \in W^m$, $m \in M$, $t \in T$;
        $\tilde{n}_{wt} := 0$ for all $w \in W$, $t \in T$;

Fast online EM-algorithm for Multi-ARTM

ProcessBatch iterates documents $d \in D_b$ at a constant matrix Φ.

matrix $(\tilde{n}_{wt})$:= **ProcessBatch** (set of documents D_b, matrix Φ)

1 $\tilde{n}_{wt} := 0$ for all $w \in W$, $t \in T$;
2 **for all** $d \in D_b$
3 initialize $\theta_{td} := \frac{1}{|T|}$ for all $t \in T$;
4 **repeat**
5 $p_{tdw} := \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ for all $w \in d$, $t \in T$;
6 $n_{td} := \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw}$ for all $t \in T$;
7 $\theta_{td} := \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ for all $t \in T$;
8 **until** θ_d converges;
9 $\tilde{n}_{wt} := \tilde{n}_{wt} + \lambda_{m(w)} n_{dw} p_{tdw}$ for all $w \in d$, $t \in T$;

BigARTM vs Gensim vs Vowpal Wabbit

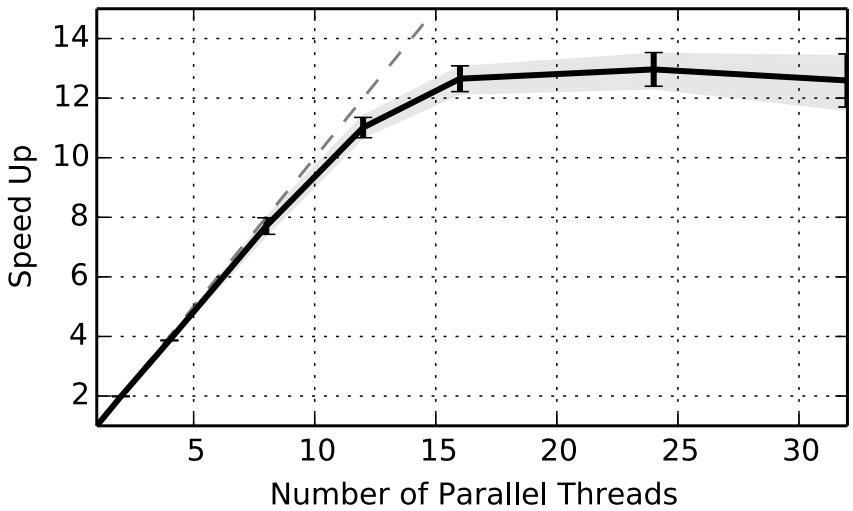
- 3.7M articles from Wikipedia, 100K unique words

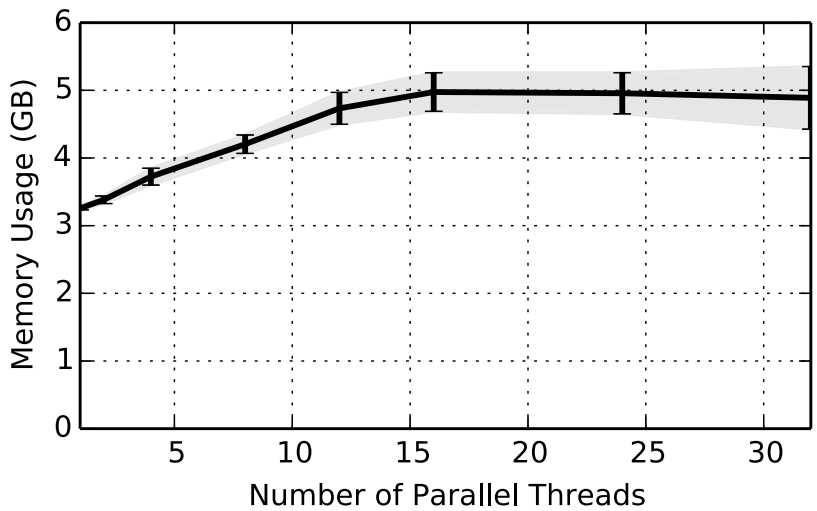
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

