BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev,
Peter Romov, Marina Dudarenko

Yandex • CC RAS • MIPT • HSE • MSU

AIST

Analysis of Images, Social Networks and Texts
Ekaterinburg • 9–11 April 2015

Contents

What is “topic”?

- *Topic* is a special terminology of a particular domain area.
- *Topic* is a set of coherent terms (words or phrases) that often occur together in documents.
- Formally, *topic* is a probability distribution over terms: $p(w|t)$ is (unknown) frequency of word w in topic t.
- Document semantics is a probability distribution over *topics*: $p(t|d)$ is (unknown) frequency of topic t in document d.

Each document d consists of terms $w_1, w_2, \dots, w_{n_d}$:
$p(w|d)$ is (known) frequency of term w in document d.

When writing term w in document d author thinks about topic t.
Topic model tries to uncover latent topics from a text collection.

Goals and applications of Topic Modeling

Goals:

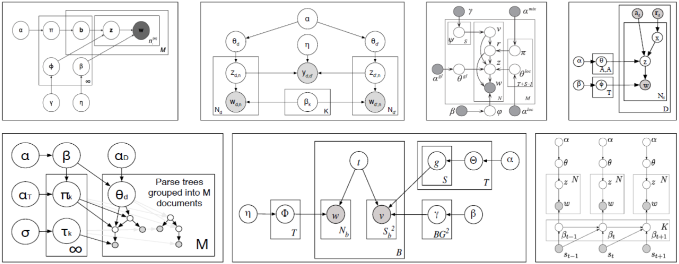
- Uncover a hidden thematic structure of the text collection
- Find a compressed semantic representation of each document

Applications:

- 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

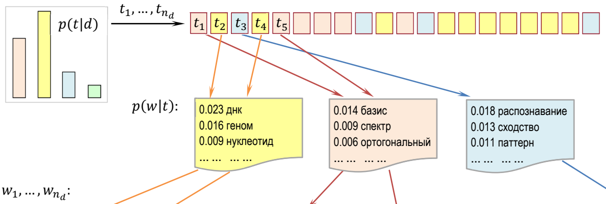


David Blei. Probabilistic topic models // Communications of the ACM, 2012. Vol. 55. No. 4. Pp. 77–84.

Generative Probabilistic Topic Model (PTM)

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).

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

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 non-negativeness and normalization constraints:

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

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

The problem of Topic Modeling is ill-posed

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

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

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

$P = \big\|p(w|d)\big\|_{W \times D}$ is known term–document matrix,

$\Phi = \big\|\phi_{wt}\big\|_{W \times T}$ is unknown term–topic matrix, $\phi_{wt} = p(w|t)$,

$\Theta = \big\|\theta_{td}\big\|_{T \times D}$ is unknown topic–document matrix, $\theta_{td} = p(t|d)$.

Matrix factorization is not unique and therefore it 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 $R_i(\Phi, \Theta) \to \max, \quad i = 1, \dots, n.$

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*.

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.

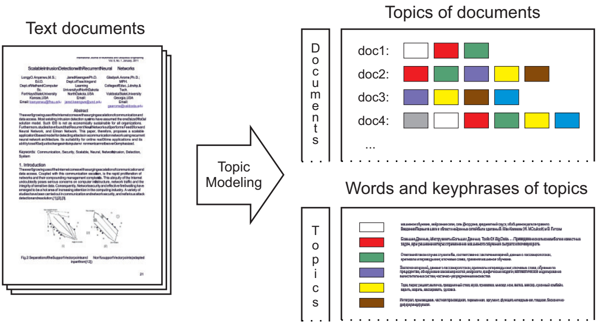
ARTM: available regularizers

- 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.

