Additive Regularization of Topic Models and its parallel implementation BigARTM.org

Anna Potapenko

National Research University Higher School of Economics,
Yandex School of Data Analysis, Moscow, Russia

May 21, 2015

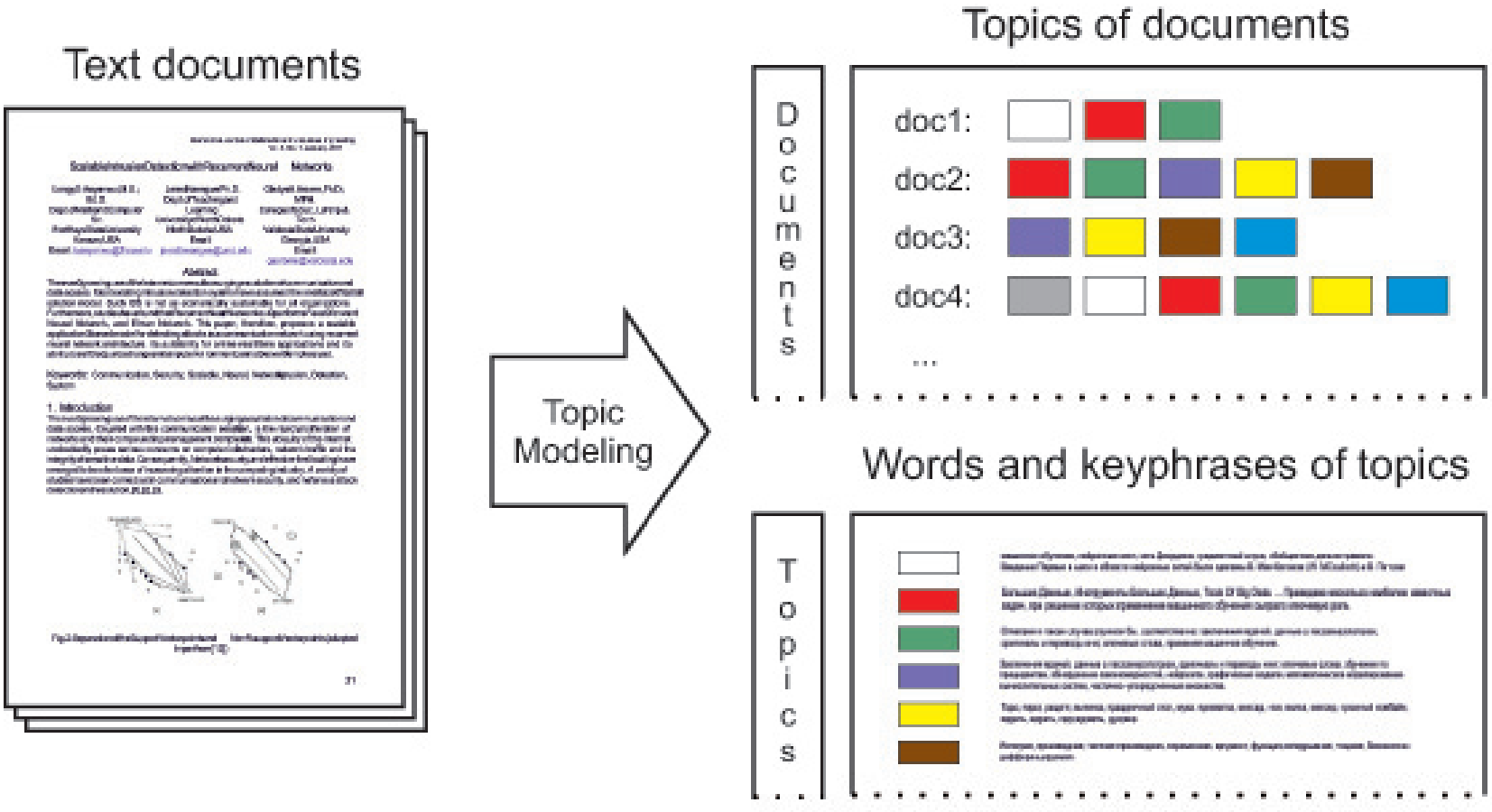
Outline:

Theory: PTM, ARTM, Multi-ARTM
- Probabilistic topic modeling
- Additive regularization of topic models
- Multimodal probabilistic topic modeling

Implementation: BigARTM.org
- Fast online EM-algorithm
- BigARTM architecture and performance
- Comparison to other packages

Experiments: making use of regularization
- Topic selection on semi-real datasets
- Sparse decorrelated topic model
- Topic modeling for classification task
- Multi-language topic model

Topic modeling – revealing a hidden thematic structure of a text collection – soft clustering of words and documents:



Probabilistic statement of the problem

Given:

- D — a collection of documents; W — a vocabulary of terms;
- $p(w|d)$ — frequencies of terms w in documents d

Find:

- $\phi_{wt} = p(w|t)$ — a distribution over terms for each topic
- $\theta_{td} = p(t|d)$ — a distribution over topics for each document

Basic assumptions:

- A *topic* is a set of coherent terms (words or phrases) that often co-occur in the documents.
- A document is a *bag of words*.
- Each *observed* word in a document has a *latent* topic.

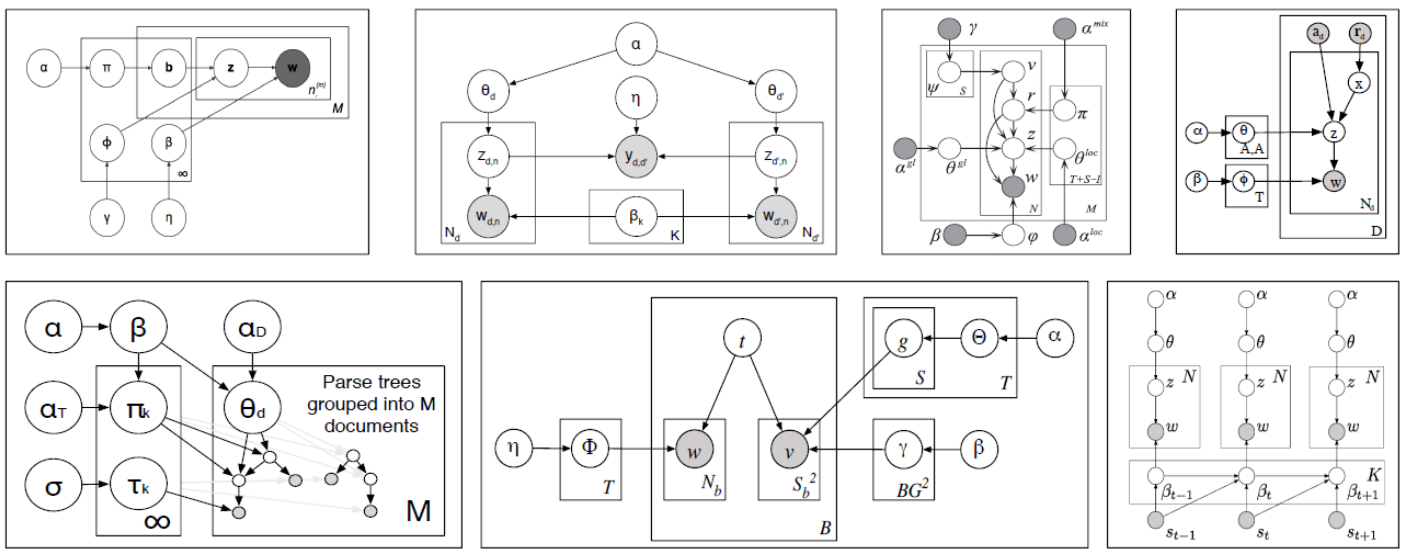
Applications of topic modeling

The compressed semantic representation of texts is used for:

- Information retrieval for long-text queries
- Semantic search in large scientific document collections
- Revealing research trends and research fronts
- Expert search
- News aggregation
- Recommender systems
- Categorization, classification, summarization, segmentation of texts, images, video, signals, social media
- and many others

Probabilistic Topic Modeling: milestones and mainstream

1. PLSA — Probabilistic Latent Semantic Analysis (1999)
2. LDA — Latent Dirichlet Allocation (2003)
3. 100s of PTMs based on Graphical Models & Bayesian Inference



David Blei. Probabilistic topic models // Communications of the ACM, 2012. Vol. 55. No. 4. Pp. 77–84.

PLSA: Probabilistic Latent Semantic Analysis [T. Hofmann 1999]

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

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

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

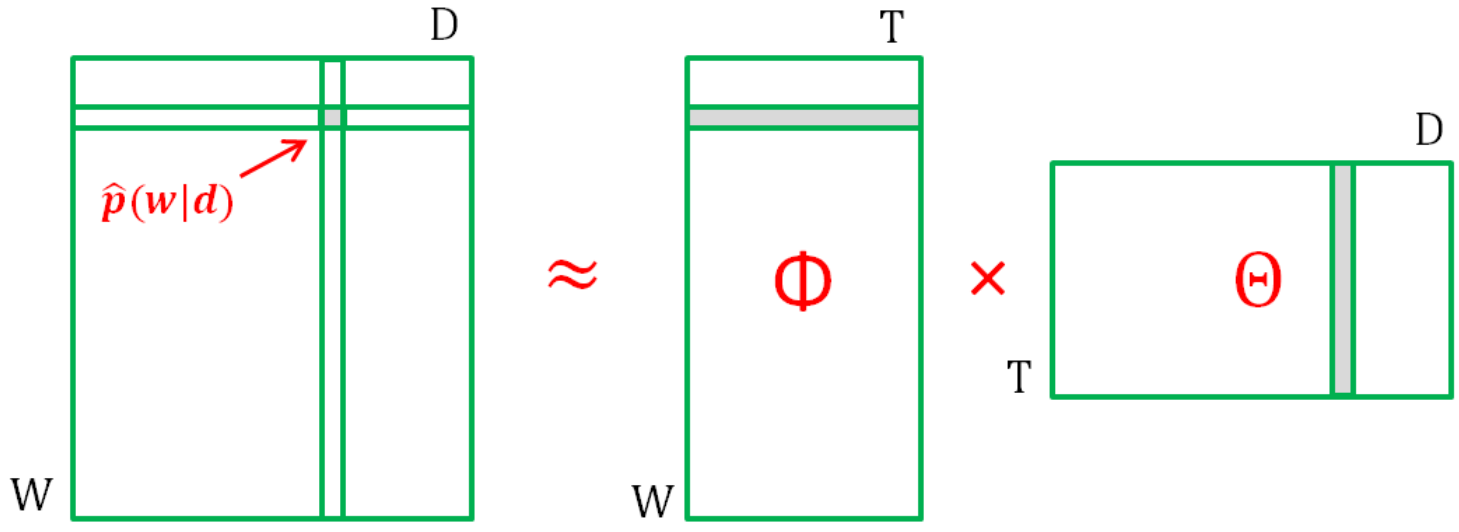
$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta},$$

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

Solution is obtained via iterations of EM-algorithm.