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

Running BigARTM in parallel

- 3.7M articles from Wikipedia, 100K unique words

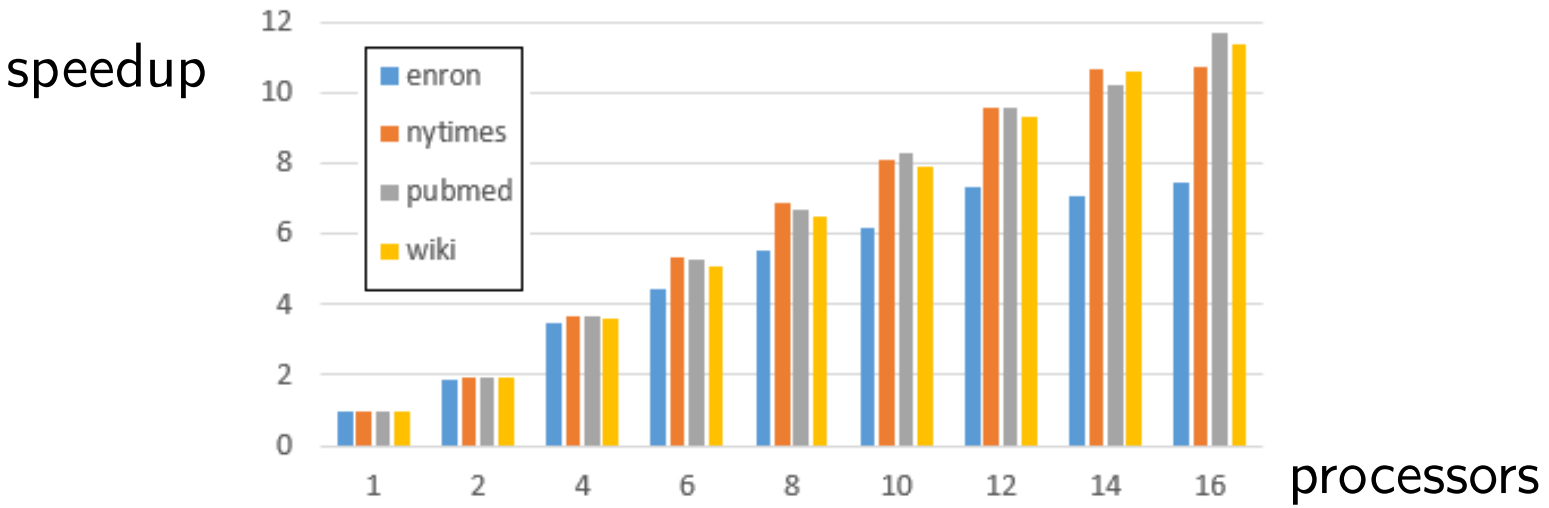




- Amazon EC2 c3.8xlarge (16 physical cores + hyperthreading)
- No extra memory cost for adding more threads

Running BigARTM on large collections

collection	$\|W\|$, 10^3	$\|D\|$, 10^6	n, 10^6	size, GB
enron	28	0.04	6.4	0.07
nytimes	103	0.3	100	0.13
pubmed	141	8.2	738	1.0
wiki	100	3.7	1009	1.2

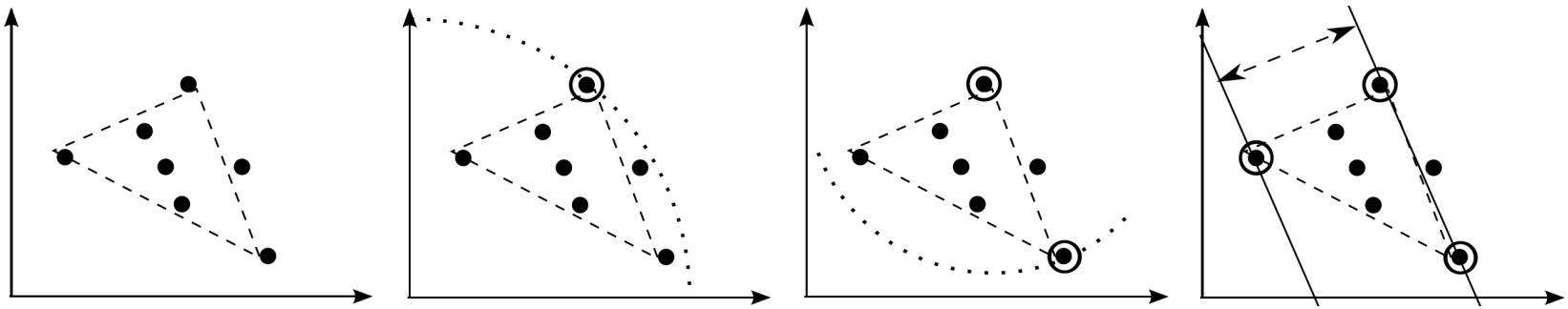


Amazon EC2 cc2.8xlarge instance:
16 cores + hyperthreading, Intel® Xeon® CPU E5-2670 2.6GHz.