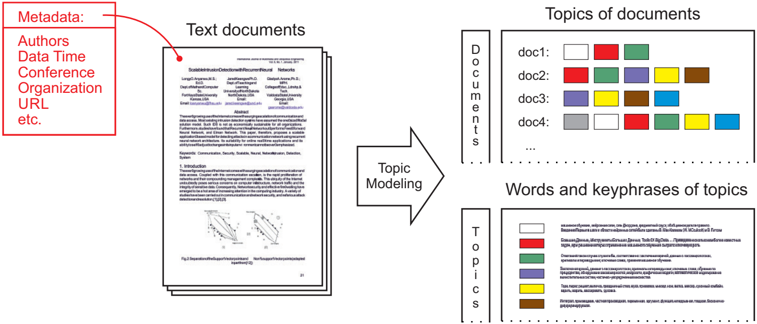
Multimodal Probabilistic Topic Modeling

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.



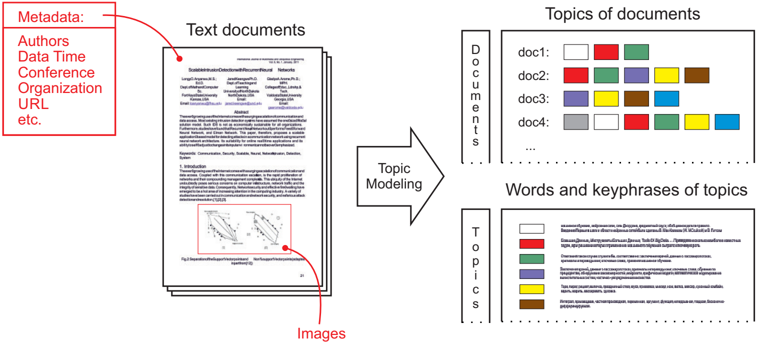
Multimodal Probabilistic Topic Modeling

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



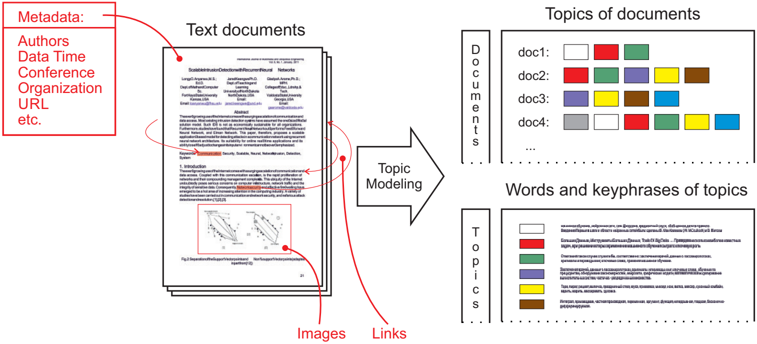
Multimodal Probabilistic Topic Modeling

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



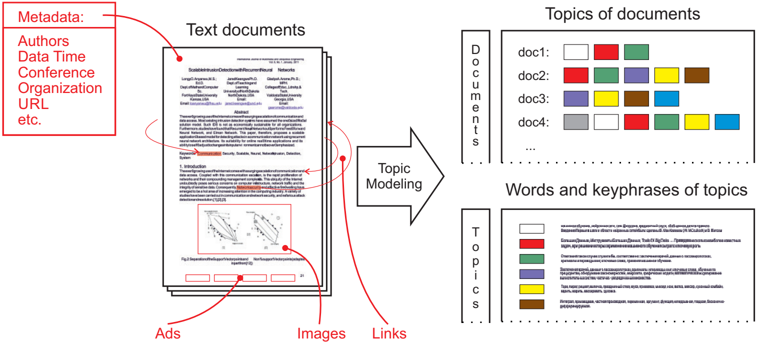
Multimodal Probabilistic Topic Modeling

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 $p(d'|t)$,



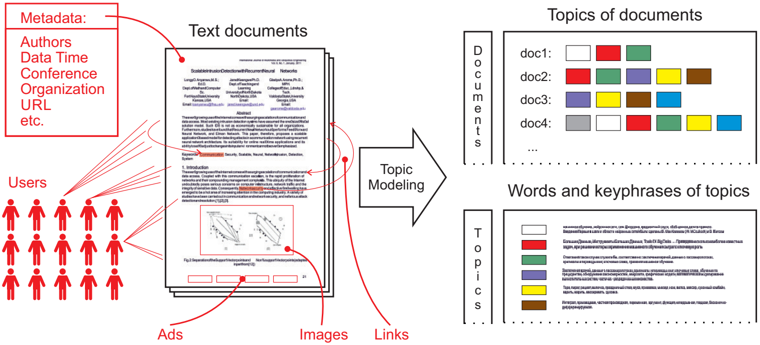
Multimodal Probabilistic Topic Modeling

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 $p(b|t)$,



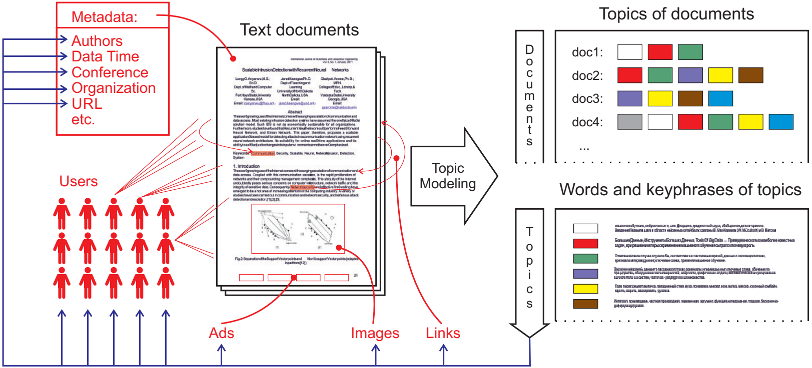
Multimodal Probabilistic Topic Modeling

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, $m \in M$
$W = W^1 \sqcup \dots \sqcup W^M$ 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*.

Multi-ARTM: multimodal regularized EM-algorithm

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

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

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

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

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

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

Online regularized EM-algorithm for Multi-ARTM

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

```
initialize φ_wt for all w ∈ W, t ∈ T;
n_wt := 0,  ñ_wt := 0 for all w ∈ W, t ∈ T;