Topic Modeling is an ill-posed inverse problem

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



Matrix factorization is not unique, the solution is not stable:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$$

for all S such that $\Phi' = \Phi S$, $\Theta' = S^{-1}\Theta$ are stochastic.

Then, regularization is needed to find appropriate solution.

ARTM: Additive Regularization of Topic Model

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

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

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

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

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, Springer CCIS, 2014. Vol. 436. pp. 29-46.

Regularized EM-algorithm

Theorem. If Φ, Θ is the solution of the regularized likelihood maximization problem, then it satisfies the following system of equations with auxiliary variables $p_{dwt} = p(t|d,w)$:

$$\begin{cases} \text{E-step:} \quad p_{dwt} = \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}}; \\ \text{M-step:} \\ \phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right); \quad n_{wt} = \sum\limits_{d\in D} n_{dw}p_{dwt}; \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right); \quad n_{td} = \sum\limits_{w\in d} n_{dw}p_{dwt}; \end{cases}$$

where $\underset{t\in T}{\text{norm}}\, x_t = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ is non-negative normalization;

Regularization: smoothing

- *Hypothesis:* the distributions are close to the fixed ones:

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

- We maximize the sum of these regularizers:

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

- According to the theorem, we obtain the M-step formulas:

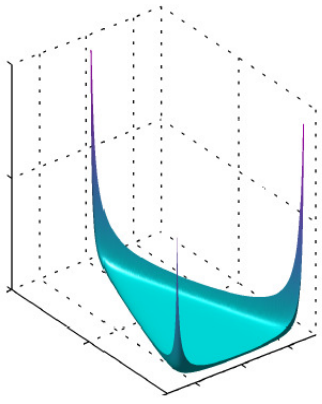
$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}\big(n_{wt} + \beta_0\beta_w\big), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} + \alpha_0\alpha_t\big).$$

Latent Dirichlet Allocation

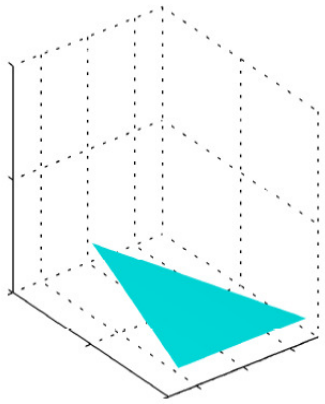
Hypothesis. Columns $\phi_t = (\phi_{wt})_{w \in W}$ and $\theta_d = (\theta_{td})_{t \in T}$ are generated from Dirichlet distribution, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

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

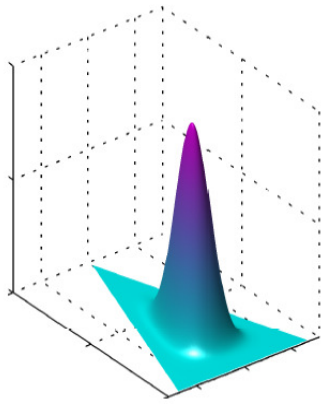
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t > 0;$$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$



$\alpha_1 = \alpha_2 = \alpha_3 = 1$



$\alpha_1 = \alpha_2 = \alpha_3 = 10$

Latent Dirichlet Allocation

Inference techniques:

- Variational Bayes
- Gibbs Sampling
- Maximum a posterior probability

M-step for LDA MAP:

$$\phi_{wt} = \underset{w \in W}{\text{norm}}\big(n_{wt} + \beta_w - 1\big), \qquad \theta_{td} = \underset{t \in T}{\text{norm}}\big(n_{td} + \alpha_t - 1\big),$$

where $\alpha_t > 0$ and $\beta_w > 0$ are Dirichlet hyperparemters.

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation //
Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

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.

Regularization: sparsing

- *Hypothesis:* the distributions are sparse (i.e. far from uniform):

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

- According to the theorem, we obtain the M-step formulas:

$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}\big(n_{wt} - \beta_0\beta_w\big), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} - \alpha_0\alpha_t\big).$$

Smoothing and sparsing together provide a generalization of LDA model with Dirichlet hyperparameters from $-\infty$ to ∞.

Regularization: topic selection

- *Hypothesis:* if a topic has a poor terminology, it is excessive.
- To make the distribution $p(t) = \sum_d p(d)\theta_{td}$ sparse maximize *KL-divergence* between $p(t)$ and the uniform distribution:

$$R(\Theta) = -\tau \sum_{t \in T} \ln \sum_{d \in D} p(d)\theta_{td} \to \max.$$

- M-step formulas (setting to zeros the rows of Θ for topics with small $n_t = \sum_{w \in W} n_{wt}$):

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

Regularization: decreasing topic correlation

- *Hypothesis:* making topics more different makes them more interpretable.
- We maximize a sum of pairwise co-variations between column vectors ϕ_t:

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

- M-step formulas (making the rows of Φ more contrast):

$$\phi_{wt} = \underset{w \in W}{\text{norm}} \Big(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \Big).$$

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

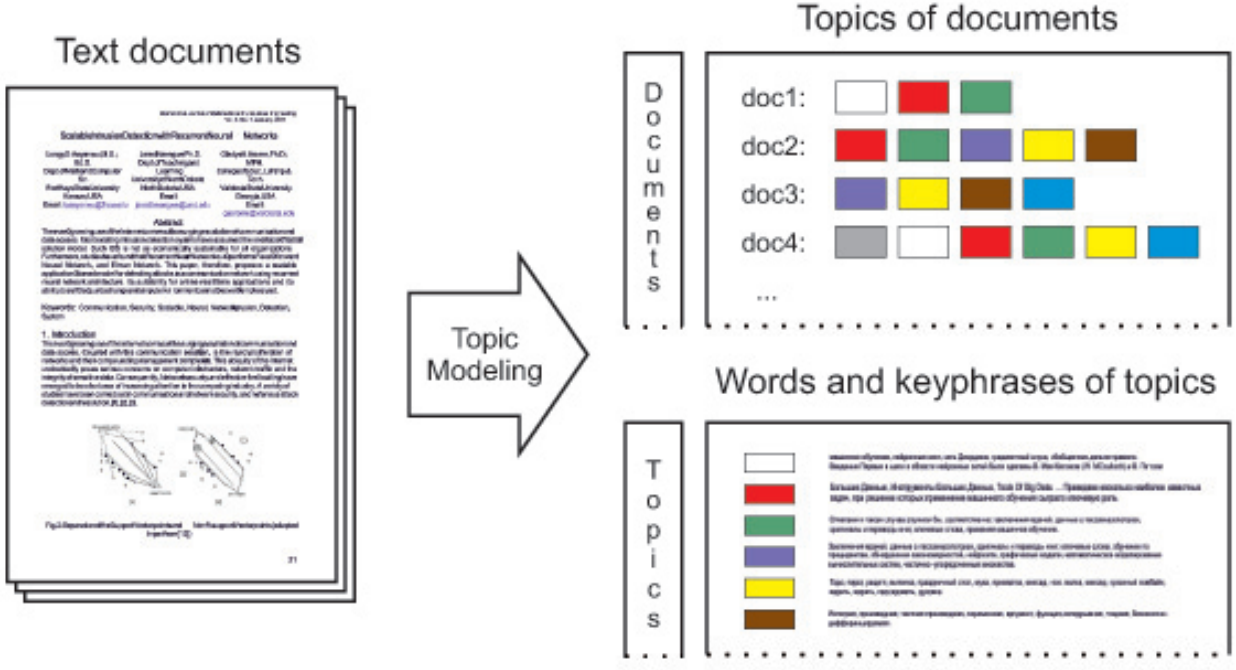
ARTM: available regularizers

- topic smoothing (equivalent to LDA)
- 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
- and many others

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Special Issue "Data Analysis and Intelligent Optimization with Applications". Springer, 2014.

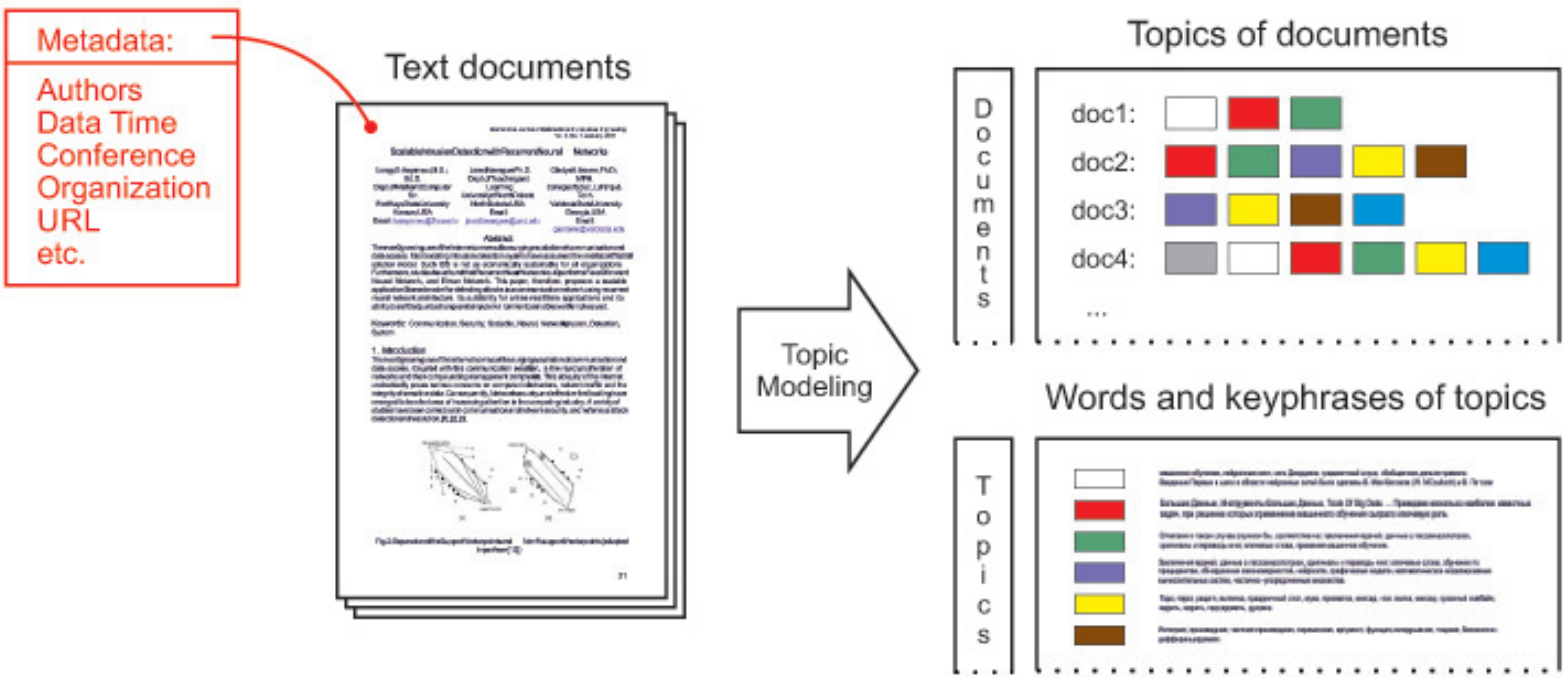
Multimodal Probabilistic Topic Modeling

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.