Arora's algorithm based on anchor words recovery

Def. The word w is an *anchor word* of the topic t if $p(w|t) = 1$.

Arora's algorithm finds Φ with the identity submatrix of anchors



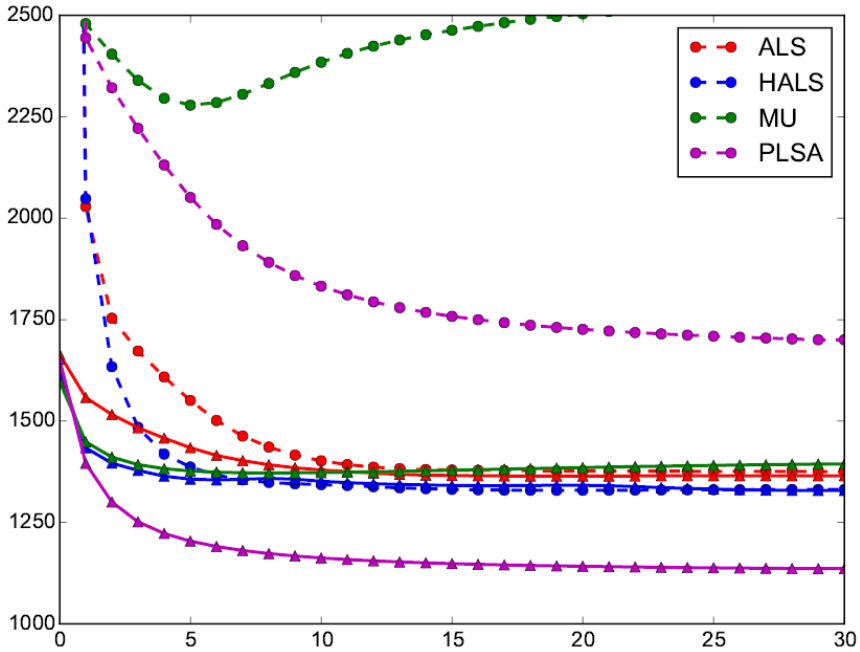
- ⊕ The fastest algorithm for Topic Modeling
- ⊕ Theoretical guarantees for polynomial time and global optimum
- ⊖ The hypothesis that $\forall t$ the anchor word exists is restrictive
- ⊖ The algorithm is not so fast for big vocabularies $|W|$

Sanjeev Arora et al. A Practical Algorithm for Topic Modeling with Provable Guarantees. ICML 2013.

Arora's algorithm for EM-algorithms initialization

ALS, HALS, MU — methods for nonnegative matrix factorization
NIPS collection: $|D| = 1500$, $|W| = 12419$, $|T| = 25$.
Perplexity $= \exp\left(-\frac{1}{n}\mathscr{L}(\Phi, \Theta)\right)$