for all batches D_b, b = 1, ..., B
    iterate each document d ∈ D_b at a constant matrix Φ:
    (ñ_wt) := (ñ_wt) + ProcessBatch(D_b, Φ);
    if (synchronize) then
        n_wt := n_wt + ñ_dw for all w ∈ W, t ∈ T;
        φ_wt := norm_{w∈W^m} (n_wt + φ_wt ∂R/∂φ_wt) for all w ∈ W^m, m ∈ M, t ∈ T;
        ñ_wt := 0 for all w ∈ W, t ∈ T;
```

$$\phi_{wt} := \operatorname*{norm}_{w\in W^m}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \text{ for all } w\in W^m,\ m\in M,\ t\in T;$$

Online regularized 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)

```
ñ_wt := 0 for all w ∈ W, t ∈ T;
for all d ∈ D_b
    initialize θ_td := 1/|T| for all t ∈ T;
    repeat
        p_tdw := norm_{t∈T}(φ_wt θ_td) for all w ∈ d, t ∈ T;
        n_td := Σ_{w∈d} λ_m(w) n_dw p_tdw for all t ∈ T;
        θ_td := norm_{t∈T}(n_td + θ_td ∂R/∂θ_td) for all t ∈ T;
    until θ_d converges;
    ñ_wt := ñ_wt + λ_m(w) n_dw p_tdw for all w ∈ d, t ∈ T;
```

$$\tilde{n}_{wt} := 0 \text{ for all } w \in W,\ t \in T;$$

for all $d \in D_b$

- initialize $\theta_{td} := \frac{1}{|T|}$ for all $t \in T$;
- **repeat**
  - $p_{tdw} := \operatorname{norm}_{t \in T}\big(\phi_{wt}\theta_{td}\big)$ for all $w \in d$, $t \in T$;
  - $n_{td} := \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw}$ for all $t \in T$;
  - $\theta_{td} := \operatorname{norm}_{t \in T}\Big(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big)$ for all $t \in T$;
- **until** θ_d converges;
- $\tilde{n}_{wt} := \tilde{n}_{wt} + \lambda_{m(w)} n_{dw} p_{tdw}$ for all $w \in d$, $t \in T$;

