Additive Regularization for Topic Modeling: theory, implementation, applications

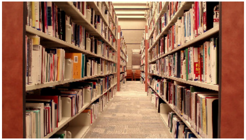
Konstantin Vorontsov
• Moscow Institute of Physics and Technology •
• Yandex School of Data Analysis •
• FORECSYS • AITHEA •

Higher School of Economics
Moscow • September 13, 2017

Contents

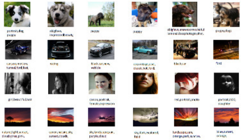
Topic modeling applications

exploratory search in digital libraries

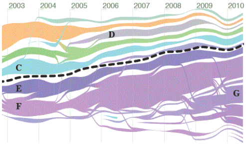


personalized search in social media

multimodal search for texts and images

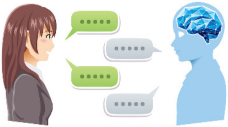


topic detection and tracking in news flows



navigation in big text collections

dialog manager in chatbot intelligence



What is a "topic" in a text collection

- *Topic* is a specific terminology of a particular domain area
- *Topic* is a set of terms that often co-occur in documents

More formally,

- *topic* is a probability distribution over terms:
  $p(w|t)$ is the frequency of word w in topic t
- *document profile* is a probability distribution over *topics*:
  $p(t|d)$ is the frequency of topic t in document d

When writing term w in document d author thought of topic t.

Topic model uncovers the set T of latent topics in a text collection.

Problem setup

Given: a set of terms W, a set of documents D,

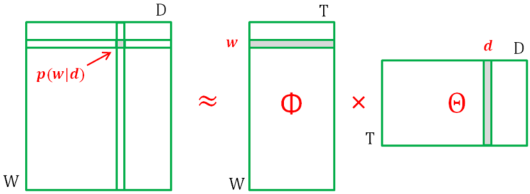
n_{dw} = how many times term w appears in document d

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

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

subject to $\phi_{wt} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$.

This is a problem of *nonnegative matrix factorization*:



PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Constrained maximization of the log-likelihood:

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

EM-algorithm is a simple iteration method for the nonlinear system

$$\begin{cases} \text{E-step:} & p_{tdw} \equiv p(t\,|\,d,w) = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \text{M-step:} & \phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum_{d\in D} n_{dw}\,p_{tdw}\Bigr) \\ & \theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in d} n_{dw}\,p_{tdw}\Bigr) \end{cases}$$

where $\operatorname*{norm}_{t\in T} x_t = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$ is vector normalization.

Well-posed and ill-posed problems in the sense of Hadamard (1923)

The problem is *well-posed* if

- a solution exists,
- the solution is unique,
- the solution is stable w.r.t. initial conditions.

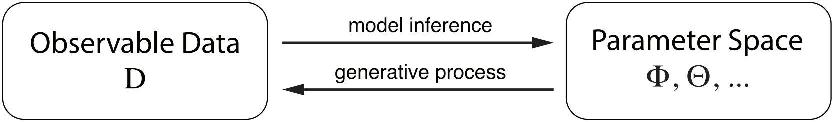
Jacques Hadamard (1865–1963)

Matrix factorization is an *ill-posed* inverse problem.
If (Φ, Θ) is a solution, then (Φ', Θ') is also the solution:

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, where $\operatorname{rank} S = |T|$
- $\mathscr{L}(\Phi', \Theta') = \mathscr{L}(\Phi, \Theta)$
- $\mathscr{L}(\Phi', \Theta') \leqslant \mathscr{L}(\Phi, \Theta) + \varepsilon$ for approximate solutions

Additional *regularizing criteria* should narrow the set of solutions.

A variety of data, parameters and requirements in Topic Modeling



More Data:

meta-data
linked data
transactional data
usage data
multilanguage data
co-occurrence data
(semi-)supervised data
linguistic data:
syntax, ontology etc.

Requirements:

topic interpretability
topic sparsity
topic diversity
topic selection
short texts

Tech. requirements:

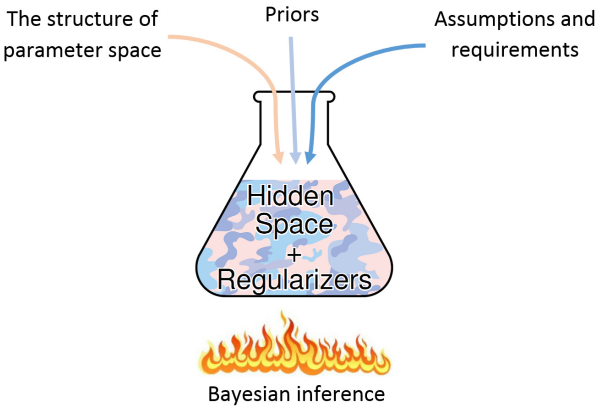
huge data
online processing
parallel processing