perplexity



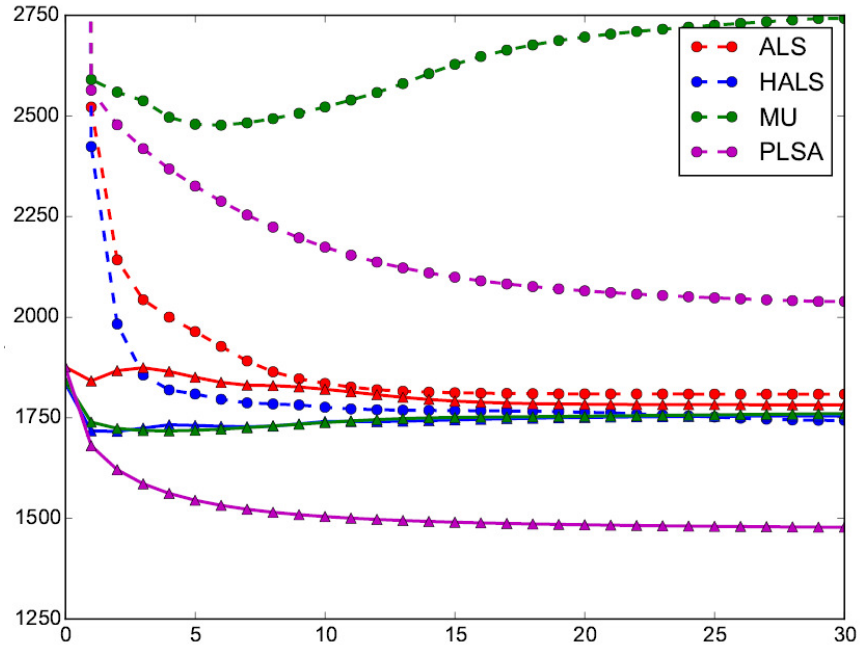
iterations

solid lines — initialization by Arora's algorithm
dotted lines — random initialization

Arora's algorithm for EM-algorithms initialization

ALS, HALS, MU — methods for nonnegative matrix factorization
Daily Kos collection: $|D| = 3430$, $|W| = 6906$, $|T| = 25$.
Perplexity $= \exp\left(-\frac{1}{n}\mathscr{L}(\Phi, \Theta)\right)$

perplexity



iterations

solid lines — initialization by Arora's algorithm
dotted lines — random initialization

Conclusions about Arora's initialization

- Arora's initialization greatly improves PLSA
  (PLSA — Probabilistic Latent Semantic Analysis is equivalent to ARTM without regularization)
- Arora's initialization does not improve quadratic loss minimizers
- PLSA is capable of improving the Arora's initialization, perhaps, because of restrictive assumptions of Arora's algorithm do not hold in the real data

Regularization for topic selection

Let us maximize KL-divergence: $\mathrm{KL}\left(\frac{1}{|T|} \,\|\, p(t)\right) \to \max$
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:

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

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

The experiments with topic selection

Real dataset: NIPS (Neural Information Processing System)

- $|D| = 1566$ preprocessed papers from NIPS conference;
- vocabulary: $|W| \approx 1.3 \cdot 10^4$; hold-out set: $|D'| = 174$.

Synthetic dataset:

- 500 EM iterations for PLSA with $|T_0| = 50$ topics on NIPS
- generate synthetic dataset (n^0_{dw}) using obtained Φ and Θ:

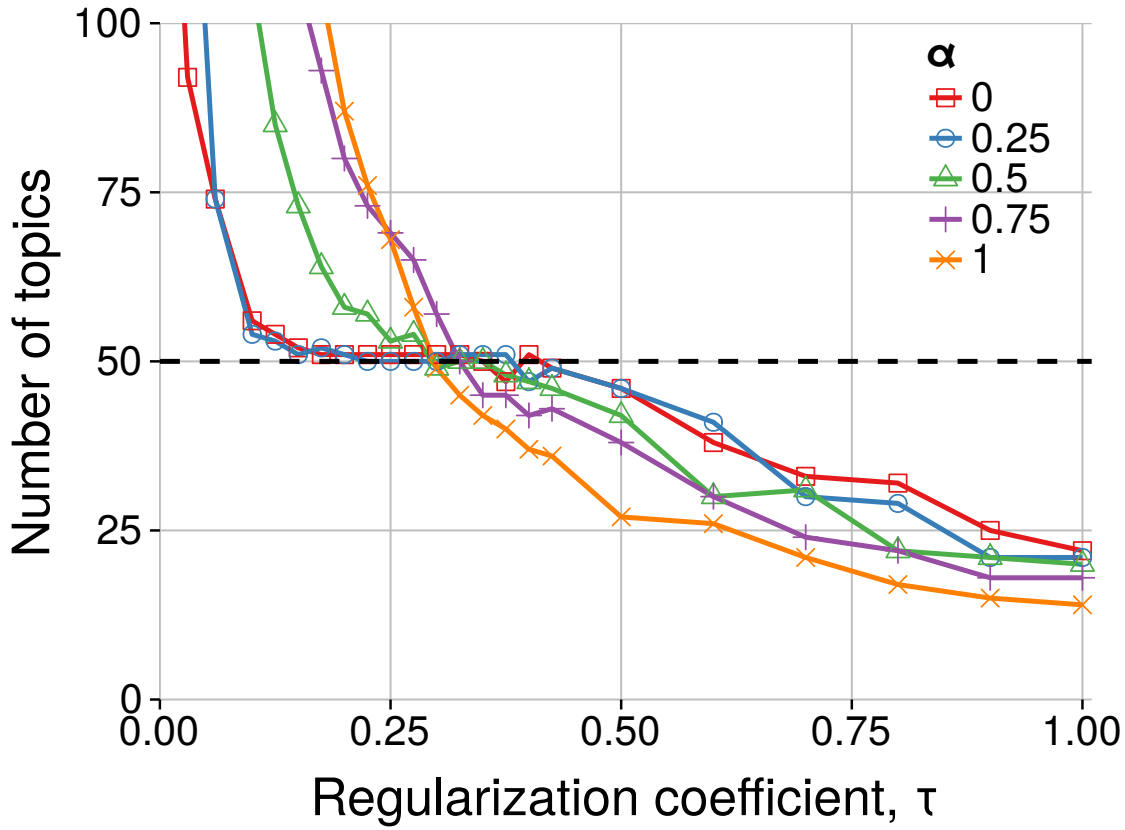
$$n^0_{dw} = n_d \sum_{t \in T} \phi_{wt} \theta_{td}$$