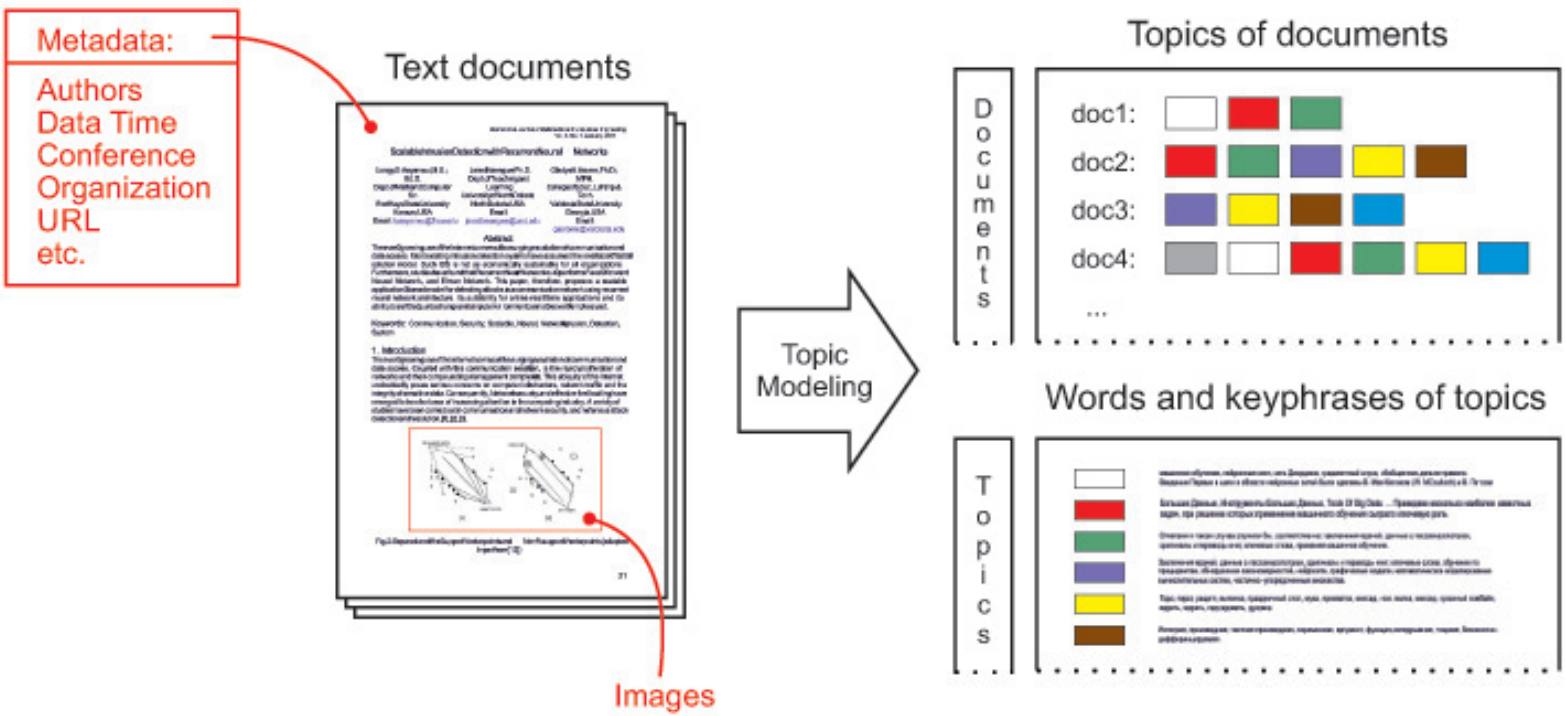
Multimodal Probabilistic Topic Modeling

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



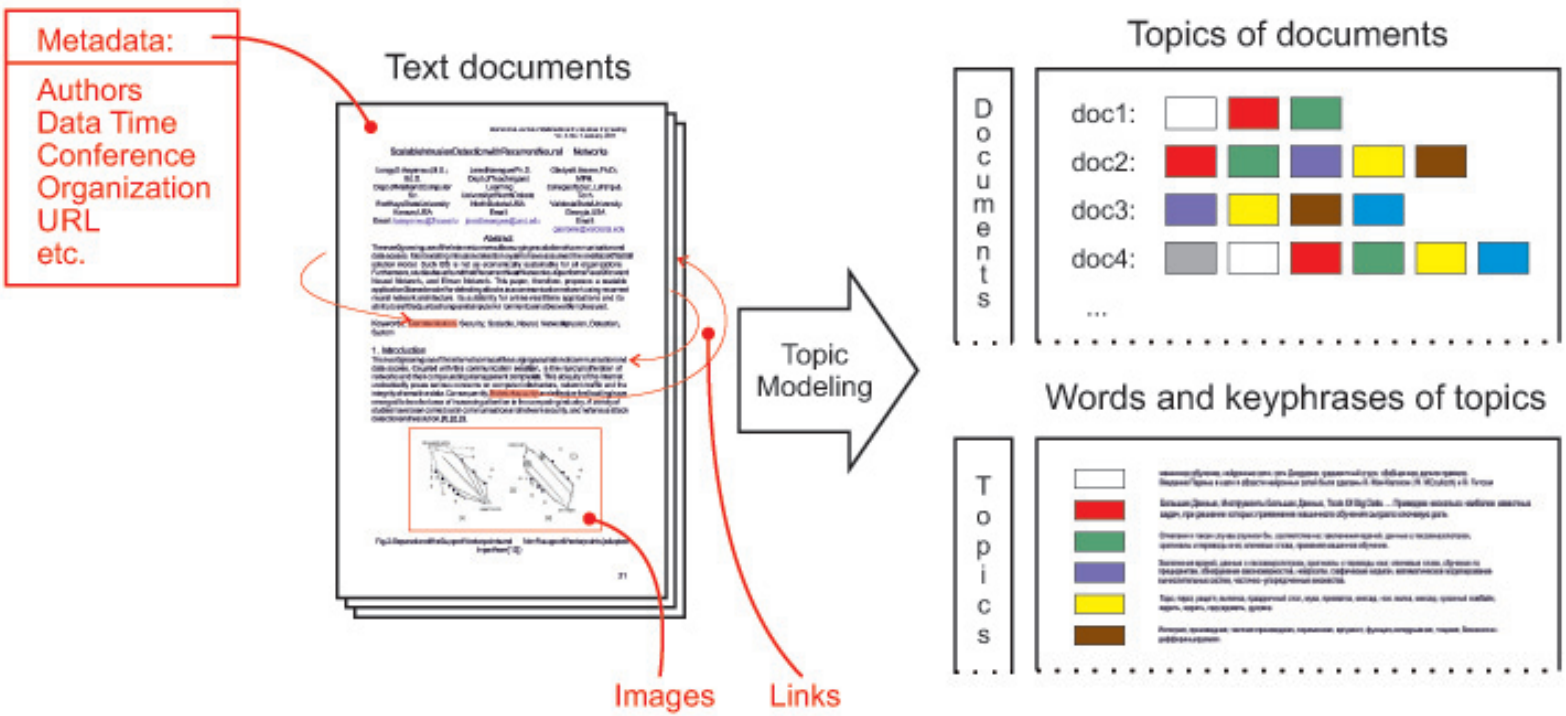
Multimodal Probabilistic Topic Modeling

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



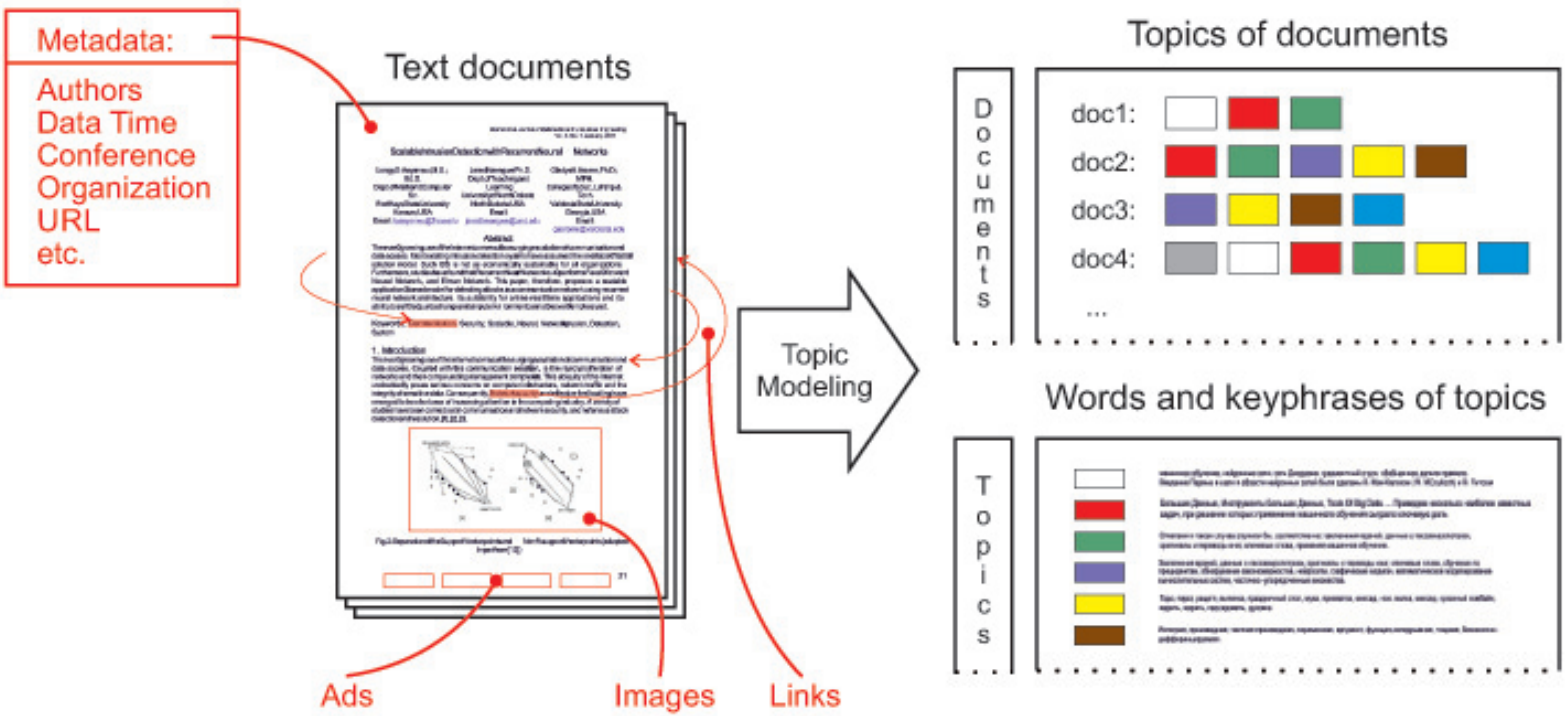
Multimodal Probabilistic Topic Modeling

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 $p(d'|t)$,