More Parameters:

temporal
hierarchical
multimodal
relational/graph
topic correlation
classification
regression
segmentation
n-gram

Bayesian approach in Topic Modeling

The *generative process* encapsulates all our knowledge about the hidden space structure, prior distributions, and requirements



The limitations of Bayesian approach for Topic Modeling

The artificial complication of the task:

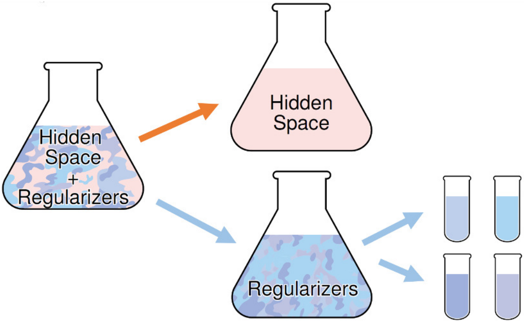
- Generative process encapsulates all we know about the problem
- Because of this, estimation of posteriors is a difficult task
- Nevertheless, posteriors are used only for point estimations
- Bayesians solve a more difficult task than it is necessary for PTM!

From this, many limitations stem:

- The solution requires a lot of math and coding for each model
- There is no way to unify models in a LEGO-style technology
- There is no easy way to combine topic models
- There is no way to impose non-probabilistic constraints
- There is no way to specify optimization criteria for the model

The classical non-Bayesian regularization for Topic Modeling

- A *simple generative process* describes the hidden space
- Regularizers describe most of the requirements and assumptions
- Regularizers can be additively mixed and interchanged



LDA — Latent Dirichlet Allocation [Blei, Ng, Jordan, 2003]

Maximum a posteriori probability (MAP) with Dirichlet prior.
The prior can be reinterpreted as cross-entropy minimization:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}}_{\text{cross-entropy regularization}} \to \max_{\Phi,\Theta}$$

EM-algorithm is a simple iteration method for the system

$$\begin{aligned} &\text{E-step:} \quad \begin{cases} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw} p_{tdw} + \beta_w\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw} p_{tdw} + \alpha_t\Big) \end{cases} \\ &\text{M-step (second and third lines)} \end{aligned}$$

ARTM — Additive Regularization of Topic Model

Maximum log-likelihood with regularization criterion $R(\Phi, \Theta)$:

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

EM-algorithm is a simple iteration method for the system

$$\begin{cases} \text{E-step:} & p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ \text{M-step:} & \phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Bigr) \\ & \theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Bigr) \end{cases}$$

K.Vorontsov. Additive regularization for topic models of text collections. 2014.

Combining topic models by adding their regularizers

Maximum log-likelihood with additive combination of regularizers:

$$\sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} + \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta},$$

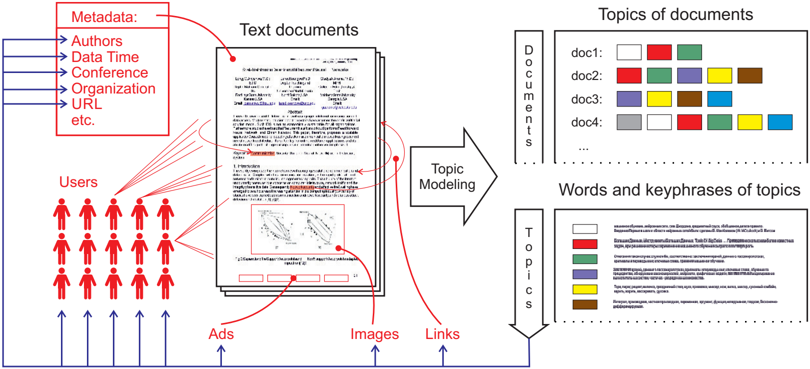
where τ_i are regularization coefficients.

EM-algorithm is a simple iteration method for the system

$$\begin{cases} \text{E-step:} & p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ \text{M-step:} & \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw} p_{tdw} + \sum_{i=1}^{n} \tau_i \phi_{wt} \frac{\partial R_i}{\partial \phi_{wt}}\Big) \\ & \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw} p_{tdw} + \sum_{i=1}^{n} \tau_i \theta_{td} \frac{\partial R_i}{\partial \theta_{td}}\Big) \end{cases}$$

Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topic distributions of terms $p(w|t)$ and other modalities: $p(\text{author}|t)$, $p(\text{time}|t)$, $p(\text{category}|t)$, $p(\text{tag}|t)$, $p(\text{link}|t)$, $p(\text{object-on-image}|t)$, $p(\text{user}|t)$, etc.