Parametric family of semi-real datesets:

- (n^α_{dw}) is a mixture of synthetic (n^0_{dw}) and real (n_{dw}) datasets:

$$n^\alpha_{dw} = \alpha n_{dw} + (1 - \alpha) n^0_{dw}$$

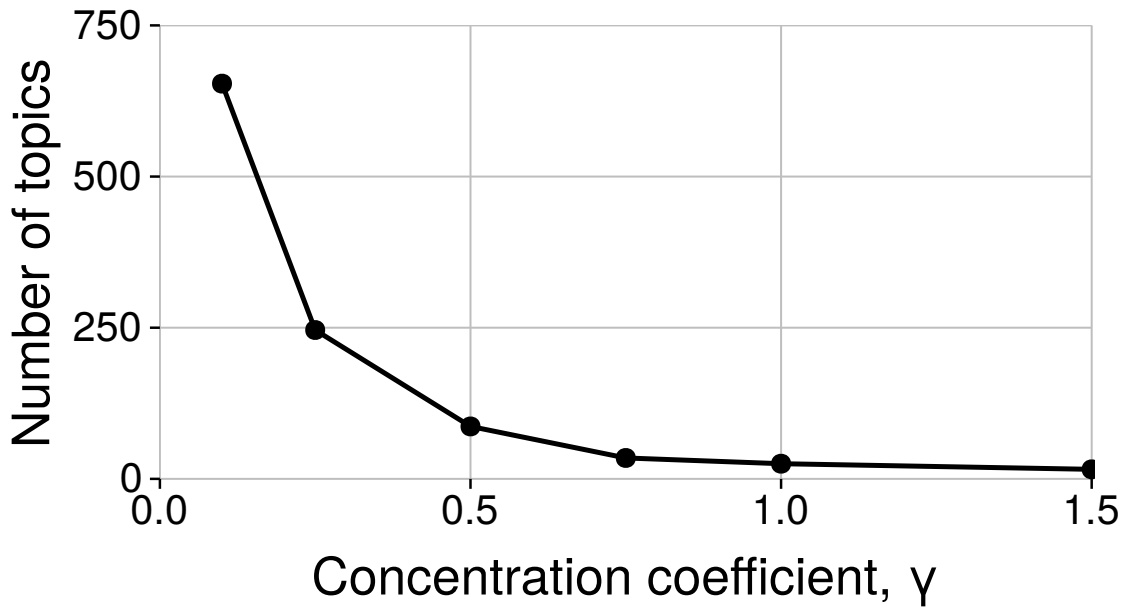
Number of topics determination



- For synthetic dataset ARTM reliably finds the truth: $|T| = 50$.
- The range of τ values leading to the correct number is wide.
- For real data the number of topics is not clear.