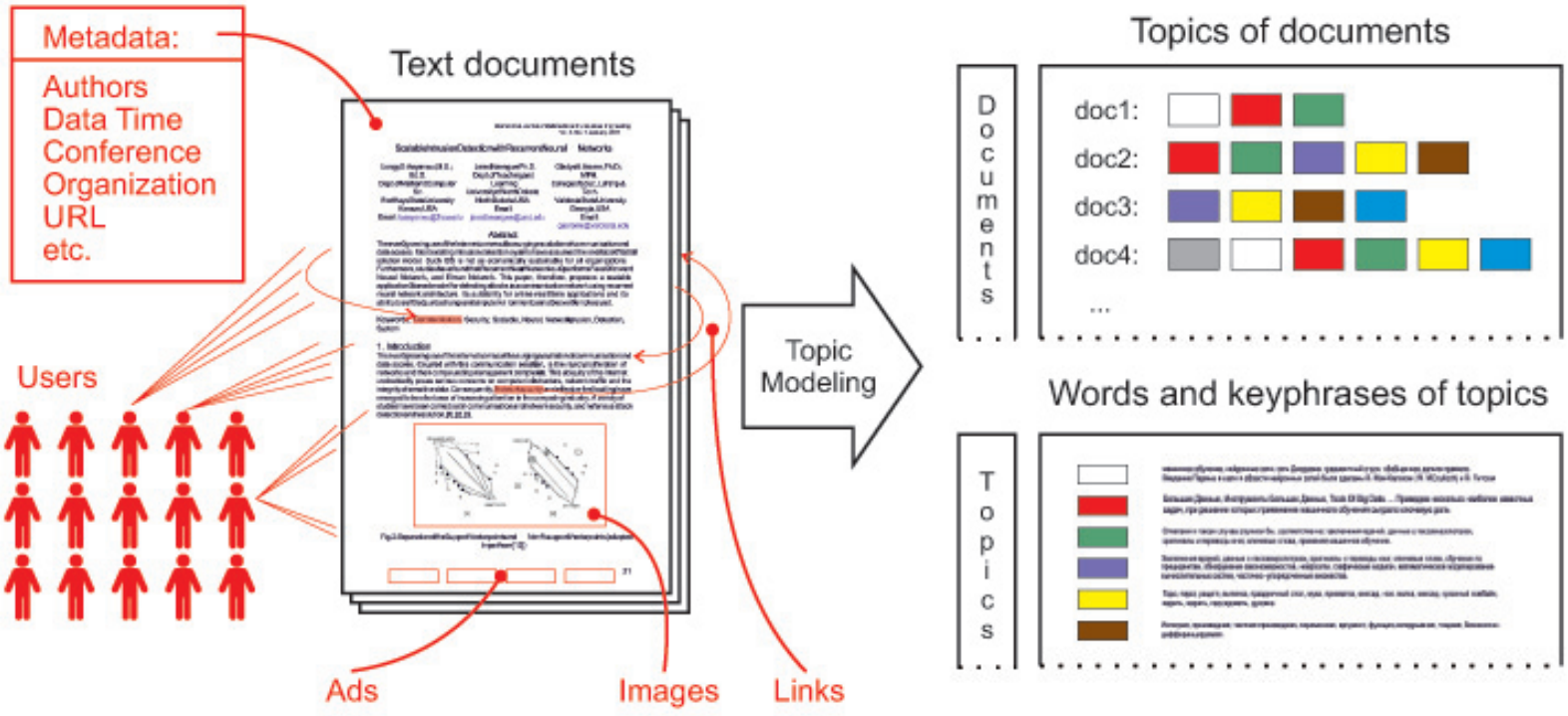
Multimodal Probabilistic Topic Modeling

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 $p(b|t)$,



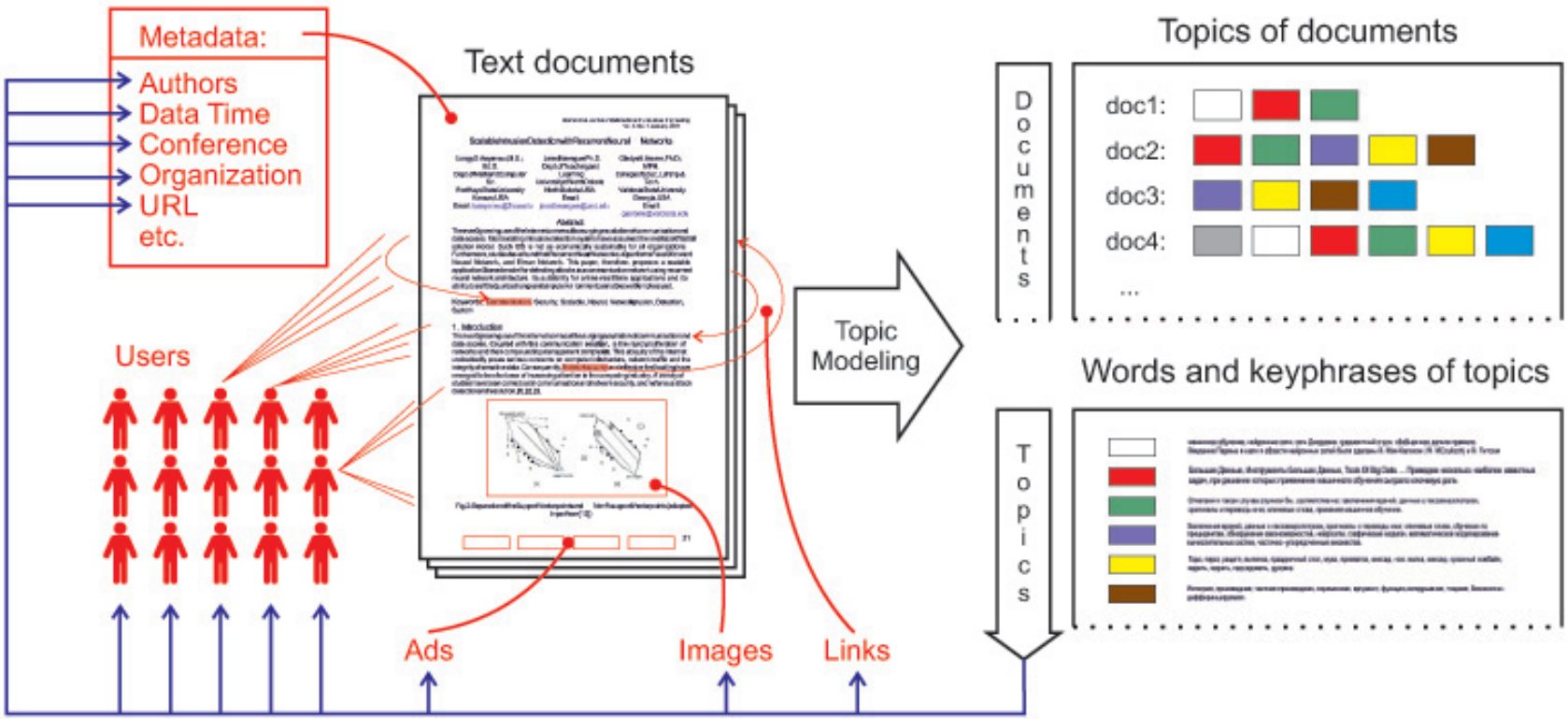
Multimodal Probabilistic Topic Modeling

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 $p(u|t)$,



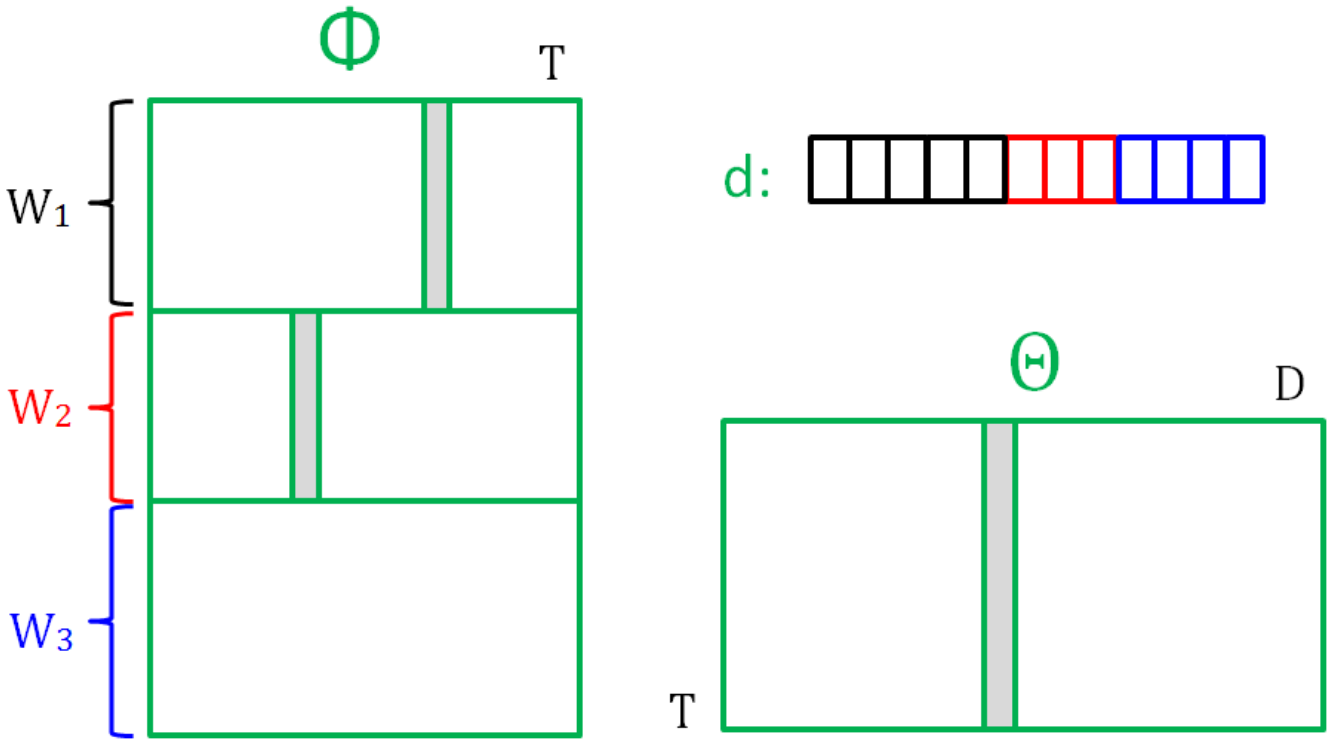
Multimodal Probabilistic Topic Modeling

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.