Multimodal extension of ARTM

W^m is a vocabulary of tokens of m-th modality, $m \in M$.

Maximum multimodal log-likelihood with regularization:

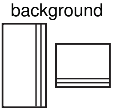
$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

EM-algorithm is a simple iteration method for the system

$$\begin{cases} \text{E-step:} & p_{tdw} = \operatorname*{norm}_{t \in T}\big(\phi_{wt}\theta_{td}\big) \\ \text{M-step:} & \phi_{wt} = \operatorname*{norm}_{w \in W^m}\Big(\sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\ & \theta_{td} = \operatorname*{norm}_{t \in T}\Big(\sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

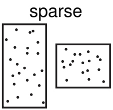
K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Ianina. Non-Bayesian additive regularization for multimodal topic modeling of large collections. 2015.

Regularizers for the interpretability of topics



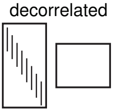
Smoothing background topics $B \subset T$:

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



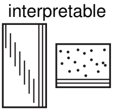
Sparsing subject domain topics $S = T \setminus B$:

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



Making topics as different as possible:

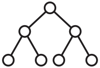
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_{w} \phi_{wt}\phi_{ws}$$



Making topics more interpretable
by combining the above regularizers

Many Bayesian PTMs can be reinterpreted as regularizers in ARTM

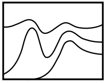
hierarchy



Hierarchical links between topics t and subtopics s:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}.$$

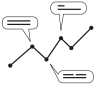
temporal



Topics dynamics over the modality of time intervals i:

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \left| \phi_{it} - \phi_{i-1,t} \right|.$$

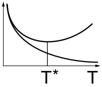
regression



Linear predictive model $\hat{y}_d = \langle v, \theta_d \rangle$ for documents:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Bigl(y_d - \sum_{t \in T} v_t \theta_{td} \Bigr)^2.$$

n of topics

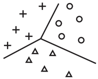


Sparsing $p(t)$ for topic selection:

$$R(\Theta) = -\tau \sum_{t \in T} \frac{1}{|T|} \ln p(t), \quad p(t) = \sum_{d} p(d) \theta_{td}.$$

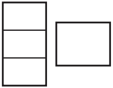
Special cases of the multimodal topic modeling

supervised



The modalities of classes or categories for text classification and categorization.

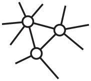
multilanguage



The modalities of languages with translation dictionary $\pi_{uwt} = p(u|w,t)$ for the $k \to \ell$ language pair:

$$R(\Phi, \Pi) = \tau \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in W^\ell} \pi_{uwt} \phi_{wt}$$

graph



The modality of graph vertices v with doc sets D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v) \in E} S_{uv} \sum_{t \in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|} \right)^2.$$

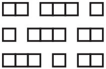
geospatial

The modality of geolocations g with proximity $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g' \in G} S_{gg'} \sum_{t \in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}} \right)^2$$

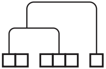
Beyond the "bag-of-words" restrictive hypothesis

n-gram



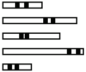
The modalities of n-grams, collocations, named entities

syntax



The modality of n-grams after SyntaxNet preprocessing

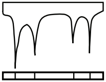
coherence



Modeling co-occurrence data n_{uv} for biterms (u, v):

$$R(\Phi) = \tau \sum_{u,v} n_{uv} \ln \sum_{t} n_t \phi_{ut} \phi_{vt}$$

segmentation



E-step regularization affecting $p(t|d, w)$ distributions for segmentation and sentence topic models

E-step regularization for bypassing the “bag-of-words” hypothesis

Maximum log-likelihood with regularizers R и $\tilde{R}$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Pi(\Phi,\Theta)) + \tilde{R}(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta},$$

where $\Pi = \bigl(p_{tdw}\bigr)_{T\times D\times W}$ is a matrix of conditionals $p_{tdw} = p(t|d,w)$.