ARTM approach: benefits and restrictions

Benefits

- Single EM-algorithm for all 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

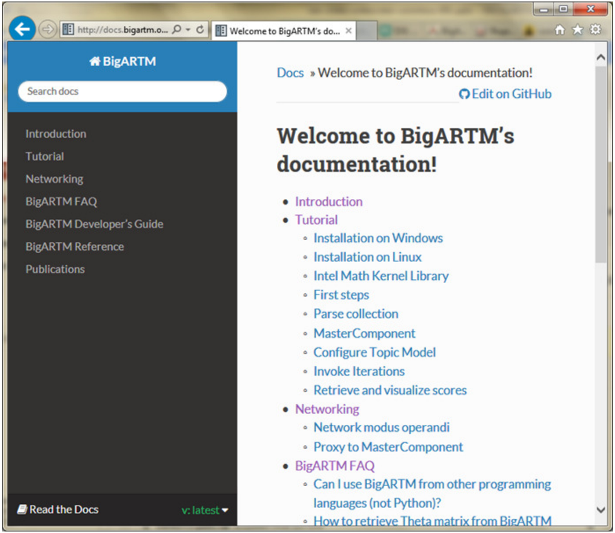
Under development (not really restrictions):

- Author-Topic Models based on 3-matrix factorization
- Further generalization of hypergraph-based Multi-ARTM
- Adaptive optimization of regularization coefficients

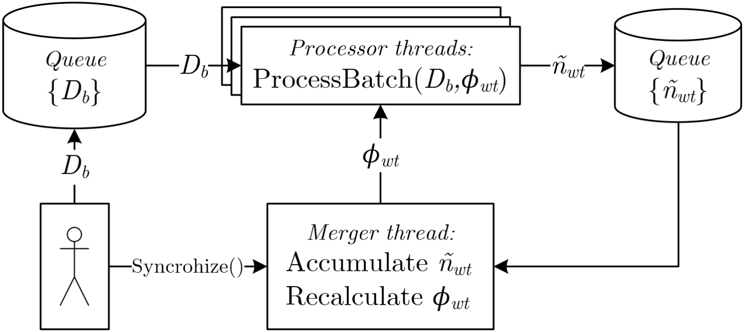
The BigARTM project: main features

- The full support of Multi-ARTM framework
- Open-source http://bigartm.org
- Parallel architecture
- Distributed storage of document collection
- 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: http://bigartm.org



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

BigARTM vs Gensim vs Vowpal Wabbit

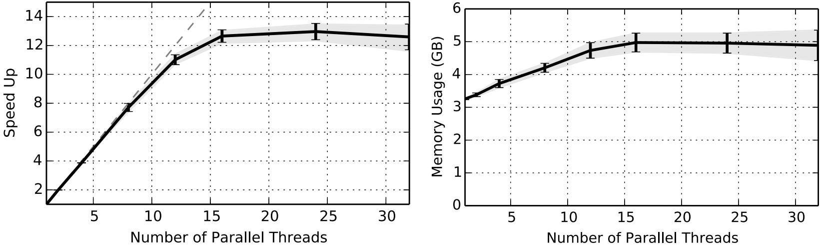
- 3.7M articles from Wikipedia, 100K unique words