Comparison to HDP topic model

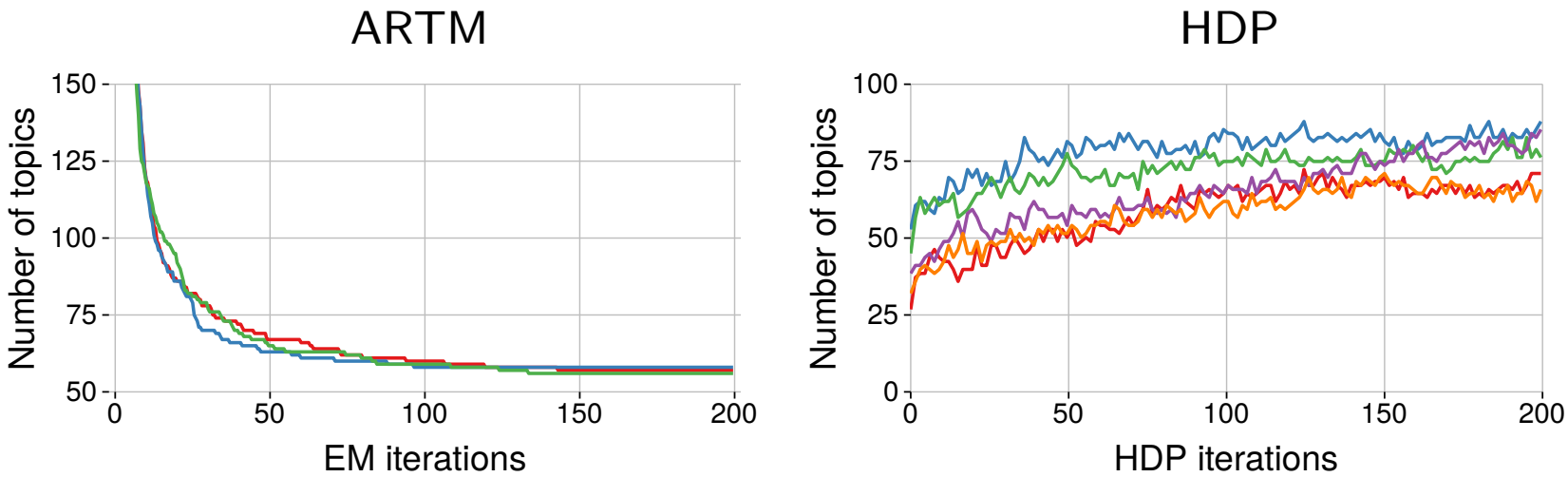
HDP (Hierarchical Dirichlet Process, Tech et. al, 2006)
is the state-of-art approach for a number of topics optimization.



- The choice of the concentration coefficient γ of Dirichlet process may lead to nearly any number of topics.

Stability of ARTM vs HDP

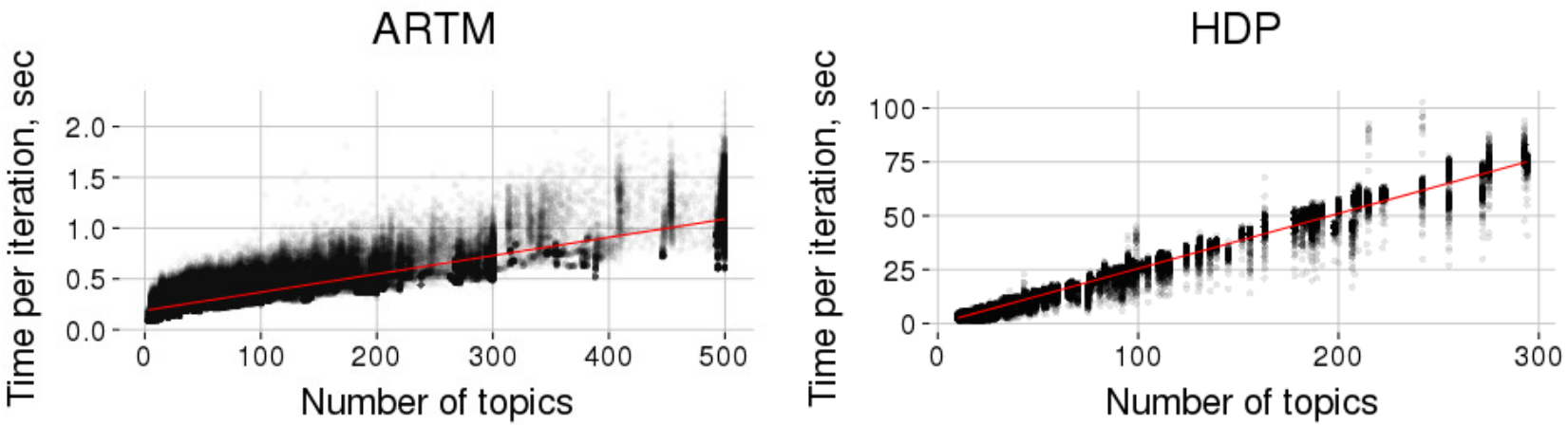
Starting ARTM and HDP many times from different initializations:



1. HDP is less stable in two ways:
   1. The number of topics fluctuates from iteration to iteration
   2. The results for several random starts significantly differ
2. The "recommended" parameters γ for HDP and τ for ARTM give the similar number of topics ≈ 60

Running time of ARTM vs HDP

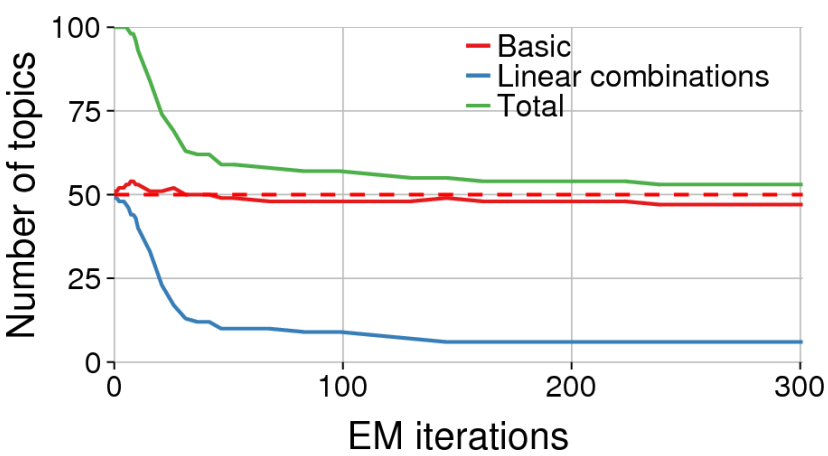
Comparing the running time per iteration (sec) of ARTM vs HDP (one iteration = one pass through the collection)

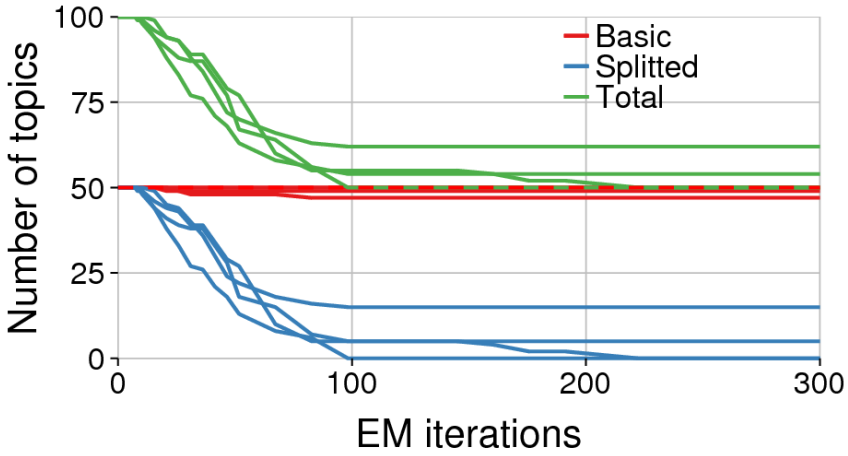


- Our method is 100 times faster!

Elimination of linearly dependent and split topics

- Add 50 linear combinations of topics in synthetic dataset.
- Add 50 subtopics (split topics) in synthetic dataset.





- Our regularizer effectively eliminates both linearly dependent and split topics from the model
- More diverse topics of the original model remain.

Conclusions about number of topics

- It seems that the "true number of topics" does not exist in real text collections.
- ARTM has a special regularizer for topic selection, which eliminates small, linearly dependent, and nested topics.
- It is faster and more stable than state-of-the-art HDP.

Challenging open problem

- How to choose a multidimensional vector of regularization coefficients τ_i and λ_m?
- How to choose a regularization path in the multidimensional regularization space?