Multi-ARTM: combining multimodality with 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



Multi-ARTM: combining multimodality with regularization

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

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

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

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

K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Dudarenko — BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // AIST'2015 (to appear).

Multi-ARTM: multimodal regularized EM-algorithm

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 $w \in W^{m(w)}$.

ARTM approach: benefits and restrictions

Benefits

- Single EM-algorithm for many models and their combinations
- PLSA, LDA, and 100s of PTMs are covered by ARTM
- No complicated inference and graphical models
- ARTM reduces barriers to entry into PTM research field
- ARTM encourages any combinations of regularizers
- Multi-ARTM encourages any combinations of modalities
- Multi-ARTM is implemented in BigARTM open-source project

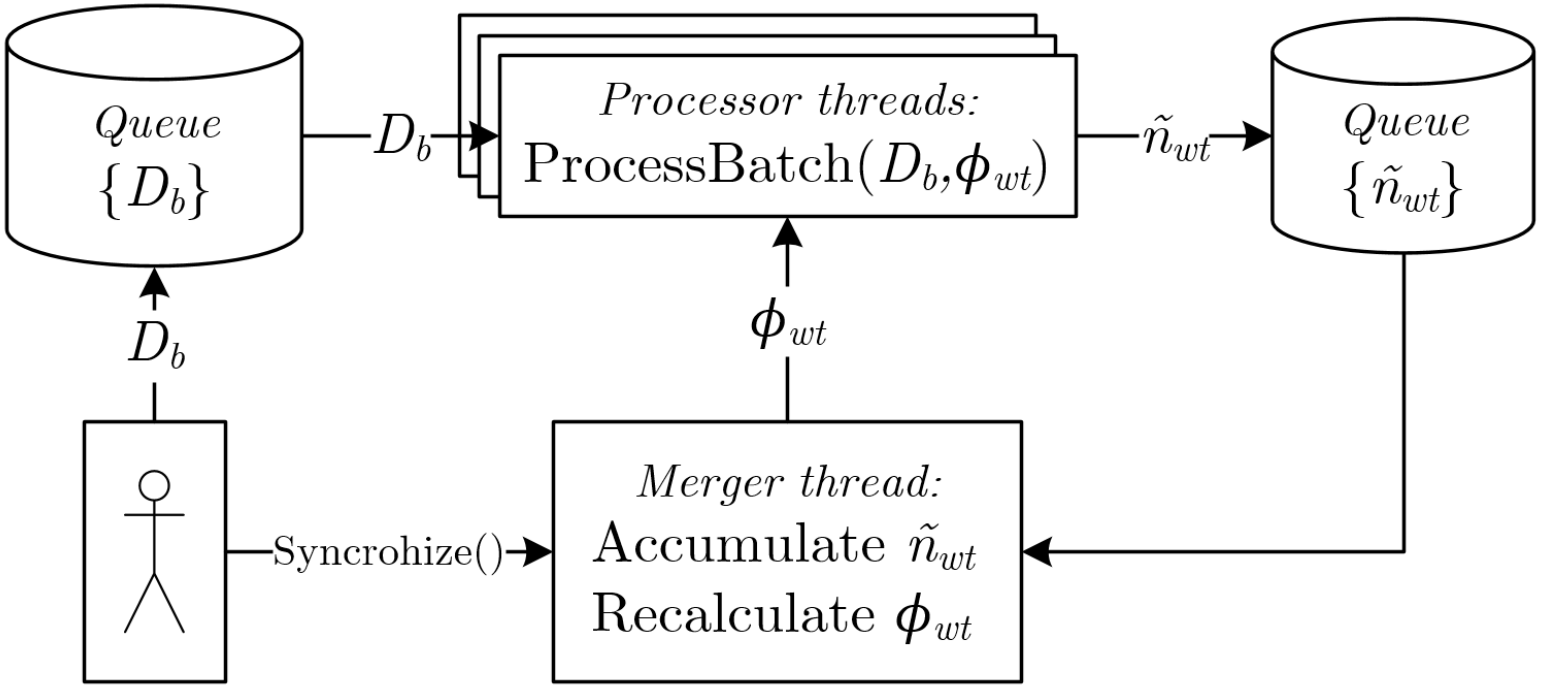
Under development (not really restrictions):

- 3-matrix factorization $P = \Phi\Psi\Theta$, e.g. Author-Topic Model
- Further generalization of hypergraph-based Multi-ARTM
- Adaptive optimization of regularization coefficients

The BigARTM project: main features

- Parallel online Multi-ARTM framework
- Open-source http://bigartm.org
- Distributed storage of collection is possible
- Built-in regularizers:
  - smoothing, sparsing, decorrelation, semi-supervised learning, and many others coming soon
- Built-in quality measures:
  - perplexity, sparsity, kernel contrast and purity, and many others coming soon
- Many types of PTMs can be implemented via Multi-ARTM:
  - multilanguage, temporal, hierarchical, multigram, and many others

The BigARTM project: parallel architecture



- Concurrent processing of batches
- 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 Multi-ARTM: master's routine

Input: collection D; discounting factor $\rho \in (0, 1]$;
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, $b = 1, \ldots, B$
  $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \textbf{ProcessBatch}\,(D_b, \Phi)$;
  if (synchronize) **then**
    $n_{wt} := \rho n_{wt} + \tilde{n}_{dw}$ 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 $w \in W$, $t \in T$;

M. D. Hoffman, D. M. Blei, F. Bach — Online Learning for Latent Dirichlet Allocation, NIPS 2010.

Fast online EM-algorithm for Multi-ARTM: node's routine

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

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

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

BigARTM vs Gensim vs Vowpal Wabbit

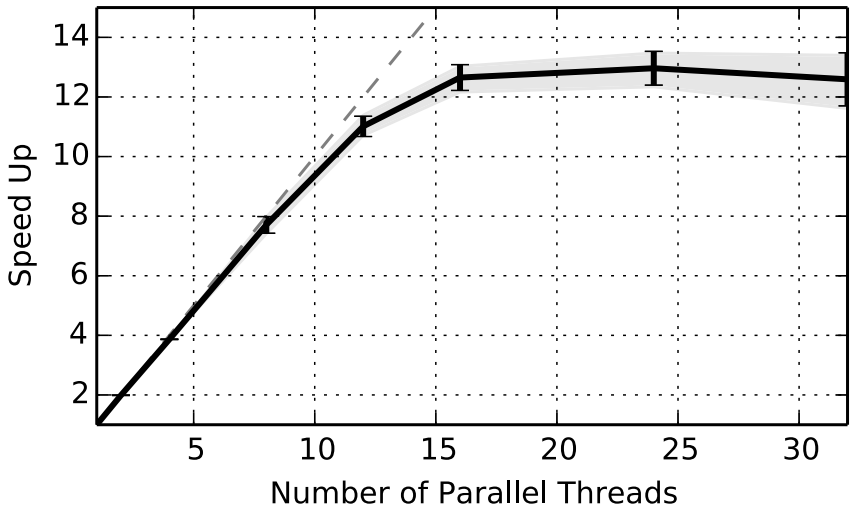
- 3.7M articles from Wikipedia, 100K unique words

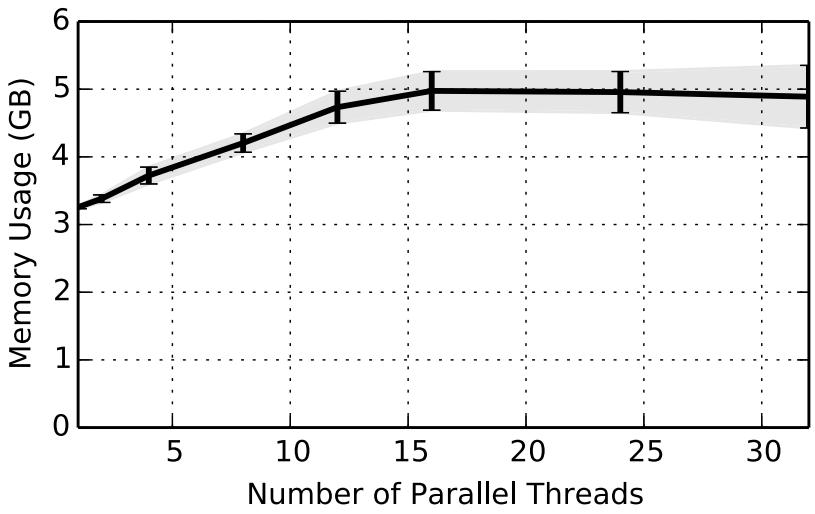
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *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

How to join the BigARTM project

BigARTM community:

1. Post questions in BigARTM discussion group:
   https://groups.google.com/group/bigartm-users
2. Report bugs in BigARTM issue tracker:
   https://github.com/bigartm/bigartm/issues
3. Contribute to BigARTM project via pull requests:
   https://github.com/bigartm/bigartm/pulls

License and programming environment

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Simple command-line API — available now
- Rich programming API in C++ and Python — available now
- Rich programming API in C# and Java — coming soon

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 topics on NIPS
- generate $\Pi_0 = (n^0_{dw})$ using obtained Φ and Θ:

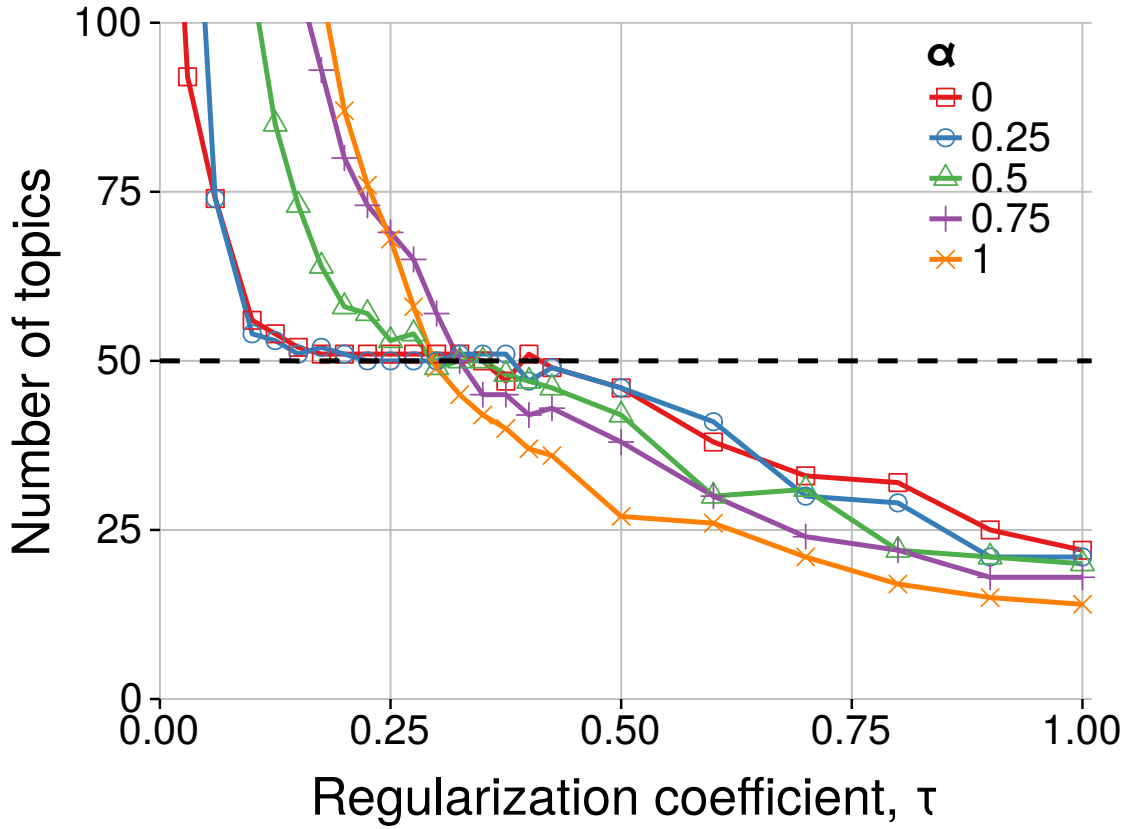
$$n^0_{dw} = n_d \sum_{t \in T} \phi_{wt} \theta_{td}$$

Parametric family of semi-real datesets:

- $\Pi_\alpha = (n^\alpha_{dw})$ is a mixture of Π_0 and NIPS n_{dw} counters:

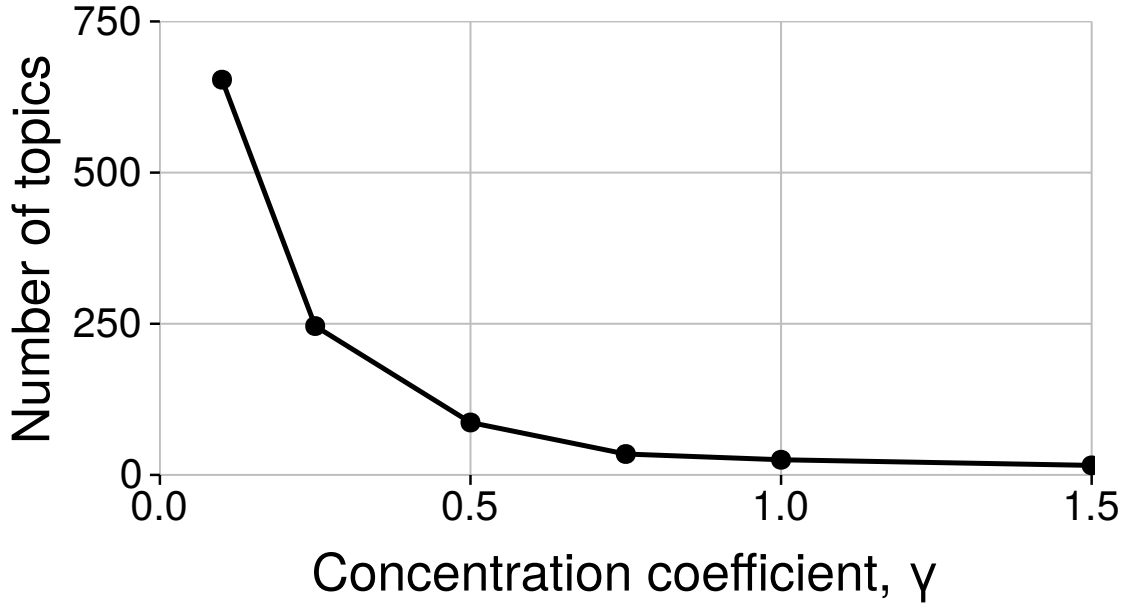
$$n^\alpha_{dw} = \alpha n_{dw} + (1 - \alpha) n^0_{dw},$$

Number of topics determination



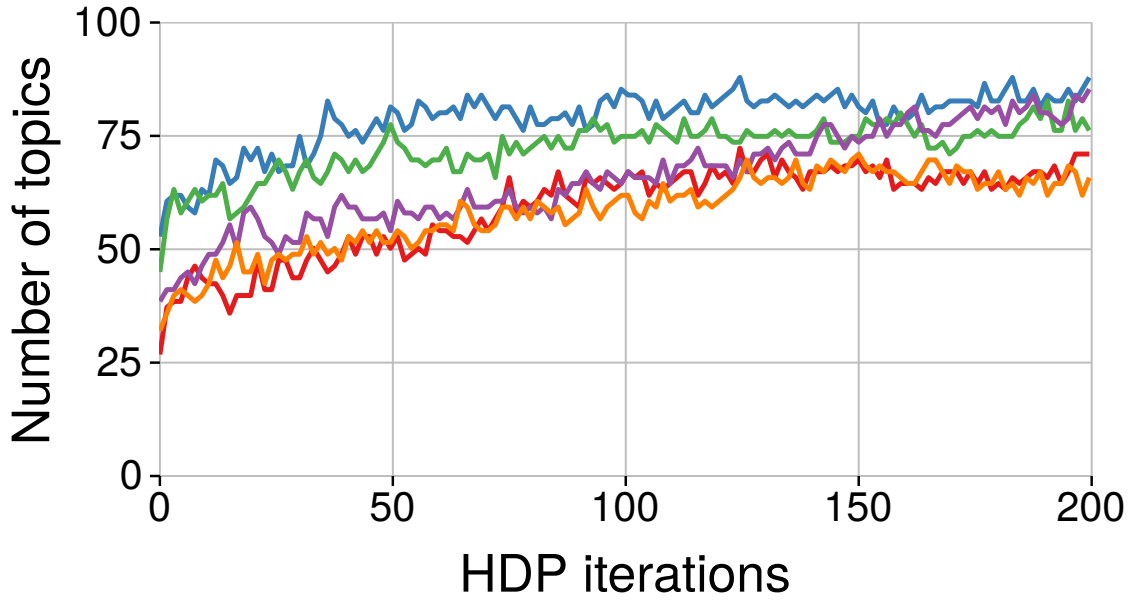
1. For synthetic datasets the proposed method reliably finds the true number of topics $|T_0| = 50$.
2. The range of τ values leading to the correct number is wide.
3. For real data the number of topics is not clear.

Comparison to HDP topic model



1. HDP (Hierarchical Dirichlet Process, Tech et. al, 2006) is the state-of-art approach for a number of topics optimization.
2. The choice of the concentration coefficient of Dirichlet process may lead to nearly any number of topics.

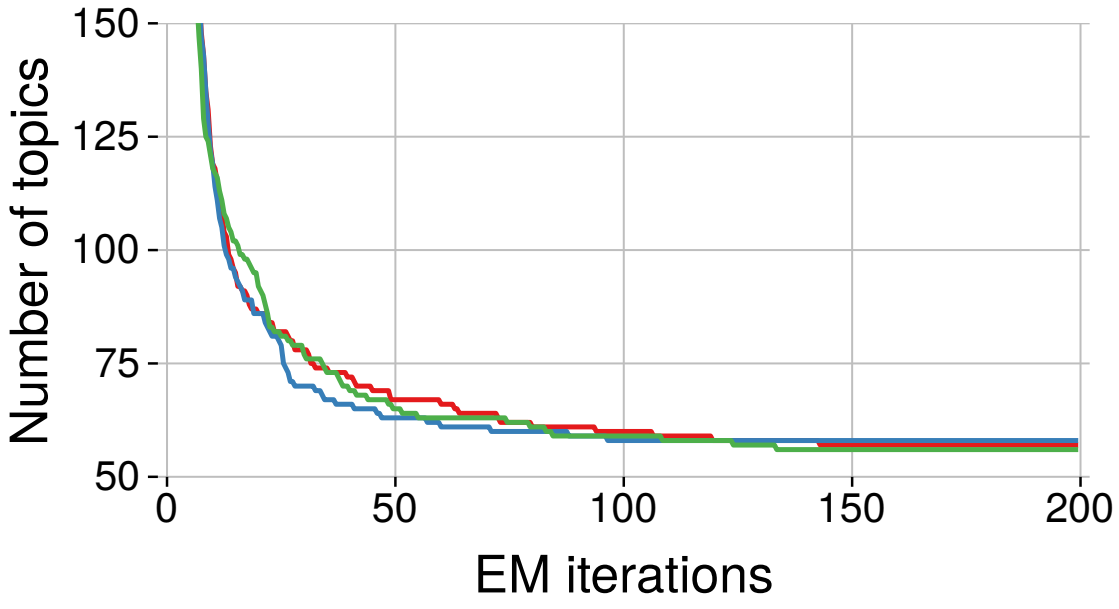
Comparison to HDP topic model



Moreover, HDP is unstable in two ways:

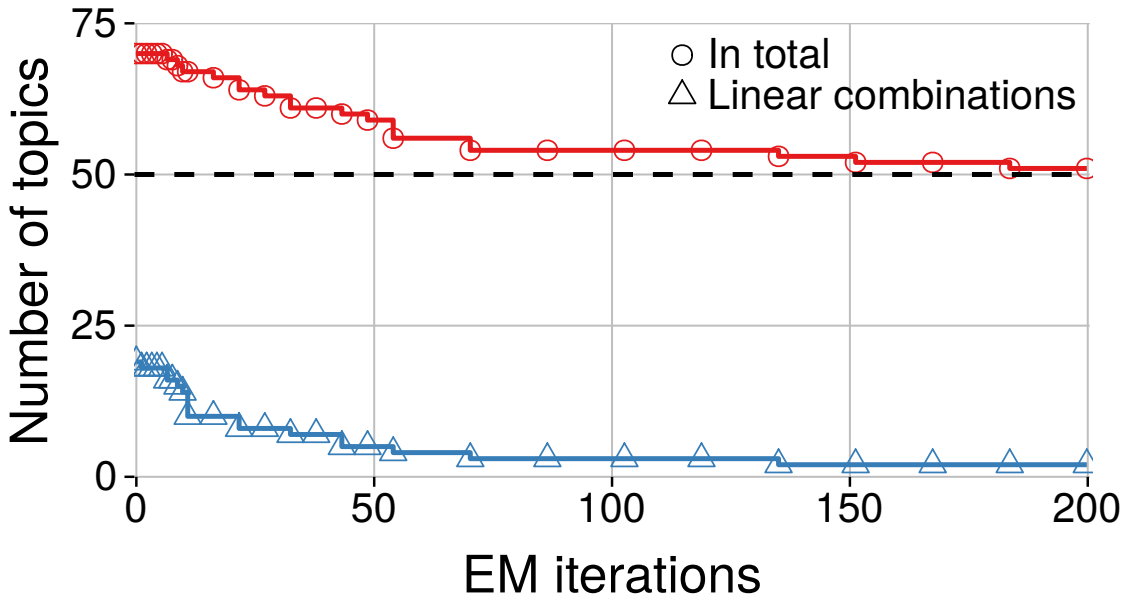
1. The number of topics fluctuates from iteration to iteration.
2. The results for several random starts significantly differ.