EM-algorithm is a simple iteration method for the system

$$\begin{cases} \text{E-step:} & p_{tdw} = \operatorname{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\ & \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \\ \text{M-step:} & \phi_{wt} = \operatorname{norm}_{w\in W}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \phi_{wt}\frac{\partial \tilde{R}}{\partial \phi_{wt}}\right) \\ & \theta_{td} = \operatorname{norm}_{t\in T}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}}\right) \end{cases}$$

BigARTM project: open source for topic modeling

BigARTM features:

- Parallel + online + multimodal + regularized Topic Modeling
- Out-of-core one-pass 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 simplifies and unifies topic modeling for applications

Stages	Bayesian Inference for PTMs	ARTM	
Requirements analysis:	Requirements analysis	Requirements analysis	
Model formalization:	Generative model design	predefined criteria	user-defined criteria
Model inference:	Bayesian inference for the generative model (VI, GS, EP)	One regularized EM-algorithm for any combination of criteria	
Model implementation:	Researchers coding (Matlab, Python, R)	Production code (C++)	
Model evaluation:	Researchers coding (Matlab, Python, R)	predefined measures	user-defined measures
Deployment:	Deployment	Deployment	

conventions: ::: not unified stages ::: ::: unified stages :::

Bayesian models require maths and coding at each stage.
Therefore practitioners rarely go beyond a basic LDA model.
ARTM breaks this barrier by unifying the modeling process.

Benchmarking BigARTM vs. Gensim and 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

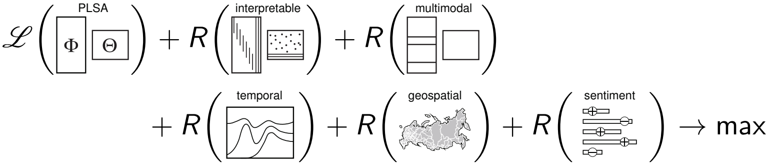
- *procs* = number of parallel threads
- *inference* = time to infer θ_d for 100K held-out documents
- *perplexity* is calculated on held-out documents.

Mining ethnical discourse in social media

Goal: finding all ethnical topics for monitoring inter-ethnic relations.

We have used 300 ethnonyms as seed words and modality.

The bag-of-regularizers:



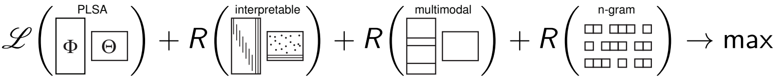
Result: the number of relevant topics augmented from 45 for LDA to 83 for ARTM.

M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.

Exploratory search in tech news

Goal: exploratory search by long text queries.

The bag-of-regularizers:



Results:

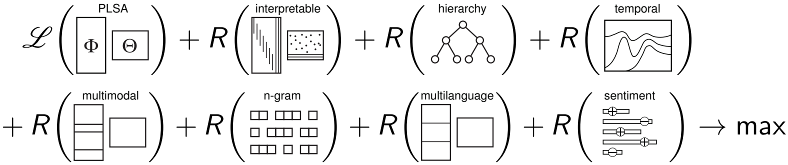
- Precision and Recall augmented from $(65\%, 73\%)$ for LDA to $(85\%, 92\%)$ for ARTM on Habrahabr.ru and TechCrunch.com tech news collections.
- Precision and Recall is comparable with assessors' quality.
- The topic-based search instantly performs the work that people typically complete in about 30 minutes.

A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Topic detection and tracking in news for media planning

Goal: the development of an interpretable hierarchical temporal dynamic topic model of the news flow.

The bag-of-regularizers:



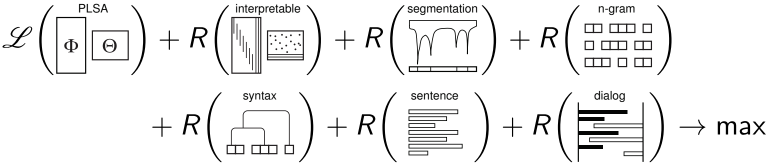
Results: ... (ongoing project)

Scenario analysis of call center records

Goals:

- determine typical scenarios of call-center dialogues between operators and customers
- elaborate the quantitative measure of how well operator works
- provide online tips for help operator handle customer's objections

The bag-of-regularizers:



Result: the quality of segmentation augmented from 40% for baselines to 75% for ARTM

Brief summary

- ARTM is a non-Bayesian regularization framework for PTM
- ARTM gives the easy way to formalize and combine PTMs
- ARTM makes it easier to understand and explain PTMs
- ARTM originates the modular “LEGO-style” PTM technology
- BigARTM: open source implementation of ARTM since 2014
- Now we are using ARTM for mining transaction data of any nature: communications, banking, e-learning.

http://bigartm.org

Welcome to use and make contributions!

References

- *K.Vorontsov.* Additive regularization for topic models of text collections. Doklady Mathematics, 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Ianina.* Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.
- *K.Vorontsov, A.Potapenko, A.Plavin.* Additive regularization of topic models for topic selection and sparse factorization. SLDS, 2015.
- *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova.* BigARTM: Open source library for regularized multimodal topic modeling of large collections. AIST, 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST, 2016.
- *M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov.* Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.
- *N.A.Chirkova, K.V.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
- *A.Ianina, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.