Difficulties:

- There is no differentiable quality measures to optimize
- All interesting measures are expensive to compute

Known recommendations and empirical findings:

- Tend them to zero (Tikhonov's theory of regularization)
- Apply interfering regularizers alternately at different iterations
- Renormalization of regularization coefficients

Renormalization of regularization coefficients

Consider the formula of the regularized M-step:

$$\phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \sum_{i=1}^{k} \tau_i \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}} \right).$$

The *impact of regularizer* R_i on a topic t and on a whole collection:

$$r_{it} = \sum_{w \in W} \left| \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}} \right|, \qquad r_i = \sum_{t \in T} r_{it}.$$

Renormalization of regularization coefficient τ_i:

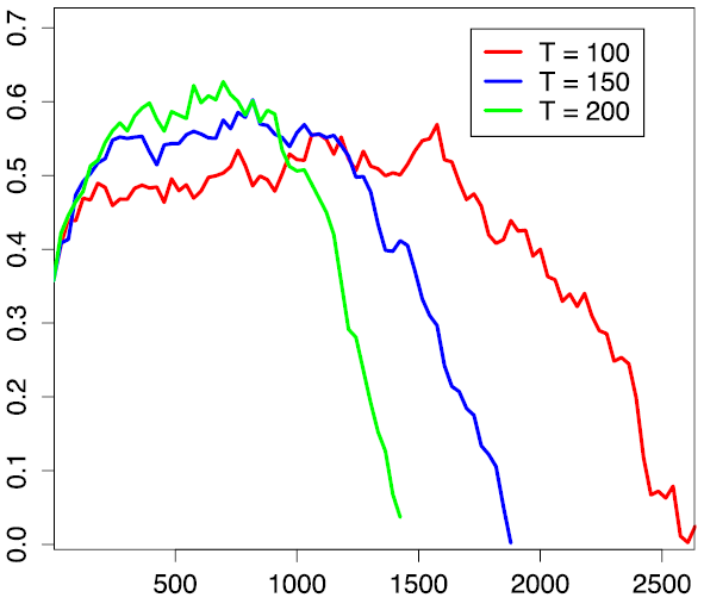
$$\phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} + \sum_{i=1}^{k} \tau_i \left(\gamma_i \frac{n_t}{r_{it}} + (1 - \gamma_i) \frac{n}{r_i} \right) \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}} \right),$$

where γ_i is *the degree of impacts individualization*.

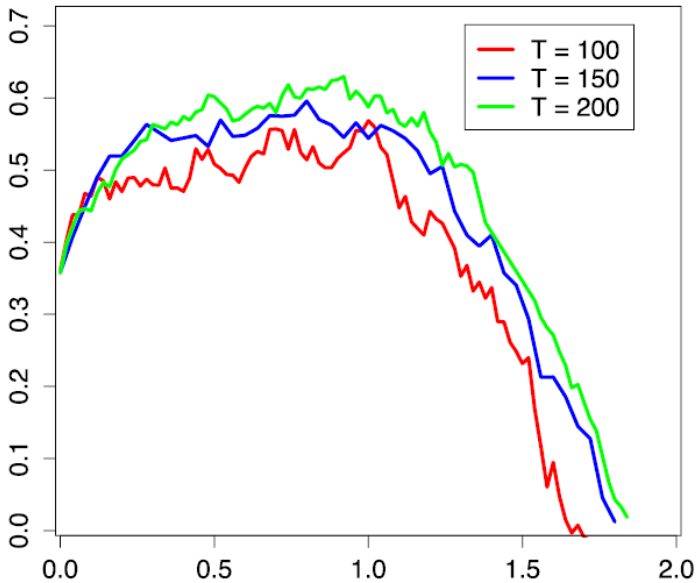
Renormalization of regularization coefficients

Collection of press releases, $|D| = 1000$
How the coherence measure depends of sparsing coefficient:

without renormalization



with renormalization



- Normalized coefficients take values in $[0, 1]$
- The recommendations for normalized coefficients can be easily transferred from one problem to another

Conclusions

- Topic Modeling is an applied area of optimization and matrix factorization in text analysis
- ARTM (Additive Regularization) is a semi-probabilistic non-Bayesian multicriteria view on Topic Modeling
- BigARTM is open source project for parallel online multimodal regularized Topic Modeling of large collections
- Open problem: can we apply Control Theory for choosing regularization path adaptively?

Contacts:
Konstantin Vorontsov: voron@forecsys.ru
Wiki www.MachineLearning.ru (in Russian)