Stability of ARTM for topic selection



1. Our method is more stable in both ways.
2. Using the "recommended"parameters γ for HDP and τ for ARTM we get the similar number of topics approx. 60.

Elimination of linearly dependent topics



1. Add linearly dependent topics: 20 convex combinations of some of 50 topics for Π_0 dataset.
2. More sparse and diverse topics of the original model remain.

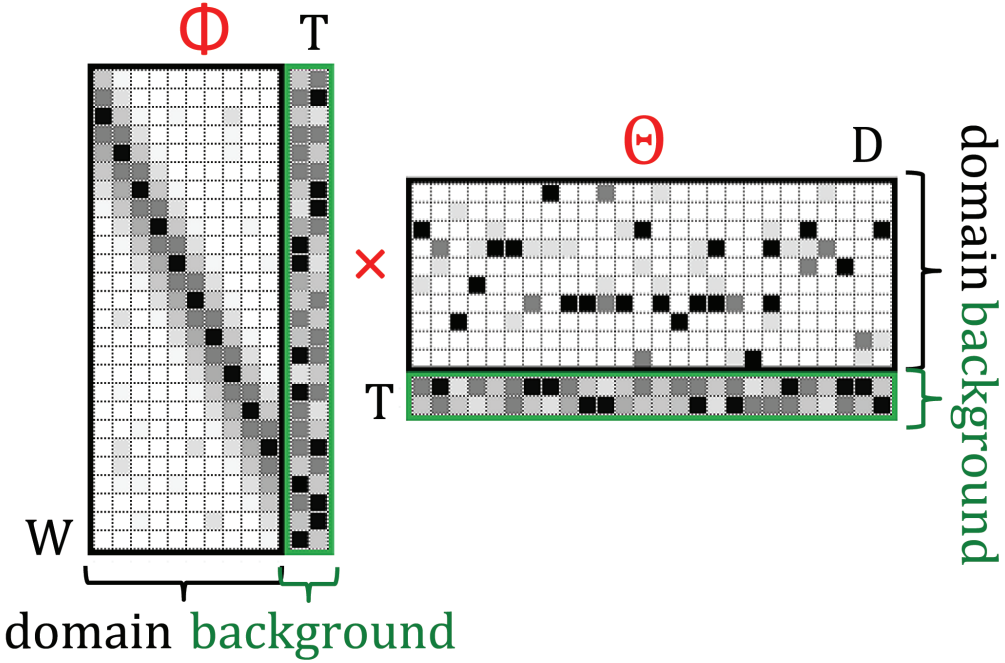
The experiments with sparse decorrelated model

Domain topics: specific terminology of domain area

1. Sparsing regularization for the parts of Φ and Θ
2. Decreasing topic correlation

Background topics: stop-words and common lexis

1. Smoothing regularization for the parts of Φ and Θ



Topics examples (top-20 words)

- **PLSA:** face, images, faces, recognition, set, image, based, hme, facial, representation, view, figure, model, experts, network, human, expert, space, examples, system
- **ARTM, domain:** face, faces, facial, cottrell, pentland, gesture, lane, emotion, person, steering, appearance, baluja, setpoint, camera, tracking, pose, pomerleau, mouth, darrell, lighting
- **ARTM, background:** model, data, models, parameters, noise, neural, mixture, prediction, set, gaussian, likelihood, networks, test, figure, training, performance, network, number, input, results

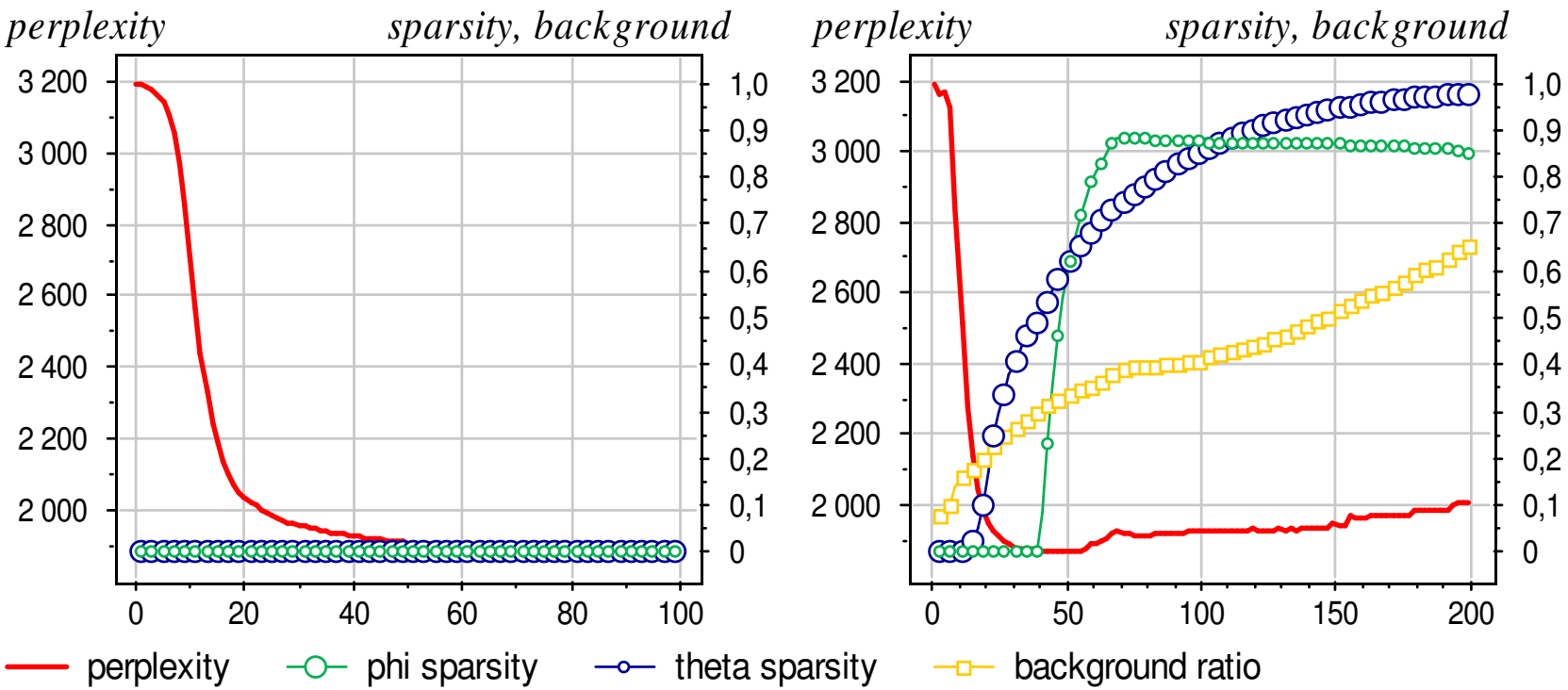
We type in red those words that are included into kernels.

Quality measures — multicriteria approach

- Held-out perplexity: $\mathcal{P} = \exp(-L/N)$
- Number of topics
- Sparsity — zeros ratio in Φ and Θ
- Topic coherence: $\frac{2}{k(k-1)} \sum_{i=1}^{10} \sum_{j=i+1}^{10} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$
- Topic kernel $W_t = \{w : p(t|w) > 0.25\}$ quality:
  - topic purity: $\sum_{w \in W_t} p(w|t)$ – the ratio of kernel in the topic
  - topic contrast: $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$ – diversity of kernels
- Ratio of word assignments to background topics

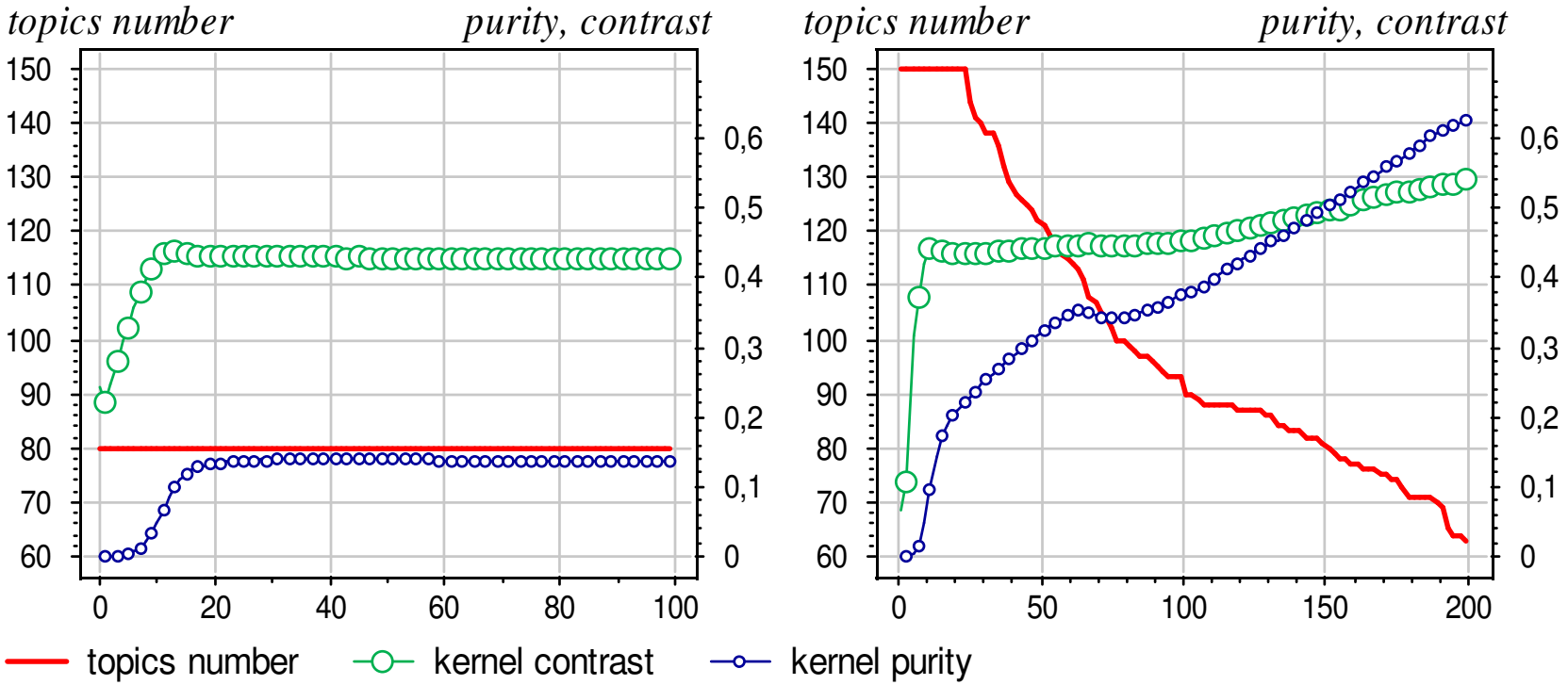
Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