	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 start using BigARTM

1. Download links, tutorials, documentation:
   http://bigartm.org
2. Compile and start examples
3. Post questions in BigARTM discussion group:
   https://groups.google.com/group/bigartm-users
4. Report bugs in BigARTM issue tracker:
   https://github.com/bigartm/bigartm/issues
5. Contribute to BigARTM project via pull requests:
   https://github.com/bigartm/bigartm/pulls

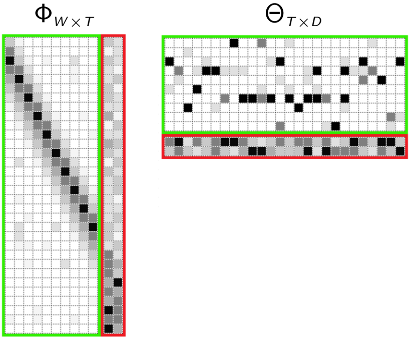
Limitations? Not really...

- 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

Combining Regularizers: experiment on 3.7M Wikipedia collection

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



Combining Regularizers: experiment on 3.7M Wikipedia collection

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

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

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

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

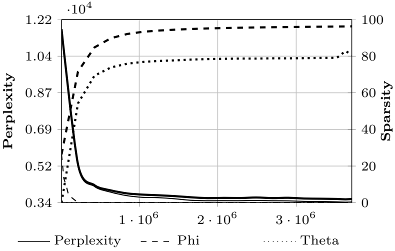
where β_0, α_0, β_1, α_1, γ are regularization coefficients.

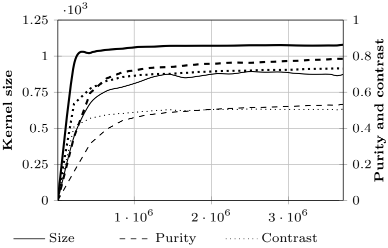
Combining Regularizers: LDA vs ARTM models

- $\mathcal{P}_{10k}$, $\mathcal{P}_{100k}$ — hold-out perplexity (10K, 100K documents)
  $\mathcal{S}_{\Phi}$, $\mathcal{S}_{\Theta}$ — sparsity of Φ and Θ matrices (in %)
  $\mathcal{K}_s$, $\mathcal{K}_p$, $\mathcal{K}_c$ — average topic kernel size, purity and contrast

Model	$\mathcal{P}_{10k}$	$\mathcal{P}_{100k}$	$\mathcal{S}_{\Phi}$	$\mathcal{S}_{\Theta}$	$\mathcal{K}_s$	$\mathcal{K}_p$	$\mathcal{K}_c$
LDA	3436	3801	0.0	0.0	873	0.533	0.507
ARTM	3577	3947	96.3	80.9	1079	0.785	0.731

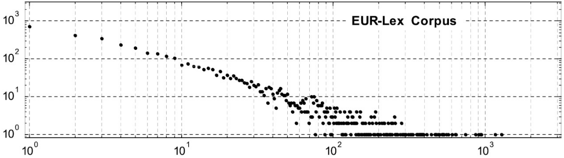
- Convergence of LDA (thin lines) and ARTM (bold lines)

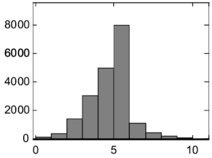




EUR-Lex corpus

- 19 800 documents about European Union law
- Two modalities: 210K 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.

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 (%, ⇑) — Area under precision-recall curve
- AUC (%, ⇑) — Area under ROC curve
- OneErr (%, ⇓) — One error (most ranked label is not relevant)
- IsErr (%, ⇓) — Is error (no perfect classification)

Results (Multi-ARTM with $|T| = 10^4$ topics):

	$\lvert T\rvert_{\text{opt}}$	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

Multi-language ARTM

We consider languages as modalities in Multi-ARTM.

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

Topic 68				Topic 79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Multi-language ARTM

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

Topic 88				Topic 251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

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

Conclusions

- 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.

http://bigartm.org

Join to BigARTM community!