Combination of sparsing, decorrelation and topic selection



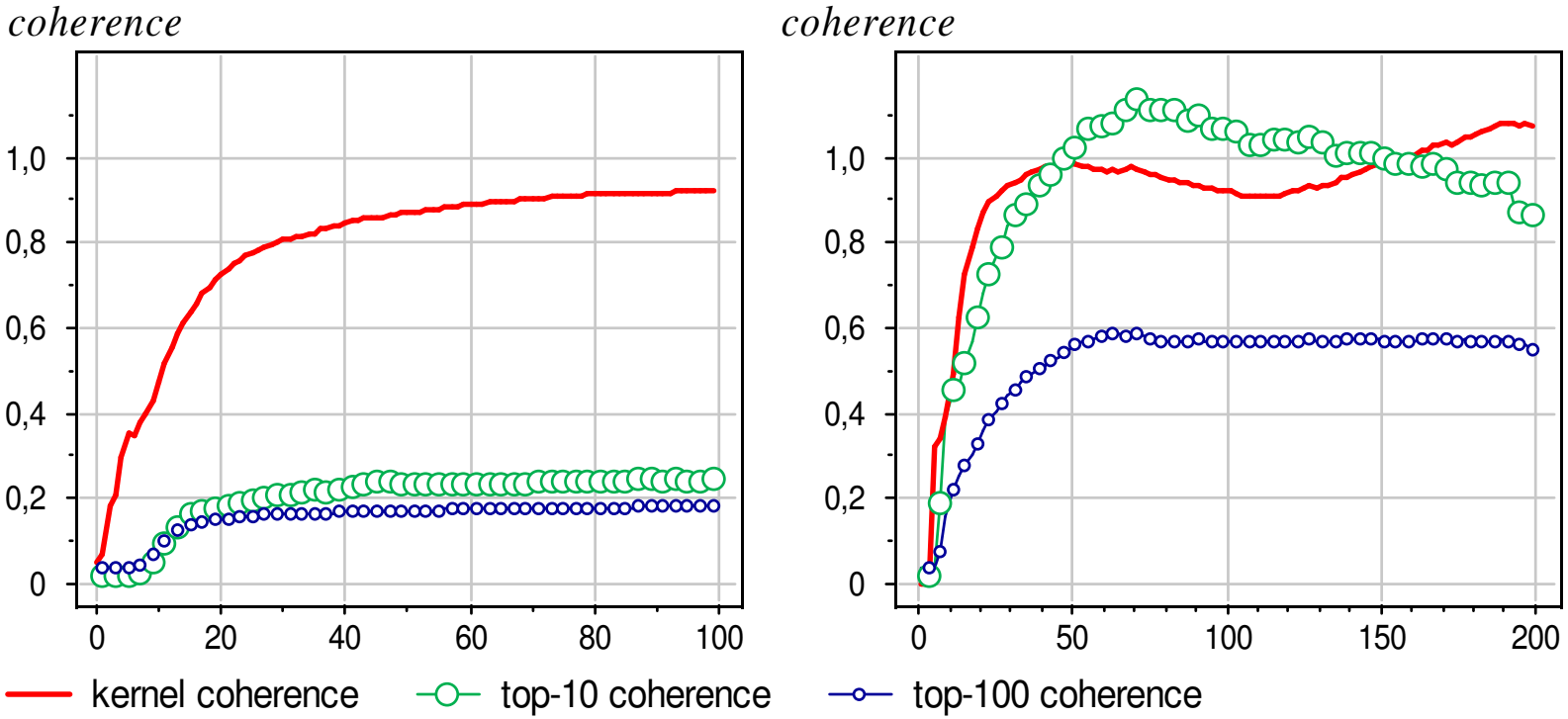
1. We achieve extremely high sparsity of Φ and Θ matrices.
2. Perplexity deteriorates mainly due to topics number decreasing.

Combination of sparsing, decorrelation and topic selection



1. The number of topics gradually decreases from 150 to 60 by eliminating the most insignificant topics at each iteration.
2. We can stop at the appropriate moment by other criteria.
3. The proposed model achieves high topics purity and contrast.

Combination of sparsing, decorrelation and topic selection

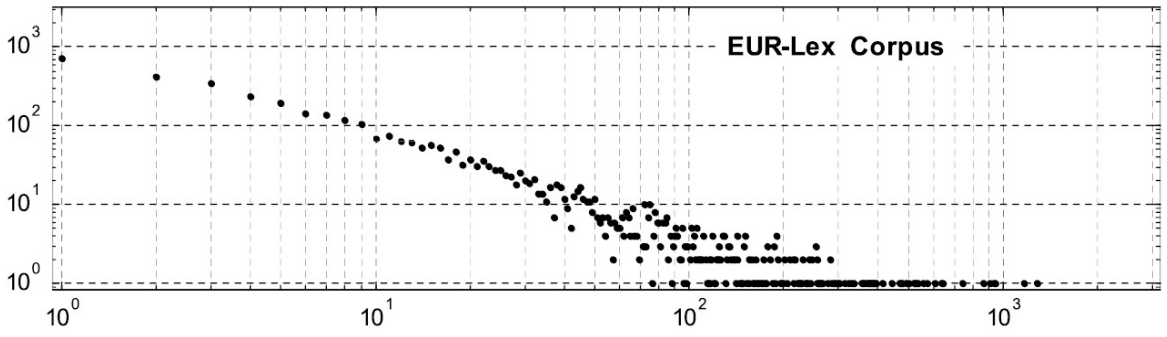


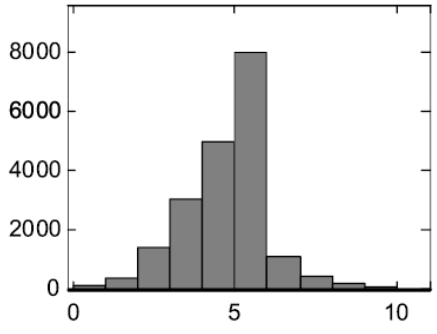
1. All kinds of coherence measures increase although it has not been explicitly incorporated into the optimization problem.

The experiments with Multi-ARTM for classification

EUR-Lex corpus

- 19 800 documents about European Union law
- Two modalities: 21K words, 3 250 categories (class labels)
- EUR-Lex is a "power-law dataset" with unbalanced classes:





- Left: # unique labels with a given # documents per label
- Right: # documents with a given # labels

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, 88(1-2). Pp. 157–208.

The experiments with Multi-ARTM for classification

Regularizers:

- Uniform smoothing for Θ
- Uniform smoothing for word–topic matrix Φ^1
- *Label regularization* for class–topic matrix Φ^2:

$$R(\Phi^2) = \tau \sum_{c \in W^2} \hat{p}_c \ln p(c) \to \max,$$

where
$p(c) = \sum_{t \in T} \phi_{ct} p(t)$ is the model distribution of class c,
$p(t) = \frac{n_t}{n}$ can be easily estimated along EM iterations,
$\hat{p}_c$ is the empirical frequency of class c in the training data.

The comparative study of models on EUR-Lex classification task

DLDA (Dependency LDA) [Rubin 2012] is a nearest analog of Multi-ARTM for classification among Bayesian Topic Models

Quality measures [Rubin 2012]:

- AUC-PR (%, $\Uparrow$) — Area under precision-recall curve
- AUC (%, $\Uparrow$) — Area under ROC curve
- OneErr (%, $\Downarrow$) — One error (most ranked label is not relevant)
- IsErr (%, $\Downarrow$) — Is error (no perfect classification)

Results:

	$\lvert T\rvert_{\text{opt}}$	AUC-PR	AUC	OneErr	IsErr
Multi-ARTM	10 000	**51.3**	98.0	**29.1**	**95.5**
DLDA [Rubin 2012]	200	49.2	**98.2**	32.0	97.2
SVM		43.5	97.5	31.6	98.1

The experiments with multi-language corpus

We consider languages as modalities in Multi-ARTM.

Collection of 216 175 Russian–English Wikipedia articles pairs.
Top 10 words by $p(w|t)$ probabilities:

Topic 68		**Topic 79**	
research	институт *(institute)*	goals	матч *(match)*
technology	университет *(university)*	league	игрок *(player)*
engineering	программа *(program)*	club	сборная *(national team)*
institute	учебный *(educational)*	season	фк *(fc)*
science	технический *(technological)*	scored	против *(versus)*
program	технология *(technology)*	cup	клуб *(club)*
education	научный *(scientific)*	goal	футболист *(footballer)*
campus	исследование *(research)*	apps	гол *(goal)*
management	наука *(science)*	debut	забивать *(score)*
programs	образование *(education)*	match	команда *(team)*

Multi-language ARTM

Collection of 216 175 Russian–English Wikipedia articles pairs.
Top 10 words by $p(w|t)$ probabilities:

Topic 88		**Topic 251**	
opera	опера *(opera)*	windows	windows
conductor	оперный *(opera)*	microsoft	microsoft
orchestra	дирижер *(conductor)*	server	версия *(release)*
wagner	певец *(singer)*	software	приложение *(application)*
soprano	певица *(singer)*	user	сервер *(server)*
performance	театр *(theatre)*	security	server
mozart	партия *(role)*	mitchell	программный *(program)*
sang	сопрано *(soprano)*	oracle	пользователь *(user)*
singing	вагнер *(Wagner)*	enterprise	обеспечение *(software)*
operas	оркестр *(orchestra)*	users	система *(system)*

All $|T| = 400$ topics were reviewed by an independent assessor, and he successfully interpreted 396 topics.

Conclusions and discussion

- ARTM (Additive Regularization for Topic Modeling) is a general framework, which makes topic models easy to design, to infer, to explain, and to combine.
- Multi-ARTM is a further generalization of ARTM for multimodal topic modeling
- BigARTM is an open source project for parallel online topic modeling of large text collections.

Conclusions and discussion

- ARTM (Additive Regularization for Topic Modeling) is a general framework, which makes topic models easy to design, to infer, to explain, and to combine.
- Multi-ARTM is a further generalization of ARTM for multimodal topic modeling
- BigARTM is an open source project for parallel online topic modeling of large text collections.

Contacts:

- Prof. Konstantin Vorontsov: voron@forecsys.ru
- Oleksandr Frei: oleksandr.frei@gmail.com
- Anna Potapenko: anya_potapenko@mail.ru